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

2018

Proceedings of the 15th International Conference for Informatics and Information Technology

Held at Hotel Bistra, Mavrovo, Macedonia
20-22 April, 2018

Editors:
Georgina Mirceva
Natasha Ilievska

ISBN 978-608-4699-08-8

Conference for Informatics and Information Technology 2018
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 in 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:
Georgina Mirceva
Natasha Ilievska

Technical editing: Georgina Mirceva and Natasha Ilievska

Cover page: Vangel Ajanovski

Total print run:150
Printed in Skopje, Macedonia, 2019

ISBN: **978-608-4699-08-8**

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

004.7:621.39(062)
004(062)

PROCEEDINGS of the 15th Conference for Informatics and Information Technology (15; 2018; Mavrovo) Proceedings of the 15th Conference for Informatics and Information Technology: CIIT 2018, April, 20-22 Mavrovo, Macedonia / editors Georgina Mirceva and Natasha Ilievska. - Skopje : Faculty of Computer Science and Engineering, 2019. - 310 стр. : граф. прикази ; 30 см

Библиографија кон трудовите

ISBN 978-608-4699-08-8
1. Mirceva, Georgina [уредник] 2. Ilievska, Natasha [уредник]

Preface

This volume contains the papers and abstracts presented at the 15th International Conference for Informatics and Information Technology (CIIT 2018) held on April 20-22, 2018 in Mavrovo, Macedonia. The conference was organized by the Faculty of Computer Science and Engineering, within the Ss. Cyril and Methodius University in Skopje, Republic of Macedonia.

In the fifteenth edition, the key CIIT conference mission remained to provide an opportunity for young researchers to present their work to a wider research community, but also facilitate multidisciplinary and regional collaboration. Despite the participation of scientists from the country, a remarkable number of participants from abroad attended the conference. Building on the success of the past fourteen conferences, this year the conference attracted a large number of submissions resulting in presentations of 44 full and short papers, 18 student papers and 5 abstracts, which were presented in eight regular sessions and three student sessions. Traditionally, the conference included student sessions presenting the work of the best undergraduate students, where they presented some of their ongoing work, or demonstrated practical implementations. Three best student papers and three best student presentations were awarded. The format of the conference allowed the participants to attend most of the talks that covered a diverse spectrum of research areas.

As a key note speaker we had the pleasure to have Prof. Ana Sokolova, associate professor at University of Salzburg, Salzburg, Austria, who gave a talk titled "Local Linearizability". We had three invited speakers covering different areas of the conference. Dimitar Nikolov, PhD, consultant at Altran AB, Malmö, Sweden, gave a talk on the topic "Ensuring correct operation in presence of faults". Mihaela Angelova, PhD, postdoctoral researcher at French Institute of Health and Medical Research, Paris, France, gave a talk on "Squamous lung carcinogenesis: immune evasion before tumor invasion". Tome Eftimov, PhD, researcher at Jožef Stefan Institute, Ljubljana, Slovenia, gave a talk titled "Are we aware of the importance of proper study analysis? Deep Statistical Comparison: a case study of meta-heuristic stochastic optimization algorithms".

Part of the conference success is owed to the support received from partners and sponsors: Ss. Cyril and Methodius University in Skopje, Sorsix, Software4Insurance, Macedonian Winemakers and S&T.

All in all, this year the CIIT conference has outgrown the role of being an excellent opportunity for young researchers to present their scientific growth, to a more premier role, that is to bring researchers together for establishing collaborative links between disciplines, for testing the ground for innovative ideas and for engaging the wider academic community.

January, 2019
Skopje

Georgina Mirceva
Natasha Ilievska

Organization

Conference chairs

Georgina Mirceva	Assistant professor - Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, Macedonia
Natasha Ilievska	Assistant professor - Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, Macedonia

Organizing Committee

Vesna Dimitrievska Ristovska	Assistant professor - Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, Macedonia
Andreja Naumoski	Assistant professor - Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, Macedonia
Petre Lameski	Assistant professor - Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, Macedonia
Ilinka Ivanoska	Assistant doctorand - Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, Macedonia
Vesna Kirandziska	Assistant doctorand - Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, Macedonia

Program Committee

Adrijan Božinovski	School of Computer Science and Information Technology, University American College Skopje, Macedonia
Aleksandar Shurbevski	Kyoto University, Japan
Aleksandra Dedinec	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, Macedonia
Aleksandra Kovachev	Alacris Theranostics GmbH, Germany
Aleksandra Mileva	Faculty of Computer Science, University Goce Delcev, Macedonia
Aleksandra Popovska Mitrovikj	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, Macedonia

Alexandru Nicolin	Horia Hulubei National Institute for Physics and Nuclear Engineering and University of Bucharest, Faculty of Physics, Centre for Theoretical Physics, Romania
Ana Madevska Bogdanova	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, Macedonia
Andrea Kulakov	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, Macedonia
Andreja Naumoski	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, Macedonia
Antun Balaz	Institute of Physics, University of Belgrade, Serbia
Biljana Stojkoska	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, Macedonia
Biljana Tojtovska	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, Macedonia
Bojan Marinkovic	Mathematical Institute of the Serbian Academy of Sciences and Arts, Serbia
Bojan Pepik	Amazon Development Center, Germany
Bojana Koteska	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, Macedonia
Boro Jakimovski	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, Macedonia
Danco Davcev	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, Macedonia
Dejan Gjorgjevikj	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, Macedonia
Dejan Kovachev	SAP Innovation Center, Germany
Dejan Spasov	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, Macedonia
Dimitar Nikolov	Altran AB, Sweden
Dimitar Trajanov	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, Macedonia
Eftim Zdravevski	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, Macedonia
Eugenia Stoimenova	Institute of Mathematics and Informatics, Bulgarian Academy of Sciences, Bulgaria
Galina Bogdanova	Institute of Mathematics and Informatics, Bulgarian Academy of Sciences, Bulgaria
Georgina Mirceva	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, Macedonia
Gjorgji Madjarov	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, Macedonia
Goce Armenski	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, Macedonia
Goce Ristanoski	Data61, Commonwealth Scientific and Industrial Research Organisation, Australia
Goran Velinov	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, Macedonia

Haris Gavranovic	Faculty of Engineering and Sciences, International University of Sarajevo, Bosnia and Herzegovina
Hristijan Gjoreski	Faculty of Electrical Engineering and Information Technologies, Ss. Cyril and Methodius University in Skopje, Macedonia
Hristina Mihajlovska	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, Macedonia
Igor Mishkovski	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, Macedonia
Ivan Chorbev	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, Macedonia
Ivan Kitanovski	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, Macedonia
Ivaylo Donchev	St Cyril and St Methodius University of Veliko Turnovo, Bulgaria
Ivica Dimitrovski	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, Macedonia
Katarina Trojacanec	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, Macedonia
Katerina Zdravkova	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, Macedonia
Kire Trivodaliev	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, Macedonia
Konstantin Delchev	Institute of Mathematics and Informatics, Bulgarian Academy of Sciences, Bulgaria
Kosta Mitreski	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, Macedonia
Lasko Basnarkov	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, Macedonia
Ljupcho Antovski	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, Macedonia
Magdalena Kostoska	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, Macedonia
Marcin Paprzycki	Systems Research Institute, Polish Academy of Sciences, Poland
Margita Kon-Popovska	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, Macedonia
Marija Mihova	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, Macedonia
Marija Slavkovik	University of Bergen, Norway
Marjan Gusev	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, Macedonia
Mihaela Angelova	Laboratory of Integrative Cancer Immunology, French Institute of Health and Medical Research, France
Mile Jovanov	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, Macedonia
Milos Jovanovik	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, Macedonia

Miroslav Mirchev	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, Macedonia
Natasha Ilievska	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, Macedonia
Natasha Stojkovikj	Faculty of Computer Science, University Goce Delcev, Macedonia
Nevena Ackovska	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, Macedonia
Ognjen Scekic	Distributed Systems Group, Vienna University of Technology, Austria
Pance Ribarski	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, Macedonia
Petre Lameski	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, Macedonia
Sahra Sedigh Sarvestani	Missouri University of Science and Technology, USA
Sasho Gramatikov	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, Macedonia
Sasko Ristov	University of Innsbruck, Austria
Simona Samardjiska	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, Macedonia
Slavcho Shtrakov	Neofit Rilsky South-West University, Bulgaria
Slobodan Kalajdziski	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, Macedonia
Smile Markovski	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, Macedonia
Smilka Janeska Sarkanjac	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, Macedonia
Sonja Gievska	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, Macedonia
Stela Zhelezova	Institute of Mathematics and Informatics, Bulgarian Academy of Sciences, Bulgaria
Stojan Trajanovski	Philips Research, Netherlands
Stoyan Kapralov	Technical University - Gabrovo, Bulgaria
Suzana Loshkovska	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, Macedonia
Tome Eftimov	Jožef Stefan Institute, Slovenia
Tsonka Baicheva	Institute of Mathematics and Informatics, Bulgarian Academy of Sciences, Bulgaria
Vangel Ajanovski	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, Macedonia
Veno Pachovski	School of Computer Science and Information Technology, University American College Skopje, Macedonia
Verica Bakeva	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, Macedonia
Vesna Dimitrievska Ristovska	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, Macedonia
Vesna Dimitrova	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, Macedonia

Vesna Prckovska	Qmenta, Spain
Vladimir Trajkovikj	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, Macedonia
Yuri Borissov	Institute of Mathematics and Informatics, Bulgarian Academy of Sciences, Bulgaria
Zaneta Popeska	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, Macedonia
Zlatko Varbanov	St Cyril and St Methodius University of Veliko Turnovo, Bulgaria

Table of Contents

Papers

Education and e-Learning

eWorld, eBusiness and eCommerce

Abstracts

Student Papers

PAPERS

Using Distributed Representations to Identify Genders and Age Groups of Twitter Users

Dario Gjorgjevski, Dejan Gjorgjevikj
Faculty of Computer Science and Engineering
Ss. Cyril and Methodius University
Skopje, Macedonia
dario.gjorgjevski@gmail.com, dejan.gjorgjevikj@finki.ukim.mk

***Abstract*—This article describes a deep neural architecture for identifying genders and age groups of Twitter users based on the text in their so-called tweets. We employ distributed representations trained from large corpora consisting of social media postings from Twitter, personal blogs, review sites, etc. To accommodate the task, we propose a modification to the Doc2Vec framework that allows us to infer a vector for each Twitter user, which is then used as one of the two inputs to a classification algorithm. The second input consists of the word vectors inferred from the words appearing the user's tweets. The word vectors are fed into a sequence of convolutional neural networks with max pooling, whose output is then concatenated with the corresponding "vector," and finally fed into a densely-connected neural network for classification. Using 1/3 of the data for validation, we obtain an accuracy of 81.30 % for genders and 61.18 % for age groups. These results are at least comparable to (and in certain cases better than) the state-of-the-art performance reported in the PAN tasks on digital text.**

***Keywords*—Author profiling, classification, deep learning, distributed representations, unsupervised learning**

I. Introduction

The huge volume of user generated data on social media makes it appealing to have systems that can profile users based on their activity. This has obvious advantages in terms of commercial exploitation, e.g., in targeted marketing and advertising; as well as in the forensic and security areas. These considerations have led to the organization of an *author profiling* task at PAN[1].

Author profiling is the task of identifying certain features of authors of text. Here, we focus on the *genders* and *age groups* of users of the Twitter website. Genders are either *male* or *female*, and the five age groups are 18–24, 25–34, 34–49, 50–64, and 65–. The dataset is provided as part of PAN, and consists of tweets in English, Dutch, and Spanish. Due to language-specific details in the preprocessing steps and time constraints, we will focus only on tweets written in English.

A. Dataset Description

The main dataset consists of 277 792 tweets written in English. These tweets were written by 436 users, with an average of 637.14 tweets per user.

Of these 436 users, 218 are male and 218 are female; in other words, the dataset is completely balanced with respect to genders. The distribution of age groups is given in Table I. There is an obvious imbalance in the distribution of age groups; however, we did not see any benefit to using a technique such as SMOTE to attempt to correct this.

Table I
Distribution of Age Groups

Age group	Number of users
18–24	28
25–34	140
35–49	182
50–64	80
65–	6

B. Dataset Format

An author in the dataset is represented by an XML file containing all of the postings that the author has written on some social media website. Each file is named with a unique hexadecimal identifier which is afterwards used as an identifier for that particular author. We parsed the XML files using the Beautiful Soup[2] Python library.

Each directory containing the XML files also contains a file called `truth.txt`, which consists of lines of the form

```
<author-id>:::<gender>:::<age-group>
```

and gives information about the gender and age group of each author.

II. Preprocessing

Before learning any distributed representations or training the classification network we preprocess the data. The preprocessing consists of three steps:

Normalization of hash tags, replies, and external links
Hash tags are replaced by a new token called `hash_tag`, replies by `at_reply`, and external links by `external_link`. Normalization is first done on the HTML level using Beautiful Soup and analyzing the `<a>` tags within each document. Any leftovers are normalized heuristically using regular expressions.

[1] http://pan.webis.de/index.html

[2] https://www.crummy.com/software/BeautifulSoup/

```
<a href="/jensinkler"
   class="twitter-atreply pretty-link js-nav"
   dir="ltr"
   data-mentioned-user-id="16515297">
  <s>@</s><b>jensinkler</b>
</a>
<a href="/werkingrl"
   class="twitter-atreply pretty-link js-nav"
   dir="ltr"
   data-mentioned-user-id="223227186">
  <s>@</s><b>werkingrl</b>
</a>
lol! I used to be a gummy candy fiend.
Back in the day...
```

(a) Original HTML content

```
at_reply at_reply lol! I used
to be a gummy candy fiend.
Back in the day...
```

(b) Normalized content

Figure 1. HTML normalization example.

Figure 1 shows an example of this normalization. The `at_reply` normalization is performed based on the presence of the `"twitter-atreply"` class.

Lexical normalization Social media postings contain many irregular linguistic constructs such as "ur" to mean "your," "lol" to mean "laughing out loud," etc. In order to convert such words and abbreviations to their regular forms, we have constructed a dictionary for lexical normalization using data provided by Han, Cook, and Baldwin [1] and the ACL 2015 workshop on noisy user-generated text[3]. Table II shows some "Internet slang" words with their normalized counterparts.

Tokenization Each text was at the end tokenized using the Punkt tokenizer [2] from NLTK, the natural language toolkit[4]. At the end, all characters were converted to lowercase.

III. Distributed Representations

In this section, we review the Word2Vec and Doc2Vec frameworks for learning distributed representations of *words* and *documents*. We note that Doc2Vec is a straightforward extension to Word2Vec obtained by considering an additional information determining the context — the entire document. As a motivating example, consider the phrase `scored a <word>`. In a document related to sport, it is likely that the word `goal` will appear next; on the other hand, in a document related to business, `deal` would be more considerably more probable to appear.

Table II
Lexical Normalization Example

Original word	Normalized word
bfs	best friends
gents	gentlemen
mutha	mother
nite	night
ntmy	nice to meet you
u	you
wkend	weekend
would	would have

[3] https://noisy-text.github.io/2015/norm-shared-task.html
[4] http://www.nltk.org/api/nltk.tokenize.html

A. Encoding Similarities Between Words

Until very recently, machine learning was dominated by *localist representations*. These representations are popularly called "one-hot." They dedicate one unit to each word, and are easy to:

- understand;
- code by hand;
- learn — this is exactly what mixture models do; and
- associate with other representations and/or responses.

Example. In vector space terms, one-hot representations are vectors with one 1 and lots of 0s:

$$\begin{bmatrix}\cdots & 0 & 0 & 1 & 0 & 0 & \cdots\end{bmatrix}.$$

There are two major disadvantages, as these vectors:

- are extremely sparse; and
- do not encode similarities whatsoever.

To elaborate further on the second point, note that

$$\forall i \neq j . \boldsymbol{w}_i \wedge \boldsymbol{w}_j = \mathbf{0},$$

even if $\boldsymbol{w}_i$ and $\boldsymbol{w}_j$ are semantically similar. This is where *distributed representations* come in. They are based on the idea that a lot of information can be obtained by representing words by means of their neighbors. In other words, we learn *contexts*.

B. Learning Word Representations

In the Word2Vec framework, every word is mapped to a unique vector, represented by a column in the matrix $\boldsymbol{W}$. The column is indexed by position of the word in the vocabulary.

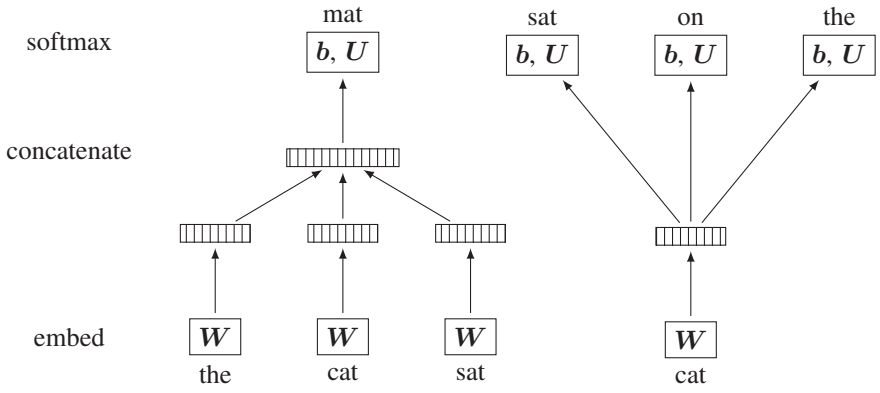


Figure 2. The continuous bag-of-words (left) and the skip-gram (right) models.

The concatenation or sum of the vectors is then used as features for prediction of the next word in a sentence.

More formally, given a sequence of training words $w_1, \ldots, w_T$, the objective of the word vector model is to maximize the average log-probability

$$\frac{1}{T}\sum_{t=c}^{T-c} \log p(w_t \mid w_{t-c}, \ldots, w_{t-1}, w_{t+1}, \ldots, w_{t+c}).$$

where c is the size of the training context (which in practice can be a function of the center word w_t). Larger c results in more training example and thus higher accuracy, at the cost of increased training time.

The prediction task is typically performed by a softmax classifier. Namely, we have

$$p(w_t \mid w_{t-c}, \ldots, w_{t-1}, w_{t+1}, \ldots, w_{t+c}) = \frac{\exp(y_{w_t})}{\sum_i \exp(y_i)}. \quad (1)$$

Each of the y_i in Eq. (1) is an un-normalized log-probability for each output word i, computed as

$$\boldsymbol{y} = \boldsymbol{b} + \boldsymbol{U}\boldsymbol{h}(w_{t-k}, \ldots, w_{t-1}, w_{t+1}, \ldots w_{t+k}; \boldsymbol{W}), \quad (2)$$

where $\boldsymbol{b}$ and $\boldsymbol{U}$ are softmax parameters, and $\boldsymbol{h}$ is constructed by a concatenation of word vectors extracted from $\boldsymbol{W}$. In practice, hierarchical softmax is preferred to softmax for fast training [3]. Additionally, the *negative sampling assumption* and *subsampling* are also applied [3].

1) Negative Sampling: The main idea behind negative sampling is given in Assumption 1.

Assumption 1 (Negative Sampling)**.** In order to learn the context, it is enough to update only the output word (*the "ground truth"*) along with a few words as negative samples.

The negative samples are drawn from a noise distribution $P_{\text{neg}}(w)$ which is determined empirically. Word2Vec uses the unigram distribution raised to the power of $3/4$.

2) Subsampling: In order to counter the imbalance between rare and frequent words, Word2Vec uses subsampling. Namely, each word in the training set is discarded with probability

$$1 - \sqrt{\frac{d}{f(w)}},$$

where $f(w)$ is the frequency of word w and d is the discard threshold, typically around 10^{-5}.

IV. Two Kinds of Models

The framework described in Section III can be used to either

1) predict a target word from source context words; or
2) predict source context words from a target word.

These two models are called *continuous bag-of-word* (CBOW) and *skip-gram*, respectively. The CBOW model is a feed-forward neural network with a single hidden layer which predicts target words from source context words. On the other hand, the skip-gram model predicts context words given a target word. The two models are illustrated on Fig. 2.

V. Learning Document Representations

Word vectors are asked to contribute to a prediction task about the next word in a sentence. So, despite the fact that they are initialized randomly, they can eventually capture semantics as an indirect result of the prediction task. We can use the same idea to learn *document vectors*, i.e., by asking them to contribute to the prediction task of the next word given many contexts sampled from the document [4]. This is known as the Doc2Vec framework.

Every document is now also mapped to a unique vector, represented by a column in the matrix $\boldsymbol{D}$. The document vector and word vectors are concatenated in order to predict the next word in a context.

Formally, the only difference is in Eq. (2) — $\boldsymbol{h}$ is now constructed not just from $\boldsymbol{W}$, but also from $\boldsymbol{D}$. The document token can be thought of as another word, acting as a memory that remembers what is missing from the current context. In contrast to the CBOW and skip-gram models, we now have the *distributed memory* and the *distributed bag-of-words* (DBOW) models, illustrated on Fig. 3. In the figure, `id` is a unique identifier for a document.

VI. The Classification Pipeline

This section describes the classification pipeline: how the word vectors are learned, how we modify the Doc2Vec framework in order to map authors to unique vectors, and

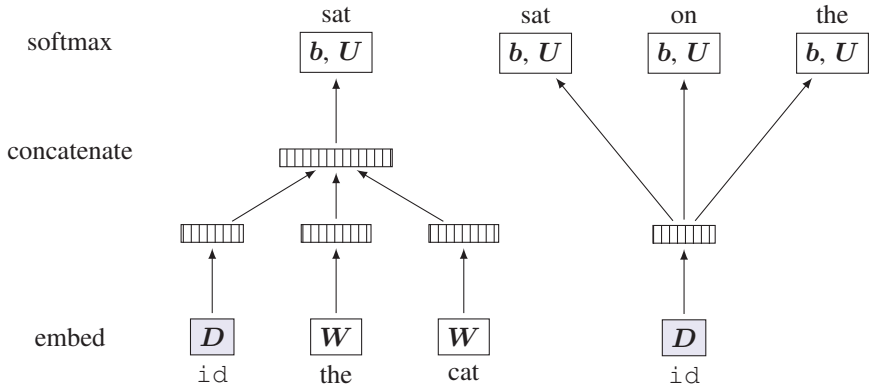


Figure 3. The distributed memory (left) and distributed bag-of-words (right) models.

how a deep neural network uses these vectors as inputs for classification.

A. Mapping Authors to Vectors

We propose a very simple modification to the Doc2Vec framework that allows us to map each author to a unique vector. Specifically, instead of considering a "document" (i.e., a social media post) as additional context, we consider the author instead. This results in the same models as on Fig. 3, except that `id` is now the *author's unique identifier*.

We used the Gensim toolkit [5] to learn the distributed representations. Data from other PAN tasks — totaling 1 047 358 social postings from Twitter, personal blogs, review sites, etc. — were used to learn word and author vectors. Both vectors are 256-dimensional, and were learned over 20 epochs by using a context window of size 8, ignoring words appearing less than 5 times. The vectors were estimated using 10 "noise words" and subsampling with rate 1×10^{-5}. Examples of similarities between word vectors are given in Fig. 4.

B. Building the Deep Neural Classifier

The classifier consists of three building blocks:

1) A convolutional neural network (CNN) used to extract features from the word vectors via 1-dimensional convolution;
2) A dense neural network used to "preprocess" the author vectors; and
3) A dense neural network leading to a softmax panel used for the final classification.

Before being fed to the CNN for feature extraction, the word sequences are padded to a length of 512, and are then embedded to a 512×256 "matrix." The CNN itself has two layers, both with 128 filters and a kernel size of 4; the first layer is max-pooled by a factor of 4, while the second by a factor of 16. Furthermore, both layers use rectified linear units (ReLU), defined as:

$$f(x) = \max\{0, x\}.$$

These units allow for faster and more effective training and have been shown to outperform units with sigmoid or hyperbolic tangent activation [6].

Independently of the feature extraction step, the author is mapped to a vector which is then fed to a dense neural network consisting of 128 ReLUs. This has the effect of "preprocessing" the vectors.

The output of feature extraction block is 896-dimensional, while the author vectors are mapped to 128 dimensions. These two outputs are concatenated and used as input to the classification block.

```
>>> model.most_similar('substantial')
[('higher', 0.72),
 ('large', 0.70),
 ('high', 0.69),
 ('significant', 0.6),
 ('huge', 0.6),
 ('reduced', 0.6),
 ('massive', 0.59),
 ('increased', 0.58),
 ('considerable', 0.57),
 ('enormous', 0.56)]
```

```
>>> model.most_similar('trump')
[('ivanka', 0.5),
 ('donald', 0.41),
 ('gould', 0.4),
 ('melania', 0.38),
 ('selby', 0.38),
 ('osullivan', 0.35),
 ('finalist', 0.34),
 ('hudson', 0.31),
 ('qingyang', 0.31),
 ('ferguson', 0.31)]
```

Figure 4. Similarities between learned word vectors.

The classification block is a dense neural network consisting of two layers of 128 and 32 ReLUs, respectively. It leads to a softmax panel to perform the final classification.

The dropout rate of the neural units in all three building blocks was set to 50 %. This was logical since word sequences are padded to lengths much longer than they usually are, so aggressive dropout makes sure that roughly half of the units are active in each step.

VII. Results and Future Work

At the end, we present our results, compare them to the state of the art, and outline possibilities for future work.

A. Results

The neural architecture was trained on 2/3 of the available data over three epochs and tested on the remaining 1/3. It was implemented in Keras[5] using the TensorFlow backend. The training took roughly 3 h and 20 min on an Intel® Core™ i5-4670 CPU @ 3.40 GHz. Accuracy on the testing set was computed to be 81.30 % for genders and 61.18 % for age groups. On the training set, accuracy was >95 % in both cases.

Comparison to the *best results* by *winning contestants* from previous PAN tasks is given in Table III. We stress that we did not have access to the same test data as the participants in the PAN task; therefore, as already mentioned, we used 1/3 of the training data for this purpose. However, recall that the distributed representations are trained in an entirely unsupervised manner. This means that our approximation of the architecture's accuracy is unlikely to suffer due to this. Even if it does, the results by other authors shown in Table III were obtained using *more* training data (as we used 1/3 for testing), so the comparison should be fair.

The most notable advantage of our approach — aside from the excellent results — is that it does not require any *feature engineering*. The distributed representations are learned in an entirely unsupervised manner, and the feature extraction from the word vectors is performed by a convolutional neural network. Training the distributed representations on additional data, which are almost trivial to obtain, will likely lead to even better performance.

B. Future Work

Firstly, we intend to improve our preprocessing step. Although the normalization dictionary that we constructed is quite sophisticated, it does not take context into account. Thus, "ur" will always be translated to "your," even though it might sometimes stand for "you are." This can be avoided by using a context-sensitive algorithm to decide which normalization to use based on POS tags and other such information. It might also be worthwhile to perform lemmatization of all texts.

We noted that over-fitting began to occur after the third epoch. This is somewhat unusual, but we leave it as future work to overcome and possibly improve on the classification We would like to explore the effects of different regularization techniques on the training of the neural architecture.

VIII. Conclusion

In this paper we presented a deep neural architecture for predicting the gender and age group of Twitter users according to the text of their posts. We proposed a modification to the Doc2Vec framework that infers a vector for each Twitter user, that is used in the classification algorithm. A distributed representation of the words contained the user's tweets in a form of word vectors is also an input to the classification algorithm.

We have observed that this approach is quite capable in predicting the gender of the author of the tweet post and very promising in predicting the age. The obtained results of 81.30 % for genders and 61.18 % for age groups are comparable and in some aspects even better than the state-of-the-art performance reported in this domain.

References

[1] B. Han, P. Cook, and T. Baldwin, "Automatically constructing a normalisation dictionary for microblogs," in *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, ser. EMNLP-CoNLL'12, Jeju Island, Korea: Association for Computational Linguistics, 2012, pp. 421–432.

[2] T. Kiss and J. Strunk, "Unsupervised multilingual sentence boundary detection," *Comput. Linguist.*, vol. 32, no. 4, pp. 485–525, Dec. 2006. DOI: 10.1162/coli.2006.32.4.485.

[3] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Proceedings of the 26th International Conference on Neural Information Processing Systems*, ser. NIPS'13, Lake Tahoe, Nevada: Curran Associates Inc., 2013, pp. 3111–3119.

[4] Q. Le and T. Mikolov, "Distributed representations of sentences and documents," in *Proceedings of the 31st International Conference on Machine Learning*, E. P. Xing and T. Jebara, Eds., ser. Proceedings of Machine Learning Research, vol. 32, Bejing, China, Jun. 2014, pp. 1188–1196.

[5] https://keras.io/

Table III
Comparison of Our Results to Winners of Previous PAN Tasks

Authors	Gender accuracy (%)	Age group accuracy (%)
This paper	81.30	61.18
López-Monroy, Montes-y-Gómez, Escalante, *et al.* [7]	62.51	43.61
Maharjan, Shrestha, and Solorio [8]	73.33	44.16
Álvarez-Carmona, López-Monroy, Montes-y-Gómez, *et al.* [9]	78.28	N/A

[5] R. Řehůřek and P. Sojka, "Software framework for topic modelling with large corpora," in *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, Valletta, Malta: ELRA, May 2010, pp. 45–50.

[6] X. Glorot, A. Bordes, and Y. Bengio, "Deep sparse rectifier neural networks," in *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, G. Gordon, D. Dunson, and M. Dudík, Eds., ser. Proceedings of Machine Learning Research, vol. 15, Fort Lauderdale, FL, USA: PMLR, Apr. 2011, pp. 315–323.

[7] A. López-Monroy, M. Montes-y-Gómez, H. Escalante, and L. Villaseñor-Pineda, "Using intra-profile information for author profiling — notebook for PAN at CLEF 2014," in *CLEF 2014 Evaluation Labs and Workshop – Working Notes Papers, 15–18 September, Sheffield, UK*, L. Cappellato, N. Ferro, M. Halvey, and W. Kraaij, Eds., CEUR-WS.org, Sep. 2014.

[8] S. Maharjan, P. Shrestha, and T. Solorio, "A simple approach to author profiling in MapReduce — notebook for PAN at CLEF 2014," in *CLEF 2014 Evaluation Labs and Workshop — Working Notes Papers, 15–18 September, Sheffield, UK*, L. Cappellato, N. Ferro, M. Halvey, and W. Kraaij, Eds., CEUR-WS.org, Sep. 2014.

[9] M. Álvarez-Carmona, A. López-Monroy, M. Montes-y-Gómez, L. Villaseñor-Pineda, and H. Escalante, "INAOE's participation at PAN'15: Author profiling task — notebook for PAN at CLEF 2015," in *CLEF 2015 Evaluation Labs and Workshop — Working Notes Papers, 8–11 September, Toulouse, France*, L. Cappellato, N. Ferro, G. Jones, and E. San Juan, Eds., CEUR-WS.org, Sep. 2015.

Towards Music Generation With Deep Learning Algorithms

Marko Docevski, Eftim Zdravevski, Petre Lameski, Andrea Kulakov
Faculty of Computer Science and Engineering
Sts. Cyril and Methodius University, Skopje, Macedonia
E-mails: mdocevski@gmail.com, {eftim.zdravevski,petre.lameski,andrea.kulakov}@finki.ukim.mk

***Abstract*—Computer music generation has application in many areas, including computer aided music composition, on demand music generation for video games, sport events, multi-media experiences, creating music in the style of passed away artists, etc. In this work we describe our approach towards music generation. We trained a deep learning model on a corpus of works of several authors. By priming the model with a snippet of an authors work we used it to create new music in their style. The dataset consists of music for guitar in midi format, containing only 1 part/instrument. We gathered more than 2000 files, of which we used from 5 to 300 per experiment. The data for the deep learning model is represented in piano roll format, a binary matrix where one axis represents the time and the other axis represents midi notes. Two deep learning architectures were evaluated, a 2-layer recurrent neural network of LSTM (Long Short Term Memory) cells and an Encoder-Decoder (Auto-Encoder) architecture for sequence learning, where both the encoder and decoder are built as recurrent layers of LSTM cells. The models were implemented in the Keras deep-learning library. The results were evaluated on a subjective basis, and with the evaluated datasets both architectures produced results of limited quality.**

***Index Terms*—music generation, midi, deep learning, recurrent neural networks, LSTM, auto-encoder**

I. Introduction

Computer music generation is an aspect of computer/artificial intelligence that has peaked interest in recent years. Opinions are split on whether computer generated music is just imitation of human work or if computers can show signs of actual creativity [1]. The opponents of computer creativity claim that no computer model can create anything novel because it would be based on a corpus of human work or human defined rule-set, which it would only recombine in arbitrarily abstract or complex way. On the other hand, most human musical composition is achieved in a similar way. People are exposed to music long before they can contribute their musical compositions, and therefore heavily influenced by other works, either consciously or subconsciously. By mimicking the human creative process and leveraging the ability of modern computers to process massive amounts of data, and inspired by the, by now legendary, article [2] by Andrej Karpathy, we would like to contribute to the discussion with our experiments with deep learning models and try to show computer creativity.

Other than adding to the philosophical discussion, we see practical applications for a music generation system. Exemplary applications include:

- Video games music, i.e., procedural generation of music based on the setting and gameplay tempo.
- Exercise music, e.g. cardio workouts, where the music's rhythm is synchronized to the users vitals and/or an exercise tempo scheme.
- Computer aided music composition, where the model would help human composers by "filling in the gaps" with candidate sequences of which the user would select the most appropriate, to augment their work or give them inspiration.
- Creating new music in the spirit or style of passed away musical performers and actors.

However, any attempt of creating a deep learning model for computer musical creativity is severely limited by the lack of a way to objectively evaluate the model. The most common approach is subjective evaluation of the results by a human, usually qualified, with musical training and experience. This not only limits evaluation of the final results, but also limits the selection and training of deep learning algorithms and can prove a severe bottleneck.

We decided to limit the scope to training models on one track/voice of polyphonic music, with an arbitrary sequence of notes or chords. We built the dataset from MIDI files due to their availability and ease of parsing.

In this paper we provide details of two experiments that we conducted, a multilayer Long-Short Term Memory (LSTM) Recurrent Neural Network (RNN) and feed-forward network, based on [3] and [4]; and a LSTM based Encoder-Decoder architecture as baseline models based on [5]; as well as outline how we aim to improve them and other experimental architecture we intend to use in future works.

The paper is organized as follows: in section II we give a brief overview of literature in the field of computer music generation; in section III we describe how we obtained the dataset, how it was processed and in subsection III-A the data representation we chose for the learning model. Then, in section IV we describe the deep learning architectures we used in the experiments, in section V we discuss and interpret

This work was partially financed by the Faculty of Computer Science and Engineering at the Sts. Cyril and Methodius University, Skopje, Macedonia. We also acknowledge the support of Microsoft Azure for Research through a grant providing resources for this work.

the obtained results, in section VI we outline how we intend to improve the current experiments and the architecture we plan to test out next, and finally in section VII, we conclude this paper.

II. Overview of Similar Approaches to Music Generation

This work was inspired by the legendary article by Andrej Karpathy - The Unreasonable Effectiveness of Recurrent Neural Networks [2]. In it Karpathy shows how a neural network consisting of 2 layers of 512 LSTM cells can generalize over several datasets of textual data: Paul Graham essay, Shakespeare plays, XML and Linux C/C++ source code. The resulting model can generate new sequences that are reasonably close to the source material and can almost fool a human, and in the case of C/C++ code some of the results compile. It is a character-level language model, i.e., it learns the texts letter by letter. Our first experiment was to adapt this model to the task of music generation.

Eck et al. in [6] describe the first application of deep learning to music generation. Due to the limited computational power at the time they defined a very simple experiment, generalization over 12 bar blues in written form, with a maximum of 8 notes per bar. They represented music in slices of note-time, i.e., one unit of time is $\frac{1}{8}$-th note long. This representation was later named as piano roll. The more complex of their two experiments consisted of learning in parallel a chord progression and a melody line, allowing them to generate both. Their architecture consists of 8 blocks of 2 LSTM cells, of which 4 are dedicated to learning melody the other 4 to learning the chord progression. The chords sections have output connections that influence the melody section but not vice versa. This work is the inspiration for most of the following research in the filed, including our own.

As an expansion to the previous work in [7] Eck et al. improve their model by using musically important temporal information from the pieces, allowing it to learn structure at different granularities. The said information is metrically significant and is provided to the network as time-delayed copies of the network inputs at intervals that depend on the piece's meter. For pieces where no meter information is provided, the authors developed a technique for algorithmic meter extraction. The architecture consists of parallel LSTM layer and standard feed-forward layer, which is supposed to speed-up local dependency discovery. We plan to investigate how and whether having metrical information embedded into the learning model impacts performance.

The architectures presented in [8] and [3] are specifically designed to tackle Bach chorales. A hybrid architecture was presented in [8] consisting of 2 LSTM subnets, of which one learns the sequence in standard order and the other in reverse order and a feed-forward subnet that learns the sequence in standard order. The three subnets are combined in a single feed-forward network. There is a copy of the architecture for each voice in the chorales. The music is generated by pseudo-Gibbs sampling procedure. They treat the 4 voices separately, as monophonic music generation problems, and combine them into a polyphonic melody. In contrast, we aim to generate mixture of chords/single notes from a single network output. To augment the dataset they transposed all the chorales to all transpositions that fit in the tonal range present in the corpus. We intend to do similar transposition schemes. The authors of [3] approached the same problem with a 3-layer LSTM network, to which they applied Grid-search hyper-parameter tuning extensively to find optimal network and parameter configurations. They used a very different data representation, encoding all the voices in a single sequence, ordered by descending pitch.

Boulanger-Lewandowski et al. in [9] attempted to combine the RNNs' ability to model sequences and Restricted Boltzmann Machines' (RBMs) ability to model probability distributions and apply them to music generation, in an architecture dubbed RNN-RBM. In this architecture the RBM part models note co-occurrence probability distributions at each time-step and the RNN part models the temporal sequences. We aim to achieve polyphony by modeling the probability of each output of the network independently of one another, even though realistically they are not, and try to avoid using a hybrid solution like RNN-RBN, due to it being difficult to train and computationally expensive.

In [4], another hybrid deep learning model, the LSTM layers try to learn the temporal dependencies in the music, followed by a set of feed forward layers that try to learn note co-occurrences (or the harmonic part). They used a piano roll very similar to our own, enriched with additional meta-data extracted from the pieces.The meta-data depends on the music being of very regular form, which our dataset is no, so we did not include any.

The work presented in [10] approaches music generation as a word-level language model, where each unit (composed of 1 to 4 bars of music) is treated as word. The words are embedded in a descriptor vector. A library of word (or units) is then created from the complete set of embeddings. A LSTM network learns sequences of the embeddings. Music is generated by first sequencing embeddings, which are then decoded to actual note sequences by selecting the most appropriate representative out of the library of units. We used character level embeddings in out second experiment.

The implementation presented in [11] is very close to the one of [2]. The authors transformed a set of folk songs in ABC notation into a modified version of the ABC notation that's more suitable for deep learning, that makes it very easy to unambiguously distinguish the various types of tokens, dubbed folk-rnn. On top of that they built a character-level model RNN deep learning model. Those tokens were encoded in one-hot vectors fed to 3 layer LSTM network, which outputs probability distribution over the unique token set. The authors of [12] used an Auto-encoder architecture to learn chord latent space embeddings. A sequence learner model generalizes over sequences of latent space embeddings. They presented a comparison of several types of chord representations. One of the compared representations is very similar to what we

used in our experiments. We discuss more in section VI.

III. MIDI Music Dataset

We created a dataset out of MIDI music files of classical guitar music, scraped from http://classicalguitarmidi.com. We chose MIDI files as the source format due to their ready availability and the variety of music available on the Internet in MIDI format. There is a number of formats that convert to MIDI, like Guitar Pro files. MIDI computer-human interfaces can allow us to extend the system to accept direct human input in a relatively straightforward way. A total of 4009 pieces were scraped, which we then filtered by number of midi tracks per file to pieces with only 1 track, resulting in a set of 2158 unique pieces from more than 70 authors (e.g., Aguado, Bach, Carcassi, Duarte, Guiliani, Mertz, Paganini and Segovia), while some files contain no author information.

The total length of all pieces is 336,878 seconds, the shortest piece is just 9 seconds, the longest 1,546 seconds, and the average length of pieces is 84 seconds. A histogram of max polyphony per piece can be seen in Figure 1. Most pieces fall between 4 and 6 notes playing at the same time, which was expected, however there are pieces where more than 6 notes are playing at the same time. This means that either more than one guitar track was encoded on the same MIDI track, or some pieces contain music played on different instruments and were mistakenly uploaded to the web site or were scraped by mistake. As seen in Figure 2, which shows tempo information from the dataset, most pieces fall in the 100 to 240 tempo range, which covers all classical tempos. Figure 3 shows the distribution of note lengths in multiples of quarter note, showing that the overwhelming majority of notes is of expected durations from $\frac{1}{16}$-ths to whole notes. However a significant number of notes are of very long duration, e.g., more than ten whole note durations, which is unusual for guitar, and is probably an instrument that can sustain a tone for longer durations, or an effect that the MIDI file author wanted to achieve and not a part of the original composition. A note range between MIDI pitch 32 (G#1) to 98 (D7) is present, which also suggests that the dataset contains parts with different instruments, as a 24 fret guitar can achieve the highest note E6 or MIDI pitch 88.

A. Data Representation for the Deep Learning Models

We chose to use piano roll representation for the data. A piano roll, in its most basic form, is a binary 2 dimensional matrix, where one dimension represents linear time, and the other the set of all possible notes, usually ordered by pitch. A non-zero value indicates a certain note played at a certain time. We used multi-hot encoding to allow for polyphony. It is the most commonly used data representation in this field, as seen in [4], [6]–[9], [13]. However we approach the piano roll transformation from a different perspective. In the other works selection and regularization steps were taken to obtain datasets consisting of 4/4 meter [4], [6]–[8] and the tempo was either ignored or used as a filtering parameter. The musical sequences were discretized according to note time, to the

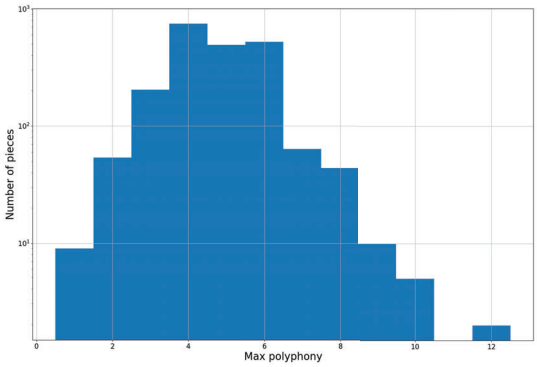


Fig. 1. Histogram of Max Polyphony.

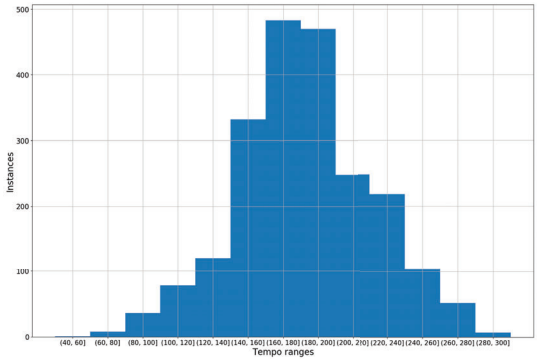


Fig. 2. Histogram of Tempos.

smallest possible note duration preset in the dataset, as seen in [6]–[8], or a preset minimal note duration, where smaller notes were simply ignored as seen in [7]. We decided to use real time instead of note time in discretizing the musical pieces, at a resolution of 16Hz (16 samples a second, i.e., 1s of music is represented as as a sequence of 16 elements). We chose this sampling frequency so we can represent $\frac{1}{16}$-th notes at a very fast tempo of 240 bpm, when the notes are properly aligned to $\frac{1}{16}$-ths of a second. This sampling technique effectively captures the vast majority of note sequences in the dataset, however it is not lossless. Notes that lie on several slices can

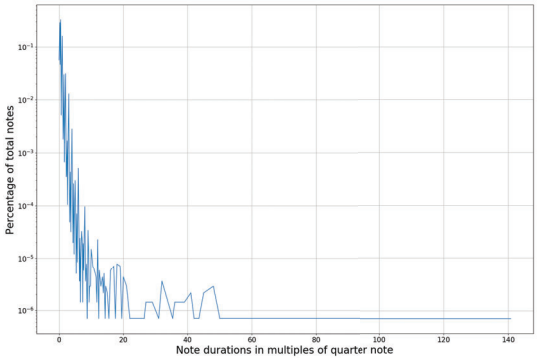


Fig. 3. Distribution of Note Durations.

be extended or contracted, or disappear entirely, depending on how far they span in the beginning and end slices. This leads to perceptive jitter/stutter when music is converted back to midi form for fast and dynamic pieces. Just like in the aforementioned works, our version of piano roll has no explicit note ending representation, which sometimes leads to several consecutive notes of the same pitch blending into one longer note. Since only a subset of the MIDI note pitches were present in the dataset, the range from MIDI pitch 32 to 98, we ignored the part of the midi specter that did not appear in any piece. MIDI files also contain velocity information for the notes played, which can be used in the piano roll representation, but we chose to ignore them as characteristics of performance and not composition, and instead focus on generating correct musical sequences.

IV. Experimental Architecture

We used the Keras deep learning library with the Tensorflow backend to implement our models. We conducted experiments with two types of architectures in mind. On the basis of the BachBot architecture presented in [3] and the architecture presented in [4], we tried 3 variations of a sequence learner:

- 2 recurrent layers with 256 LSTM cells with 0.2 dropout followed by fully connected output layer
- 3 recurrent layers with 1024 LSTM cells with 0.4 dropout followed by a fully connected output layer. Very close to the implementation in [3].
- 3 recurrent layers with 1024 LSTM cells with 0.4 dropout followed by 2 fully connected feed-forward layer, the second being an output layer. The second feed-forward fully connected layer was inspired by [4].

The activation function of the output layer is sigmoid, the model was trained with the ADAM optimizer using binary cross-entropy loss function. In Keras parlance the models were not stateful, relying on the Truncated Back-Propagation Through Time (TBPTT) to provide the network with the ability to learn sequences. We tried values between 16 and 64 timesteps for TBPTT (or between 1 and 4 seconds of look-back). The more timesteps the more context the model has, but it also becomes both harder to train, needs more training time to minimize training loss and requires more resources. We chose 16 timesteps as the value for TBPTT timesteps.

The other architecture we tried is inspired by the one described in [5], a Encoder-Decoder technique for sequence to sequence learning, where we limited the output sequence to be of length 1. We tried the following configuration:

- An encoder of one LSTM layer of 256 cells, a decoder of one LSTM layer of LSTM cells, with a fully connected output layer with sigmoidal activation.

This model was also trained with the ADAM optimizer using binary cross-entropy loss function. Each of the outputs from the final layer is treated separately, as if its activation probability is statistically independent from the rest, even though that is not strictly true. An output is considered active if the sigmoid activation value is greater than 0.5. This allow us to have multiple outputs activated at each time-step, allowing for polyphony, just like in the "bilinear model" defined in [12].

V. Results and Discussion

Training was done on a desktop PC equipped with an Intel i7 processor, 32GB of RAM and a GTX 750Ti NVidia GPU with 2GB of VRAM. All models were trained up to 100 epochs or until they did not improve on the loss function for more than 5 epochs. We used a small subset of 5 pieces from the whole dataset to do a baseline comparison between the model configurations. The pieces were chosen by familiarity to the experimenters to give us the ability to spot elements and patterns in the generated data. They were transformed to piano roll representation at a sampling frequency of 16Hz and all of them were concatenated into a single sequence with no special symbols for start or end of subsequence. With an input sequence length of 2 seconds, or 32 time slices, we got 26336 samples for training. No data was withheld for validation. The longest training time per epoch was 12 minutes for the most complex variant of the first architecture. Generation is done in a iterative fashion, i.e. a feed-forward generation strategy. We start by giving the models a primer, a 5 second sequence from the start of each of the pieces. The network is fed from the primer one sample at a time, discarding the predictions until the original primer boundary is reached. Afterwards the predictions are concatenated to the primer. We generate sequences of fixed length of 60 seconds. An overview of the results is given in Table I. An example generated sequence visualized in it's piano roll representation is shown in Figure 4. This is the best case result we have achieved, where the model repeats a subsequence several seconds in length. It was generated by the simplest model.

TABLE I
Overview of the generation results by architecture

Architecture	Result
2 layer LSTM	Up to 2 seconds of valid sequence of polyphonic music followed by either a constant polyphonic tone, a repeating pattern or silence
3 layer LSTM	Silence
3 layer LSTM and 2 feed-forward layers	Silence
Encoder-Decoder Architecture	A constant tone of several notes for the whole duration of the generated sequence

One limitation that affects all models is the character of the data. The music is not structured or very similar from piece to piece. That coupled with the polyphony make it hard for the model to generalize over the dataset.

The 2 layer LSTM model is limited in expressiveness due to the small network capacity and over-fits the data. It learns small sequences extremely well and tends to get stuck in loops when they get activated from the primer. In other cases it doesnt get activated at all and results in silence.

The 3 layer LSTM model on the other hand, is too expressive and does not have enough quality data to learn, so it never gets activated at all. This was clear from the training process,

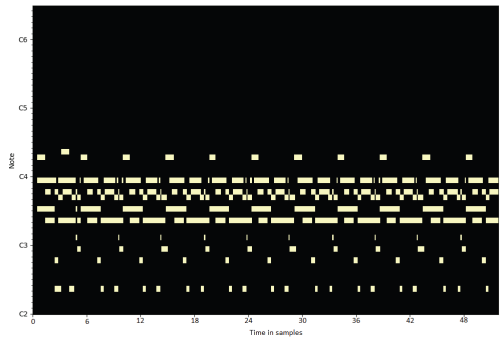


Fig. 4. Exemplary generated sequence.

where the 3 layer LSTM model plateaued much more quickly and at a higher loss value than the 2 layer LSTM model. The model with 3 LSTM and 2 feed forward layer had the same issues as the regular 3 layer LSTM model.

The Encoder-Decoder model had the same problems as the 2 layer LSTM model, however it has a more limited arbitrary memory capability when compare to the pure LSTM model. This leads to over-fitting even shorter sequences and therefore gets stuck repeating the same note. According to [12], the way we structured the outputs, with multi-hot encoding and independence assumption of the output probabilities, the model is expected to perform poorly compared to one-hot encoding or more advanced encoding schemes.

To get a time estimate for training we ran the simplest variant of the first architecture with a subset of 300 randomly chosen pieces from the dataset. An epoch took 52 minutes to train, which would take the model several days to be fully trained. This proves impractical for training models on the complete dataset.

VI. Future Work

We suggest the following ways in which the results could be improved:

- Adjust the parameters of the models, the layer sizes, activations, dropout values, input sequence sizes. Investigate applicability of Grid Search, as used in [3], or other optimization algorithms and techniques [14].
- Augment dataset with transpositions to every key present in the dataset. This is a technique used in [8], [10], and [13]. It will result in providing more samples that are musically correct, as well as sequences with the same relative pitch movement, allowing the models to better capture patterns across the dataset.
- Create a more efficient training method as well as an efficient batching method so that we can train the models on much larger subsets of the dataset.
- Try a different data representation as seen in [3] or [13].
- Try an architecture that learns to embed the input data into a latent space, i.e. chord embedding, and then have a multi-layer LSTM model learn embedding sequences similar to [12].
- Transform the problem to univariate predictions to greatly simplify the learning difficulty, like in [13].

VII. Conclusion

The work so far did not fulfill the set goal of generating a 60 second long sequence of polyphonic music. We discussed our interpretation of the limitations of the models we used in section V. They need further refinement before being able to generate actual musical sequences. We gave a brief overview of how the results could be improved. In section V we also noted that training on the complete dataset is not practical on the current hardware. However thanks to Microsoft Corporation, we have received a grant for Azure Cloud for use with machine learning, which we intent to leverage after refining the architecture.

References

[1] F. Ghedini, F. Pachet, and P. Roy, "Creating music and texts with flow machines," *Multidisciplinary Contributions to the Science of Creative Thinking*, pp. 325–343, 2015.

[2] Andrej Karpathy, "The Unreasonable Effectiveness of Recurrent Neural Networks," 2015. [Online]. Available: https://karpathy.github.io/2015/05/21/rnn-effectiveness/

[3] F. Liang, M. Gotham, M. Johnson, and J. Shotton, "Automatic Stylistic Composition of Bach Chorales with Deep LSTM," *Proceedings of the 18th International Society for Music Information Retrieval Conference*, pp. 449–456, 2017.

[4] D. D. Johnson, "Generating polyphonic music using tied parallel networks," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 10198 LNCS, pp. 128–143, 2017.

[5] K. Cho, B. van Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation," 2014.

[6] D. Eck and J. Schmidhuber, "A First Look at Music Composition using LSTM Recurrent Neural Networks," *Idsia*, pp. 1–11, 2002.

[7] D. Eck and J. Lapalme, "Learning musical structure directly from sequences of music," *University of Montreal, Department of Computer . . .*, no. 1983, pp. 1–12, 2008.

[8] G. Hadjeres, F. Pachet, and F. Nielsen, "DeepBach: a Steerable Model for Bach Chorales Generation," 2016.

[9] N. Boulanger-Lewandowski, "Modeling High-Dimensional Audio Sequences with Recurrent Neural Networks," p. 145, 2014.

[10] M. Bretan, G. Weinberg, and L. Heck, "A Unit Selection Methodology for Music Generation Using Deep Neural Networks," pp. 1–13, 2016.

[11] B. L. Sturm, J. F. Santos, and I. Korshunova, "Folk Music Style Modelling By Recurrent Neural Networks With Long Short Term Memory Units," *Ismir*, 2015.

[12] S. Madjiheurem, L. Qu, and C. Walder, "Chord2Vec: Learning Musical Chord Embeddings," no. Nips, 2016.

[13] C. Walder, "Modelling Symbolic Music: Beyond the Piano Roll," 2016.

[14] P. Lameski, E. Zdravevski, R. Mingov, and A. Kulakov, "Svm parameter tuning with grid search and its impact on reduction of model overfitting," in *Rough Sets, Fuzzy Sets, Data Mining, and Granular Computing*. Springer, 2015, pp. 464–474.

Deep Learning: The future of chemoinformatics and drug development

Ljubinka Sandjakoska
Faculty of Computer Science and Engineering
UIST St Paul the Apostle
Ohrid, Macedonia
ljubinka.gjergjeska@uist.edu.mk

Ana Madevska Bogdanova
Faculty of Computer Science and Engineering
SS Cyril and Methodius University
Skopje, Macedonia
ana.madevska.bogdanova@finki.ukim.mk

***Abstract*—The recent growth of using machine learning methods in solving complex problems in life sciences highlights the need of its refinement. Deep learning, as new and perspective sub-field of machine learning, has a big contribution to the refinement of the existing machine learning approaches. Deep learning concept dramatically improves the state-of-the-art in the field. Improving is done by solving the conventional techniques' problem of limitation to processing the raw data. In this paper, we will give an answer to the question – why deep learning? The key concepts will be given also. We emphasize the usage of deep learning in chemoinformatics and pharmaceutical research and its advantage over previous machine learning methods and standard methods for molecular dynamics simulations and drug development. A matter of importance of this paper is to summarize the existing approaches in applications of deep learning in chemoinformatics and to identify a new perspective as a future research problems with views that are needed to be further tested.**

Keywords—machine learning, deep learning, chemoinformatics, drug development

I. INTRODUCTION

Nowadays, in the area of big data, extracting relevant information is more than necessary and very challenging, especially in solving non-linear problems of complex systems in real times. The concept of machine learning is not so new, but its advanced techniques and tools are promising in dealing with the newest data issues. Due to the enormous growth in chemical databases deploying of improved and optimized methods drawn from machine learning is more than necessary. Many machine learning methods have already been used, over the two past decades, in the field of chemoinformatics [1]. The wide usage is grounded on the theoretical framework that machine learning provides for the discovery and prioritization of bioactive compound with desired pharmaceutical effects and their optimization as drug-like leads. The most used machine learning method in computational chemistry and computer-aided drug design are artificial neural networks [2]. Since deep learning refers to artificial neural networks, the continuum in application is logical.

The principles of deep learning, given in [3], start the modern way of solving problems by introducing a learning algorithm for supervised deep feed-forward multilayer perceptron, which units had polynomial activation functions combining additions and multiplications in Kolmogorov-Gabor polynomials. The first description of deep neural network [4], with eight layers trained by the "group method of data handling", open the era of learning the creation of hierarchical, distributed, internal representations of incoming data. A lot of problems can be identified in the fundamental learning in neural networks. Here will be discussed only a few problems that we think have influence to the domain. The problem of limitation to processing the raw data is overcome by using deep learning. Deep learning, as new and perspective sub-field of machine learning, also try to deal with the problems that arise from amount, heterogeneity and complexity of the data. It is based on learning several levels of representations, corresponding to a hierarchy of features, where higher-level concepts are defined from lower-level ones, and the same lower-level concepts can help to define many higher-level concepts [5]. Deep learning is characterized by adaptability of modeling abstraction over multiple levels. Most of the standard learning methods use shallow-structured learning architectures. The using of this architectures required careful engineering and considerable domain expertise to design a feature extractor that transformed the raw data into a suitable internal representation or feature vector from which the learning subsystem could detect or classify patterns in the input [6]. Deep learning refers to machine learning techniques that deploy supervised and/or unsupervised strategies to automatically learn hierarchical representations in deep architectures. Deep learning methods effectively exploit complex, compositional nonlinear functions, learn distributed and hierarchical feature representations, and make effective use of both labeled and unlabeled data [7]. In order to avoid over-fitting problems conventional learning architectures work with a few parameters with high preference unlike to the deep learning algorithms that include a large number of hidden neurons, often resulting in large number of free parameters which require specific training techniques. Namely, in deep learning the techniques that are used for training attempt to simulate the network to learn the most general of the possible weight combinations, using advanced regularizations such as "early termination" [8] and "dropout" [9]. In this context should be mentioned that deep learning offer improvement in using enhanced activation functions [10], such as parallel computing in on graphics processing units [11]. These

advantages of deep learning architectures over shallow-structured architectures is key and promising for solving complex chemical tasks. The new concept of large-scale deep learning include two aspects: large volume of data and large models. In chemoinformatics the both aspects are considered. The first aspect of large volume of data is more characteristic for drug discovery research, where deep learning should deal with a problems such as listing a candidate compound for drug, predicting the drug-target relationship, predicting protein secretion capabilities etc. The second aspect of large models usually is applied in standard chemical simulations, such as predicting molecular properties and relationships between them or analyzing particle accelerator data.

In the following paper, first we will present briefly the key concepts and categorization of deep learning. In the next session we provide an overview of the recent studies that use deep learning in the field of life sciences or more precisely in chemistry and pharmaceutical discovery. In the third part towards deep learning on chemoinformatics data also will be presented views that are needed to be further tested. In the last section the concluding remarks are given and we will identify a new perspectives as a future research problems.

II. DEEP LEARNING

A. Key concepts

Deep learning, stated as a particular kind of machine learning that achieves great power and flexibility by learning to represent the world as nested hierarchy of concepts, with each concept defined in relation to simpler concepts, and more abstract representations computed in terms of less abstract ones [12], give the directions of its advanced applications over standard machine learning techniques. There are several key points of choosing deep learning solutions. The first key point is determined by the question - why deep learning? The answer is related with its increscent of the performance as the scale of data increases. Namely, deep learning understand perfectly big amount of data, but this implies hardware dependencies. In fact, deep learning heavily depend on high-end machines. Since, deep learning algorithms perform a large amount of matrix multiplication operations, the process of its optimization should be done using the GPUs. This is one of the differences between standard machine learning and deep learning. Physics seems to dictate that any future efficient computational hardware will have to be brain-like, with many compactly placed processors in 3-dimensional space, sparsely connected by many short and few long wires, to minimize total connection cost (even if the "wires" are actually light beams) [13]. Also, identification and extraction of features in reducing the complexity of the data, such as a hand-coding of the feature (usually done by the domain expert in order to make the patterns more visible for the machine learning algorithms) is avoided in deep learning. The task of creating new feature extractor for every problem is not characteristics of deep learning. Issues such as overfitting and computation time are challenges to deep learning. The deep learning is capable to solve these issues because of the adding layers of abstraction which are responsible for modeling the rare training data dependences. Different methods for regularization [14] can be also used to overcome the problem of overfitting in standard artificial neural network. As a solution of this problem sometimes deep learning offers cropping and rotating as an augmentation method for increasing the size of training sets. This actually, depicts one of the refinements and the major step ahead of deep learning compared to the conventional machine learning.

In order to present the key concepts of deep learning let consider deep feed-forward neural network architecture, which is most basic form of fully connected network, with one input layer, two hidden layers and one output layer. (Fig.1)

During a forward pass, the pre-activation $z_j^{(1)}$, which is a linear combination of the preceding layer's output values, is calculated for each non-bias neuron in the first hidden layer. Next, a nonlinear activation function $az_j^{(1)}$ is applied to compute the output values of $h_j^{(1)}$. Once the neurons' activations for one layer are calculated, they serve as input values for the following layer until one finally arrives at the network's output values o_l. Then, the output values are evaluated against the true output values t_l using an error-function $C(o_l)$. To resolve the indirect dependency of the error function on the network's weights, the error is back-propagated through the network using the chain rule for derivatives. Other training techniques may be applied.

B. Categorization of Deep Learning

In the literature several approaches of categorization of deep learning are presented. In general the classes of deep learning depend on how the architecture and techniques are intended for use [5]. Also, different categories of deep learning architectures can be viewed according to its interpretation - in terms of universal approximation theorem or probabilistic inference. According to the intention for use there are three classes of networks: deep networks for unsupervised or generative learning; deep networks for supervised learning and hybrid deep networks. (Tab.1)

Deep networks for unsupervised learning can be either generative or non-generative in nature. Its main characteristic is meaningful generation of samples by sampling from a network. Deep networks for supervised learning usually are known as discriminative deep networks since the capability of providing discriminative power for classification purposes. The third category refers to deep architecture that either comprises the use of both generative and discriminative model components. Hybrid architecture exploits the generative component as a help tool in discrimination.

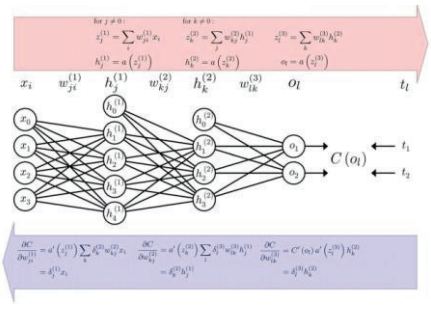


Fig.1. Deep feed-forward neural network architecture

TABLE I. CATEGORIZATION OF DEEP LEARNING

Deep networks for unsupervised or generative learning	Energy-based deep models Deep Belief Networks (DBN)
	Deep Boltzmann machine (DBM)
	Restricted Boltzmann Machine (RBM)
	Sum-Product Network (SPN)
	Recurrent Neural Networks (RNN)
	Dynamic or temporally recursive generative models
Deep networks for supervised learning	Deep-structure Conditional Random Field (CRF)
	Deep-Stacking network (DSN) tensor variant
	Deep-Stacking network (DSN) kernel version
	Convolutional Neural Network (CNN) -Time-Delay Neural Network (TDNN) -Hierarchical Temporal Memory (HTM)
Hybrid deep networks	Architectures with generative and discriminative model components

III. TOWARDS DEEP LEARNING ON CHEMOINFORMATICS DATA

Life science community introduce the concepts of deep learning in a variety of problems such as predicting of: drug-target interactions, molecular activity, toxicity, reactivity, protein contact; also virtual screening, quantitative structure–activity relationship (QSAR), proteome mining, ADMET (absorption, distribution, metabolism, excretion and toxicity) profiling properties etc. Usually deep learning techniques are adopted as complementary to standard software for molecular simulations, but today tendency go in line by the twist in technology in order to increase the efficiency of drug development Deep architectures require careful analysis and not partial but complete definition regard to respected domain. The wide application of deep learning architectures in chemoinformatics is related to several advantages, specific for chemoinformatics. The first advantage refers to its capability for multi-task learning [15], which allows multi-label information, therefore can utilize the relation between the targets and allows to share hidden unit representations among prediction tasks. The second advantage is automatically construction of complex features [16] because of the ability to provide hierarchical representation of a compound, where the high levels represent more complex concepts [17].

Although neural networks have a long history in drug discovery and design, the risks of over-training and "black-box" phenomena motivate the scientists for its substitution with some alternative techniques such as supported vector machines [18] [19], naïve Bayesian classifier, decision trees [20]. The alternatives for conventional neural network not give the wanted results and the novel concepts of deep neural network came in the focus. Namely, in [21] a comparative analysis is done, between deep learning to the following machine learning methods used for drug-target prediction: logistic regression, k-nearest neighbor, binary kernel discrimination and supported vector machines. Also, re-implementation of some commercial products is done in order to be depicted the advantages of deep learning with the respect to the area under ROC curve averaged over the targets. The results in the paper [21] show that deep learning outperformed all other methods and commercial products. Furthermore, in solving the task of toxicity prediction [22], the prediction of arbitrary number of toxicological effects at the same time without the need to train the single classifier for each one is allowed by deep learning architecture. Boosting the performance on task with few training examples is because of the exploiting representations learned in the multi-task environment. The advances of using deep learning, as an effective method in chemoinformatics, are approved in predictive modeling of compound-protein interactions from descriptors of ligands and proteins. In addition, the authors of [19] detect the incensement of the calculation time and computer memory in using supported vector machines for chemical genomics-based virtual screening (CGBVS)which predicts compound-protein interactions. The proposed deep learning approach [19], emphasizes that deep learning does not require learning all input data at once because the network can be trained with small mini-batches. The experiments show that deep architecture outperforms the original CGBVS with a quarter million compound-protein interactions and cross-validation results which show accuracy up to 98.2 % (σ<0.01) with 4 million compound-protein interactions. The nature of biological data, which are characterized by nonlinearity and imbalance, noise contaminations and diversity motivated the authors of [23] to apply deep architecture in compound-protein interaction prediction, by tuning the hyperparameters, in order to boost the prediction. Moreover, deep learning approaches are proved as good prediction tool of drug release polylactide-coglycolide microspheres in drug formulation process. [24] Recognition of pharmacological properties of multiple drugs across different biological systems and conditions is also done by deep learning neural net trained on transcriptomic data [25]. Another successful deployment of deep architectures is in predicting aqueous solubility of drug-like molecules. Namely the authors of [26] conclude that the performance of the deep learning methods matches or exceeds the performance of other state-of-the-art methods according to several evaluation metrics. Also, this implementation depicts

one specific advantage of deep learning, such as its reliability to identify only the minimal suitable molecular descriptors. This is possible because suitable representations are learned automatically from the data. In addition, prediction models for drug induced liver injury were developed using DL architectures [27]. Beside the prediction, which is in the focus, another benefits from the deep analysis is identification of important molecular features that are related to drug induced liver injury and the risk in humans. The prioritization of the experiments during the drug discovery process is advanced using deep learning for QSAR [28].The goal is reducing the experimental work that needs to be done. In [28, 29] ADME profiling properties are obtained using deep learning, also biological targets [28,30] and transporters identification [28].

In the resent years, the advancement of reinforcement learning (RL) and deep learning result with synergy of the both approaches. The integration of reinforcement learning algorithms with deep architectures give new possibilities for solving standard problems characteristic for both concepts. In [31] is given an overview of deep reinforcement learning with presenting two possibilities for its integration: first is in approximating RL functions using deep NN architectures and the second is in training deep NN using RL. In [32] are given comparative test results, of each DL technique when applied to the biological data. The results are promising and show a great opportunities of employing deep RL in drug development.

IV. CONCLUSION

Deep learning is more than a trend and it is expected to become more important than it is today. The advantages mention in this paper give possibilities for improvement in pharmaceutical research and drug development. The improvement lead to low costs and decreasing the number of in-vivo experiments. Research community in the area of implementation of machine learning in life sciences believe that the future role of deep learning may not just be only a high-performance prediction tool, but perhaps as a hypothesis generation device as well.

There are some views that are needed to be further tested in order to get answer to the questions such as: is there some method for preparing the complex data to faster deep training/testing in using parallelism like GPUs; how to find appropriate architecture; is it possible to avoid the phenomena of “black-box”; or is it possible to find solution of the problem of vanishing gradients etc.

We believe that some of the problems of standard machine learning methods, such as Naïve Bayesian and support vector machine when are used for prediction of blood brain barrier penetration, skin permeability, mutagenicity, counterfeit drug detection, docking, can be solved by deep learning. This open new prospective for applying deep architectures in chemoinformatics, as a future research problems.

REFERENCES

[1] J.B. Mitchell, “Machine learning methods in chemoinformatics”,Wiley Interdisc Rev Comput Mol Sci 4, pp. 468-481.

[2] J. Zupan, and J. Gasteiger “Neural Networks in chemistry and drug design”, Wiley-VCH, Weinheim, 1999.

[3] A. G. Ivakhnenko and V. G. Lapa, “Cybernetic Predicting Devices”, CCM Information Corporation, 1965.

[4] A. G. Ivakhnenko, “Polynomial theory of complex systems”. IEEE Transactions on Systems, Man and Cybernetics, 4, 1971, pp. 364–378.

[5] L. Deng and D. Yu, “Deep learning: methods and applications”, Microsoft Research, Now publishers, 2014.

[6] Y. LeCun, Y. Bengio and G. Hinton, “Deep learning” Nature, 521, May 2015.

[7] Y. Marc'Aurelio Ranzato, L. Boureau, and Y. LeCun, “Sparse feature learning for deep belief networks”, in Proc. Adv. Neural Inf. Process. Syst., vol. 20, 2007, pp. 1185-1192.

[8] A. Tropsha, “Best Practices for QSAR Model Development, Validation, and Exploitation”, Mol. Inf. 2010, 29, pp.476– 488.

[9] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov, “Improving neural networks by preventing co-adaptation of feature detectors”, Cornell University Library 2012, arXiv:1207.0580.

[10] F. Agostinelli, M. Hoffman, P. Sadowski and P. Baldi, “Learning activation functions to improve deep neural networks”, Cornell University Library 2014, arXiv: 1412.6830.

[11] R. Raina, A. Madhavan, and A. Y. Ng, “Large-scale deep unsupervised learning using graphics processors”, Proc. 26th Annual Int. Conf. Mach. Learn. ICML 2009, pp. 873-880.

[12] I. Goodfellow, Y. Bengio and A. Courville, “Deep Learning”, MIT Press, 2016.

[13] J. Schmidhuber, “Deep Learning”. Scholarpedia, 10(11):32832, doi:10.4249/scholarpedia.32832 2015.

[14] Y. Bengio, N. Boulanger-Lewandowski, andR. Pascanu, “Advances in optimizing recurrent networks”. IEEE International Conference on Acoustics, Speech and Signal Processing, 2013, pp. 8624-8628.

[15] R. Caruana, “Multitask learning,” Machine Learning, vol. 28, no. 1, 1997, pp. 41–75.

[16] Y. Bengio, A. Courville, and P. Vincent, “Representation learning: A review and new perspectives,” IEEETrans Pattern Anal Mach Intell, Feb 2013.

[17] Y. Bengio, “Deep learning of representations: Looking forward,” in Proceedings of the First International Conference on Statistical Language and Speech Processing, Berlin, Heidelberg, pp. 1–37, Springer-Verlag, 2013.

[18] E. Byvatov, U. Fechner,J.Sadowski and G. Schneider, “Comparison of support vector machine and artificial neural network systems for drug/nondrug classification.” J Chem Inf Comput Sci. 2003 Nov-Dec;43(6), pp. 1882-1889.

[19] M. Hamanaka,K. Taneishi,H. Iwata, J. Ye, J.Pei, J. Hou,and Y. Okuno, “CGBVS–DNN: Prediction of compound–protein Interactionsbased on deep learning”, Mol Inf 36, 2016, pp. 1–10.

[20] J.B. Mitchell, “Machine learning methods in chemoinformatics”, Wiley Interdisc Rev Comput Mol Sci 4, 2014, pp. 468–481.

[21] T. Unterthiner, A. Mayr, G. Klambauer, M. Steijaert, J. K. Wegner, H. Ceulemans, and S. Hochreiter, “Deep Learning for drug target prediction”,Conference Neural Information Processing Systems Foundation, Workshop on Representation and Learning Methods for Complex Outputs, Montreal, Canada, December, 2014.

[22] A. Mayr,G. Klambauer, T.Unterthiner, and S Hochreiter, “DeepTox: Toxicity prediction using deep learning”, Front. Environ. Sci. 3:80, 2016, doi: 10.3389/fenvs.2015.00080.

[23] K. Tian, M. Shao, Y. Wang, J. Guan, and S. Zhou. “Boosting compound-protein interaction prediction by deep learning”, Methods, Volume 110, November, 2016, pp. 64-72 http://dx.doi.org/10.1016/j.ymeth.2016.06.024

[24] H.M. Zawbaa, J. Szlek,C. Grosan, R.Jachowicz, andA.Mendyk, “Computational intelligence modeling of the macromoleculesrelease from plga microspheres-focus on feature selection” PLoSONE 11:e0157610, 2016.

[25] A. Aliper, S. Plis,A Artemov,A. Ulloa, P. Mamoshina, and A. Zhavoronkov, “Deep learning applications for predicting pharmacologicalproperties of drugs and drug repurposing using transcriptomicdata”, Mol Pharm 13, 2016 pp. 2524–2530.

[26] A. Lusci, G. Pollastri, and P. Baldi,“Deep architectures and deep learning in chemoinformatics: the prediction of aqueous solubility for drug-like molecules”, J Chem Inf Model., 53(7), 2013, pp.1563–1575.

[27] Y. Xu, Z. Dai, F. Chen, S. Gao, J.Pei, and L. Lai,"Deep learning for drug-induced liver injury. J Chem Inf Model. 2015;55(10), pp. 2085–2093.
[28] J. Ma, R.P. Sheridan, A. Liaw, G.E. Dahl, and V.Svetnik, "Deep neural nets as a method for quantitative structure-activity relationships", J Chem Inf Model. 55(2),2015, pp. 263–274.
[29] T.B. Hughes, G.P. Miller, and S.J. Swamidass,"Modeling Epoxidation of Drug-like Molecules with a Deep Machine Learning Network" ACS Cent Sci. 2015;1(4), pp. 168–180.
[30] T. Unterthiner, A. Mayr, G. Klambauer, and S. HochreiterToxicity prediction using deep learning. https://arxiv.org/pdf/1503.01445.pdf.
[31] Y. Li,"Deep reinforcement learning: An overview", 2017, https://arxiv.org/abs/1701.07274.
[32] M. Mahmud, M. S. Kaiser, A. Hussain, and S. Vassanelli, "Applications of Deep Learning and Reinforcement Learningto Biological Data", https://pdfs.semanticscholar.org/0ea1/491b8d12a41b4c257ecabca5ae7cce05835f.pdf

An Off-Lattice HP Model for Protein folding with three componential evaluation function

Ivan Todorin
Faculty of engineering
South-West University "Neofit Rilski"
Blagoevgrad, Bulgaria
ivan_todorin@swu.bg

Ivan Trenchev
Faculty of engineering
South-West University "Neofit Rilski"
Blagoevgrad, Bulgaria
trenchev@swu.bg

***Abstract*— Biological activity of a protein is determined by its 3D structure. If we can predict the 3D structure of proteins, in case we know the primary structure, than only proteins with expected properties have to be synthesized for drug design. We modify off-lattice HP Model such as not to have constant constraint for size and evaluation function equal to count of contacts, but to use variable constraint for size and evaluation function formed by multiplication of subfunctions for three energy minimizing components - secondary structure, hydrophobic contacts and S-S bridges.**

Keywords: HP folding prediction, Drug design, Heuristic algorithms, Off-lattice models

I. INTRODUCTION

The 3D structure of proteins is the major factor that determines their biological activity. The synthesis of new proteins and the crystallographic analysis of their 3D structure is very slow and very expensive process. If we can predict the 3D structure of many proteins, than only proteins with expected properties have to be synthesized. That will increase the number of known structures in the databases for proteins, and they can be used for drug design. The prediction of the 3D structure of proteins, if we know only the primary structure – the amino-acid sequence, is a protein folding problem. The reason for this process of folding in water environment is the interaction between water molecules and between amino-acids and water molecules. As water molecule has higher polarity than amino-acids, there is a minimum of energy when the protein is folded, not to spoil water to water interconnections [1, 2]. The way of folding is determined by the polarity or the hydrophobicity of different amino-acids, so the 3D structure with minimum energy is the real case. There is less energy when more hydrophobic (H) amino-acids are in contact in the core of the folded 3D structure and more polar (P) amino-acids are in contact with water. As we know the amino-acid sequence and the hydrophobicity of every amino-acid, we can predict the 3D structure – this method is called HP folding. It is needed to calculate the impact of all contacts between amino acids, using chosen scoring function, for many possible 3D structures [3]-[5].

II. MODEL DESCRIPTION

In this chapter we will look at our model of predicting the three-dimensional form of proteins, a variant of the HP model, in which:

1. Alpha carbon atoms are positioned with three-dimensional coordinates starting at (0.0.0) and (0.0.1), following the amino acid numbers in the amino acid sequence, randomly choosing the position of each subsequent alpha carbon atom of possible peptide chain twists taken over 10 degrees of accuracy. The distance between the alpha carbon atoms of adjacent amino acids is a unit of normalized distance, and the form of the protein is simplified only by these coordinates, and placement of the amino acids will mean the localization of these atoms defining the spinal structure of the peptide chain. Use the random number generator of the C ++ programming environment by taking the remainder of the fission to 72 and let it be the number r, a $r_1 = r / 12$ and $r_2 = r \bmod 12$, then:

$$x_i=x_i\text{-}1+\sin(90\text{-}30r_1),$$

$$y_i=y_i\text{-}1+\sin(30r_2)\cos(90\text{-}30r_1),$$

$$z_i=z_i\text{-}1+\cos(30r2)\cos(90\text{-}30r_1)$$

where (x_i, y_i, z_i) where (x_i, y_i, z_i) has the coordinates of the newly introduced amino acid, and $(x_{i-1}, y_{i-1}, z_{i-1})$, $(x_{i-1}, y_{i-1}, z_{i-1})$ and (xi-2, yi-2, zi-2) are the coordinates of the previous amino acids.

2. The following protein-based limitations are observed when placing amino acids:

- cannot be located closer than 0,7 standard quadratic distance in each of the three orthogonal directions to another amino acid:

$$(x_i-x_j)^2+(y_i-y_j)^2+(z_i-z_j)^2>0,7$$

as (x_i, y_i, z_i) и (x_j, y_j, z_j) are the coordinates of the two amino acids

- can not be located further than the center of the molecule from a distance set by the variable extension parameter in each of the three orthogonal directions:

$$(x_i-x_c)^2+(y_i-y_c)^2+(z_i-z_c)^2<s.l$$

as (x_i, y_i, z_i) are the coordinates of the inserted amino acid, (x_c, y_c, z_c) are the coordinates of the center of the molecule built to date, l is the length of the peptide chain, s is the variable of the allowed extension

- In case of an unacceptable position, a new random angle of twist is set, so that all 36 options are exhausted. If the next amino acid can't be placed, the build of the current conformation is

terminated, the failed conformation counter is updated and the next one is started. This is accomplished in a number of embodiments implemented in another cycle with a number of embodiments, the percentage of failed conformations [5], [6] (possible three-dimensional shapes) in the execution of the internal loop determines the new value of the allowable extension variable for the next iteration of the outer loop.

3. In order to form the value of the estimation function, it is necessary to determine which amino acids are in contact - since the side radicals may be between the two regions of the peptide chain, then the proximity of their alpha carbon atoms to the twice the normalized distance between adjacent alpha carbon atoms - so another amino acid can not be placed between them. The proximity check is defined by the formula:

$$(x_i-x_j)^2+(y_i-y_j)^2+(z_i-z_j)^2<4$$

as (x_i, y_i, z_i) и (x_j, y_j, z_j) are the coordinates of the two amino acids

The evaluation function is a work of three sub - functions that take into account the degree of construction of the following:

1) the secondary structure - the neighborhood of amino acids and the formation of hydrogen bonds in the spinal structure - the value is their number:

$$SF = \sum_{(i,\, j\in C)} a_{i,j}$$

where SF is the value of the evaluation sub-function, a_{ij} are the amino acids with numbers i and j, and C is the plurality of the pairs in the contact.

2) hydrophobic amino acid contacts - the value is the sum of the hydrophobicity values of the adjacent hydrophobic amino acids:

$$SF = \sum_{(i,\, j\in C)} (H_i + H_j),$$

where SF is the value of the evaluation sub-function, Hi and Hj are the hydrophobicity values of the amino acids with numbers i and j, and C is the plurality of the pairs in contact.

3) disulphide bridges - the value is the plus plus one, as there may be no such:

$$SF = 1 + \sum_{(i,\, j\in C)} a_{i,j},$$

where SF is the value of the evaluation sub-function, a_{ij} are the amino acids with numbers i and j, and C is a plurality of pairs of amino acids cysteine in contact.

Compared to the classic HP model, it has a higher computational complexity, but gives a more realistic build of the peptide chain in space - amino acid positioning follows [5. 10] the natural principle of peptide linking, similar to the curve motion in chess but three-dimensional. It is thus possible to build three-dimensional shapes (conformations) that resemble the real forms in nature.

.

III. SEQUENCE OF WORK

The primary structure data is read from a text file, the letter sequence is taken from the PDB from the FASTA file format - the amino acid sequence for which the three-dimensional shape is sought. Hydrophobicity values of amino acids are sequentially recorded in an array and their number is stored.

The initial parameters for the coefficient of admissible spread, the step with which it will change, and the percentage of failed conformations above which will change will be entered from the keyboard. Better results are obtained when the coefficient is high - more than 90% fail, and successful conformations with a smaller coefficient of admissible stretching are more compact and thus more likely to have a higher value of the evaluation function. This is what will be explored - with what parameters you get better results. In the course of calculation, it starts with the most compact forms and gradually increases the space for the placement of the molecule.

The first two amino acids occupy the coordinates (0,0,0) and (0,0,1), after which each of the following is located:

- a number is taken from the pseudo-random number generator, which determines the angle of twist of the peptide bond, checks the flag at that torsion angle, and if it is not true, calculates the new position in the space

- the proximity limit is checked to match the amino acids present, and if the condition is not fulfilled, the flag at this torsion angle is set to true, and the unfulfilled angle counter is incremented by one, then switching to the a new pseudo-random number[7]

-check the implementation of the limit to move away from the center of the coordinates of the amino acids present so far, and if the condition is not met, the flag at that torsion angle is set to true and the counter of non-defeating angles increases by one, then passes to taking a new pseudo-random number

- if the new amino acid is successfully positioned, its coordinates are written in the corresponding array, the flags and the counter of the corners are reset, the center of coordinates of the existing amino acids is updated and the position of the next amino acid [9]

- if the next amino acid cannot be placed, the building of the current conformation is terminated, the counter of the failed conformations is updated and the next one is started.

For each successful build, the value of the evaluation function is calculated. This happens by pairing the amino acids, without the neighboring ones and those through one, and if the distance between them allows for interaction (hydrogen bonding, hydrophobic contact of residues or disulfide bridge), then the first evaluation subfunction is increased by one, in hydrophobicity and two amino acids - the second evaluation sub-function increases with their hydrophobicity values, and with two amino acids Cysteine - the third evaluation subfunction is increased by one. The overall evaluation function is a work of the three sub-functions - thus assessing the simultaneous construction of all the structures of the protein as it should be in the real case [5], [8].

Data for each conformation is recorded in the files, which has the highest value of the evaluation function so far, and the best conformation repeat counter is updated upon reaching the same value. If the last best conformation has repeated many times, it increases the likelihood that it is close to the optimal solution. In one file, the coordinates, the spread parameter, and how many times the last one was found are recorded. A second file records all contacts on numbered amino acid pairs and a total number of contacts. In the third file, only the hydrophobic contacts - number and numbered pairs, as well as the number of disulfide bridges - are recorded.

IV. RESULTS

The difference in analysis with the lattice models is that greater freedom of turning gives a larger number of possible structures. Due to the observed prevailing significance of the failed percentage parameter in the step and the initial value of the coefficient of expansion s, it will change. Also there will be two versions of the evaluation sub-function for the hydrophobic contact - with the parameter of the influence of water wtw = 8 and wtw = 0 - in the first case contacts between polar amino acids bear a positive value, although significantly less than the hydrophobic, whereas in the second variant has a negative value.

The data refer to the study of the 1UUB protein with 56 amino acids, model 1 in (the PDB has 19 models) whose coordinates and contacts, as well as the explanation for its choice, are given in the previous chapter where the study is on the same protein to be equally comparable data [2], [8], [10].

Initially, according to the first goal of finding a conformation with the highest valuation function at different parameters and the same computational burden, in this case the testing of 1,000,000,000 random conformations (if possible in this model 36^{56}), the requirement for a maximum distance between amino acids to form contact was set to 1.4 times the standard distance and not 2 times as given in the model description as the next studies on the second purpose (verification of the model) showed that contact should be considered as 2 times norm earlier away - close enough for the formation of all connections lowering total energy. Nevertheless, these results serve to achieve the first goal [2], [5]- [8].

The original structure has 125 contacts, which are couples of amino-acids, and for every obtained structure we compare how many couples are the same. If many couples of amino-acids from different parts of the peptide chain are placed together in the same way – contacts match, then the 3D form is close to the original.

With 90% failed and wtw = 8, the following best structure with 179 contacts and 26 matching the 125-frame structure was obtained

With 60% failed and wtw = 8, the following best structure with 140 contacts and 37 matches with the 125 framed structure.

At 50% failed and wtw = 0, the following best structure with 140 contacts and 47 matching the 125 frame structure

With 40% failed and wtw = 0, the following best structure with 123 contacts and 40 coinciding with the 125 framed structure.

In comparison to other results of using Off-Lattice models, we have not very different success in prediction – matches of contacts, but the difference is that using variable constraint for 3D size and changing it according to the percent of failed structures, we find out that at proper percent of failed structures the program woks faster, and if we have limit of only 1,000,000,000 random structures, results are better [11].

Our data will be used to predict the three-dimensional structure of the protein with quantum methods. The startup file will be generated to support the NAMD program. The prediction of the 3D structure is not a clear. This method is probable.

V. CONCLUSIONS

If we use variable constraint for 3D size and changing it according to the percent of failed structures, we can predict 3D forms faster and by the present computational resources, it will reflect the number of investigated structures.

On the other hand, the achieved degree of matching between predicted and real structures is not enough to determine completely their biological activity, but can predict some possibility – for some activities they have potential and for others is clear they have not. Models with greater complexity, representing protein structure more realistically, have to be developed, but they have to work fast enough to have reason of their use – milliards of proteins have to be processed.

ACKNOWLEDGMENT

This work is partially supported by the project of the Bulgarian National Science Fund, entitled: "Bioinformatics research: protein folding, docking and prediction of biological activity", code NSF I02/16, 12.12.14.

REFERENCES

[1] A. Kolinski, J. Skolnick, “Monte Carlo simulations of protein folding. i. lattice model and interactionscheme”, Proteins, vol. 18, pp. 338–352, 1994.

[2] B. Berger, T. Leighton, “Protein folding in the hydropho-bichydrophilic (HP) model is NP-complete, Journal of Computational Biology”, vol. 5, pp. 27–40, 1998

[3] N. Stoeva, “The Right of the Personal Data Protection - Nature and Guarantees”, Proceedings of the International Scientific Seminar “Intellectual Property in Bulgaria - Perception, Awareness and Behavior” Trencheva, T. (compl.), Za bukvite-O Pismeneh, Sofia, pp. 89-104, 2018.

[4] R. Mavrevski, “Selection and comparison of regression models: estimation of torque-angle relationships”, C. R. Acad. Bulg. Sci. 67, pp. 1345-1354, 2014.

[5] H.-J. Böckenhauer, D. Bongartz, “Protein folding in the HP model on grid lattices with diagonals”, Discrete Applied Mathematics vol., 155, pp. 230–256, 2007.

[6] L. Fukshansky, S. Robins, “Bounds for solid angles of lattices of rank three”. Journal of Combinatorial Theory, Series A vol., 118 pp. 690–701, 2011.

[7] Y Wai, T. David, P.Landau, “Monte. Carlo simulations of the HP model (the “Ising model” of protein folding)”, vol., 182 (9), Computer Physics Communications, pp. 1896-1899, 2011

[8] A. Albrechta, A.Skaliotisb, K.Steinhöfelb, “Stochastic protein folding simulation in the three-dimensional HP-model”, Computational Biology and Chemistry. vol., 32 (4), pp., 248-255, 2008.

[9] K. Silpaja, Chandrasekar, M.V. Sangaranarayanan, “Exact enumeration of conformations for two and three dimensional lattice proteins”, Computer Physics Communications, vol., 199, pp. 8-11, 2016.

[10] R. A. Laskowski, G.J. Swaminatha, “Problems of Protein Three Dimensional Structures Reference Module in Chemistry”, Molecular Sciences and Chemical Engineering, 2013

[11] M. Traykov, S. Angelov and N.Yanev, “A New Heuristic Algorithm for Protein Folding in the HP Model”, Journal of Computational Biology, vol. 23, (8), pp.45 -51, 2016

[12] B. Verhalen, R. Dastvan, S. Thangapandian, Y. Peskova, H. A. Koteiche, R. K. Nakamoto, E. Tajkhorshid, and H. S. “Mchaourab Energy Transduction and Alternating Access of the mammalian ABC Transporter P-glycoprotein”. Nature, vol. 543: pp. 738-741, 2017.

Protein Binding Sites Prediction by Making Fuzzy-based Selection of the Features

Georgina Mirceva, Andreja Naumoski and Andrea Kulakov
Institute for intelligent systems
Faculty of computer science and engineering, Ss. Cyril and Methodius University in Skopje
Skopje, Macedonia
georgina.mirceva@finki.ukim.mk, andreja.naumoski@finki.ukim.mk, andrea.kulakov@finki.ukim.mk

***Abstract*—In the processes that occur in the living organisms, protein molecules are one of the main components. If we gather knowledge how the proteins influence these processes, we can use that knowledge for designing new drugs in order to regulate these processes. The literature provides plethora of methods that can be used to find out the functions of proteins, and we put our focus on the methods that make functional annotations of proteins by analyzing the binding sites where an inspected protein structure interacts with another protein structure. In this research paper, we introduce a novel approach for determination of the binding sites of proteins. First, we extract the features of the amino acid residues of the protein, then we make fuzzy-based selection of the features, and finally, we build prediction model by applying fuzzy-based classifier. For model induction, we use classification method that is based on the fuzzy set theory, in order to avoid big sensitivity to slight changes in the data. Regarding selection of the features, besides classical techniques, we also consider technique that is based on the fuzzy set theory, with the intention of performing more appropriate selection of the features. Finally, we present some experimental results obtained from the evaluation of our approach.**

Keywords—protein structure; protein binding site; fuzzy logic; feature selection

I. INTRODUCTION

One of the areas in bioinformatics is proteomics that focuses on analyzing protein structures and investigating the functions of these structures in the processes in organisms. The study of proteins is very essential because they are among the most important components in cells, and control various cell processes. Namely, the pharmacologists that design new drugs use the knowledge gathered about the functions of protein molecules, and based on that knowledge they design these drugs that are used for treating particular diseases. Due to the high importance of knowing and understanding the protein molecules, the researchers have developed various methods for discovering protein functions. Although there are various experimental methods for analyzing proteins and finding their functions, they cannot provide this knowledge with a reasonable speed. The innovations in technology lead to rapid increase in the speed of determination of proteins' structure. Hence, the gap between the known protein structures that are functionally annotated, and those that are not annotated yet increases. This shows that the development of computational methods for protein functional annotation is very essential.

The literature offers numerous computational methods for annotating protein structures with protein functions. There are methods [1] that aim to find out which proteins have similar sequence or structure with the inspected protein, and based on the functions of these proteins, they perform functional annotation of the examined protein. These methods are based on the knowledge that there are lots of proteins that have same predecessor and same functions. Due to this, it is also supposed that those parts of the protein molecule that remain the same during evolution define the functions of the proteins. So, the researchers have proposed many methods [2] that analyze protein structures and find the parts of proteins that remain the same, so called conserved parts of the protein. Later, the conserved parts of proteins could be examined, and based on their characteristics the functions of the protein structure could be discovered. In the processes in cells of living organisms, protein structures interact and based on these interactions we can determine the corresponding functions that they will have. In that direction, many studies focus on finding out which are the pairs of protein structures that interact, and these studies result in developing graphs that show the pairs of interacting proteins. These graphs are called protein-protein interaction networks, where the nodes in the graph correspond to the protein structures that are examined. In recent time, there are many research studies [3] that aim to annotate protein structures based on the knowledge from these protein-protein interaction networks by using various algorithms that could be used for analyzing graphs. Regarding the interactions between protein structures, from the experimental methods that are performed to analyze proteins, besides the knowledge gathered that shows which are the pairs for interacting proteins, these methods also provide knowledge where the interaction occur. Namely, for both proteins in the interaction we can identify which are the parts of the structure that is touched by the other structure. In this way, the binding sites of proteins could be determined, where contact with another protein takes place. However, these methods are expensive and time consuming, and require intensive human work. Therefore, the researchers work on developing faster computational methods [4] for finding the binding sites of proteins. The research presented in this paper is focused on the last group of methods. Particularly, in this paper we aim to propose an approach that could be used for binding sites detection. The gathered knowledge could be used to perform functional annotation of protein structures by inspecting the features of the binding sites that are identified with our approach.

Protein structures are composed of amino acid residues that make some particular conformation in the three-dimensional space, and further they are composed of atoms. The examination of the protein structure in order to detect its binding sites could be made by studying the features of its amino acid residues. Various features are used in the literature, but in this study, we use the features that are most often used, i.e. Accessible Surface Area (ASA) [5], Relative ASA (RASA) [6], depth index (DPX) [7], protrusion index (CX) [8] and hydrophobicity [9]. As it was described, protein structures have a number of constituting amino acid residues, which are further composed of several atoms. In the process of extracting the features, various sets of atoms could be considered. For example, we can consider all atoms, or we can divide the atoms based on whether they are positioned on the backbone of the protein or not, or we can divide them based on their type.

After extracting the features of the residues, we can generate a prediction model that will predict which of these residues would be constituents of the binding sites of the protein where interaction with another protein may occur. One way to do that is to apply some of the classical methods for generating classification model. During evolution, some minor changes occur in the protein structures. Although the novel structure is similar with its predecessor, these small changes could lead to different decision in the classification process since the classical methods for classification make crisp decisions. Therefore, in order to generate a prediction model it could be better to use some classification method that is based on the fuzzy set theory, thus overcoming the problems that occur in the classical set theory. The literature provides various fuzzy-based classification methods. Fuzzy decision trees (FDTs) are proposed in [10], [11], optimizations of FDTs are provided in [12] and [13], while [14] gives evidence about their strengths and weaknesses with respect to the classical decision trees. The methods for building FDTs are limited to use a particular type of fuzzy aggregation operator in the model induction process. This is not a case with the fuzzy pattern trees (FPTs), which are proposed in [15], where various fuzzy aggregation operators are combined for model induction. In [15], the model is built in a bottom-up direction, starting from the tree leaves towards the root node. Later, a top-down approach for building FPTs is proposed [16], so by changing the direction of model generation smaller changes are made to the current model. In [16], another termination criterion is proposed in order to fit the model's complexity in accordance to the problem's complexity. In this study, we use the bottom-up FPT approach for generating the prediction model. In [17], we already used the bottom-up and top-down FPTs approaches for building proteins' binding sites prediction models. The samples in the dataset correspond to the amino acid residues, so the number of samples is very high. Therefore, it is very important to reduce the dimensionality of the dataset. For that purpose, in [17] we used various feature selection techniques (FSTs) in order to select the features that are most useful. By selecting the most important features, besides reduction of the dataset, also the complexity of the model could be reduced and also the prediction accuracy could be increased. The reduction of the dimensionality of the dataset and the reduction of the model's complexity result in reduction of the time needed for training the model and making predictions for unknown proteins. Protein binding sites prediction could be further advanced by applying an appropriate fuzzy-based FST. Namely, before using some classification method based on the fuzzy set theory, instead of using some classical FST we may use a FST based on the fuzzy set theory. In that way, the feature selection is expected to provide better filtering of the most relevant features. In this paper, we propose a novel approach for protein binding sites prediction that is based on the fuzzy set theory. Namely, besides using fuzzy-based classification method for generating models, we use a fuzzy-based FST that is presented in [18], [19].

The rest of this research paper is organized as follows. Section 2 provides brief explanation of the proposed approach. The results obtained from the evaluation of our approach are shown and discussed in Section 3. And finally, in Section 4 we make final conclusions and we identify several directions for further improvements.

II. The Proposed Approach for Protein Binding Sites Prediction

The proposed approach has three steps. First, several features of the residues of the protein structure are extracted. Then, in the second step a selection of these features is performed. In this study, we make selection of the features in three ways. Finally, a prediction model is generated in the third step by using the bottom-up FPT approach.

A. First Step - Extracting the Features of the Amino Acid Residues

In this paper we consider the features of the amino acid residues that are most widely used for binding sites prediction, i.e. 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] is a very useful feature because it indicates whether a given residue of the protein could be touched by some residue of the interacting protein. This feature is calculated by applying the algorithm proposed in [5], where a 'probe' sphere with a radius of 1.4 Å is rolled over the surface of the examined protein. As it was previously described, protein structures are composed of amino acid residues, which are further composed of atoms. Therefore, this algorithm estimates the ASA for each amino acid residue. A large number of residues are in the protein interior, and the interacting protein would not touch them in the interactions, meaning that they are not real candidates for forming binding regions, and can be removed for training the model and making predictions. For that purpose, in the dataset we take into account only the residues for which at least 5% of their surface is touched by the 'probe' sphere [20].

For each amino acid, it is known which are its constituting atoms and how they are folded in the three-dimensional space. The number of atoms varies for different amino acids, so instead of ASA, the Relative ASA (RASA) [6] could be estimated to reduce the effect of having larger or smaller amino acid. RASA is calculated by dividing the estimated ASA and the standard ASA [6] for the examined amino acid. We use the NACCESS program [6] for estimating ASA and RASA.

The depth index (DPX) [7] gives evidence how deep is a given atom from the surface of the protein. It is measured as Euclidean distance to the nearest atom that is touched by the 'probe' sphere, which are actually the atoms with ASA greater than zero.

The protrusion index (CX) [8] shows the propensity of an atom. For that purpose, the non-hydrogen atoms are examined, and for each of them we count how many heavy atoms they have in their surroundings. The surroundings of an atom is actually a sphere around that atom with radius of 10 Å [8]. Then, the volume occupied by these atoms is estimated by multiplying the number of counted heavy atoms in the inspected surroundings and the average volume of atoms. The protrusion index CX is measured as a ratio between the remaining volume (the non-occupied volume) and the occupied volume. This means that the atoms that have a small number of heavy atoms in their neighborhood, would have a high protrusion index, and vice versa.

The hydrophobicity [9] is a well-known characteristic of the amino acids, which shows their preferences to be in the interior of the protein structure or towards the surface of the protein structure. In this study, we use the hydrophobicity scale that is proposed in [9].

The features described above, except hydrophobicity, are calculated for each atom. On the other side, the samples in the dataset correspond to the amino acid residues, which are composed of several atoms. Therefore, the features of the residues are calculated by aggregating the values of the features of its atoms. In the aggregation of ASA and RASA, various sets of atoms are used: set of all atoms, set of the backbone atoms, set of the side-chain atoms, set of the polar atoms and set of the non-polar atoms. Regarding aggregation of DPX and CX, the minimal, maximal and mean values are considered. There is no need for aggregation regarding the hydrophobicity feature because it is a characteristic of the amino acids, so for the samples in the dataset, which are amino acid residues, we directly have the value of this feature. Because we filter only the residues that are on the protein surface, therefore for each residue in the dataset the minimal DPX is zero, so we discard this feature.

B. Second step – Selection of the Features

After extracting the features of the amino acid residues, next we make selection of the features. In our previous studies, we first started using total ASA, average DPX, average CX and hydrophobicity. Some experiments by using this set of features are presented in [17]. In this paper, we also use this set of features. However, the selection of these features was not made by using some computational algorithm. Therefore, in [17], we applied various FSTs that are found in the literature. These techniques are based on the classical set theory. As we described in [17], the FSTs could be wrapper or filter techniques. With wrapper techniques, the method that is used for generating classification model is used during the feature selection phase, so it is applied on various sets of features and for each set, the performance of the classification model is measured. In this way, the objective function is optimized, for example the classification accuracy is maximized. On the other side, filter techniques try to optimize some other objective function. For example, with the FST we may try to optimize the correlation of the features and the class attribute, while the final objective function that is used for evaluation of the performance of the prediction model could be its classification accuracy. Among the filter FSTs, some consider features individually, while others consider a set of features and inspect the whole set of features. In our previous study [17], we presented and applied various FSTs. In this study we focus on the filter FSTs that individually ranks the features.

From the ranking of the FSTs made in [17], generally the technique denoted as InfoGain, which is based on the information gain measure showed best performance. As we said before, the classification models in this study are made by applying the bottom-up FPT approach. In [17], by applying this classifier in combination with various filter FSTs, generally most accurate model was obtained by using the technique where the features are ranked based on the information gain. Therefore, in this research we consider this technique from the techniques that are based on the classical set theory.

In order to improve the predictions, further we aim to apply a FST that would make more appropriate selection of the features. Since we are using a fuzzy-based classifier for generation of the models, we apply the fuzzy decision reducts technique [18], [19], which is based on the fuzzy set theory. We use the implementation of this technique provided by its authors, by using the default set up.

C. Third step – Generating Classification Model

In this paper, the generation of the model is done by applying the bottom-up FPTs approach [15], where for all classes, separate prediction tree is generated. In this study, there are two classes, classes for binding and non-binding residues. After generating trees for all classes, during testing, for each residue of the inspected protein the similarities with all trees are calculated and the residue is classified in the class with highest similarity. In this study, the calculation of similarity is based on the Root-Mean Squared Error (RMSE).

For generation of FPTs, first, fuzzification of the dataset is made. For that purpose, the triangular, trapezoidal and Gaussian membership functions (MFs) are used. In the fuzzification, each feature is labeled with fuzzy terms. In the process of feature selection, we selected the four most relevant features. Lets use N=5 fuzzy terms for each of these features. In this way, the number of fuzzy terms is 4*5=20. For each fuzzy term, we obtain a separate tree at level 0, so called primitive trees. These trees could be considered as models. However, they are very simple and would not provide acceptable predictions, so the model induction continue on the next, upper level, by combining these primitive trees. First, for each primitive tree, we calculate the similarity between the inspected fuzzy term and the examined class. After calculating the similarity for each primitive tree, the primitive tree that lead to highest value for similarity is identified, and it is further combined with the remaining primitive trees. For that purpose, various fuzzy aggregation operators could be considered. In this study we use AND and OR fuzzy aggregation operators. The same procedure is repeated further on the upper levels until

a stop criteria is satisfied. In this study, the generation of the tree stops when the number of levels with parent nodes is 5. We can obtain simple model or general model, based on that whether only primitive trees are considered to be aggregated with the current tree, or also the trees from the upper levels are considered for aggregation. In this study we consider all trees for aggregation, thus obtaining general model. As it was described before, separate FPT is generated for each class, and test samples are classified in the class that lead to highest value of similarity.

III. EVALUATION

The dataset used for evaluation of the proposed approach holds knowledge gathered from the Biomolecular Interaction Network Database (BIND) [21]. This database is filled with experimentally determined binding sites of proteins. We do not consider all proteins from the BIND database because it holds data for many of them, whereas many of them have some homologous in the database. Therefore, as it is commonly used in proteomics, first the most representative protein chains are filtered based on the similarities of their sequences. In this study, in the test set we consider the chains with less than 10% similarity in the sequence, similarly as we did in [17]. The training set is formed by considering the chains with less than 20% sequence similarity, by considering only the chains that are not in the test set. As it was described in section 2, in the dataset we take into consideration only the amino acid residues that are on the surface of the protein structure. In this way, we obtain 115579 samples for training and 625939 samples for testing the models. In our case, the classes are not uniform, because the class that corresponds to the non-binding residues has 86.42% of the samples, according to the knowledge from the BIND database. Therefore, we sample the set with training samples to 27% of its size, thus obtaining *train100* set. The sampling is done without replacing the samples, and to obtain uniform distribution of the classes. Regarding the test set, it is important to note that it is not sampled towards uniform distribution, so it remains unbalanced. Therefore, we should use an evaluation measure that is suitable for unbalanced sets. In this paper, we use the AUC-ROC evaluation measure, which is appropriate for unbalanced test sets. After balancing of the training set, we applied min-max normalization in the interval [0, 1] of the features in both sets.

The fuzzy-based FST that is used has significantly higher memory requirements, so we were not able to apply it on the entire training set. Therefore, we make further sampling of the training set towards 90% and 10% of its size, thus obtaining *train90* and *train10* sets. Here, the sampling is made in the same manner (without replacement and by keeping uniform class distribution). Regarding model induction, we made experiments by using the entire training set, *train100*, and by using *train90* and *train10* sets. We made experiments by using FPTs with triangular, trapezoidal and Gaussian membership functions. Regarding the set of features, we examined three variants. The first set of features contains total ASA, average DPX, average CX and hydrophobicity, which we used in our former research. The second set of features is obtained by applying the classical FST where the features are ranked according to the information gain, and then the four most relevant features are selected. The third set of features is obtained by applying the fuzzy decision reducts technique, by selecting the four features that have highest rank. In all three cases, we have sets with four features.

Table 1 presents the experimental results for AUC-ROC by using the first set of features, which contains total ASA, average DPX, average CX and hydrophobicity. The results obtained by using the FST based on information gain measure are given in Table 2, while Table 3 shows the results achieved by the fuzzy decision reducts FST. The bolded results are those that are highest by using particular sets for feature selection and training the model, and a given set of features. The underlined results are those which are highest by using particular sets for feature selection and training, and a given membership function for building FPT.

The results shows that generally by using the fuzzy decision reducts technique for feature selection the highest AUC-ROC is obtained. By using the set of total ASA, average DPX, average CX and hydrophobicity, generally lowest results are obtained, except when using Gaussian membership function and using the *train90* set for both feature selection and training the prediction model. From the results given in Table 1 and Table 2, it is interesting to note that when using the first two sets of features lower results are obtained when using the entire set for training (*train100*) instead of using some sample of the set (*train90* or *train10*), which is due to overfitting of the model. On the other side, when using the third set of features, which is attained by making selection with the fuzzy decision reducts technique, generally this is not the case.

TABLE I. THE RESULTS FOR AUC-ROC BY USING THE TOTAL ASA, AVERAGE DPX, AVERAGE CX AND HYDROPHOBICITY

Set for training	Membership function		
	Triangular	*Trapezoidal*	*Gaussian*
train90	0.559	0.566	**0.586**
train100	0.546	**0.566**	0.546
train10	0.557	**0.566**	0.557

TABLE II. THE RESULTS FOR AUC-ROC BY USING THE FEATURE SELECTION TECHNIQUE BASED ON INFORMATION GAIN

Set for feature selection	Set for training	Membership function		
		Triangular	*Trapezoidal*	*Gaussian*
train90	*train90*	0.577	0.566	**0.579**
train10	*train100*	0.551	**0.566**	0.554
train10	*train10*	0.562	0.543	**0.563**

TABLE III. THE RESULTS FOR AUC-ROC BY USING THE FUZZY DECISION REDUCTS TECHNIQUE

Set for feature selection	Set for training	Membership function		
		Triangular	*Trapezoidal*	*Gaussian*
train90	*train90*	**0.570**	**0.570**	0.567
train10	*train100*	**0.569**	0.563	**0.569**
train10	*train10*	0.564	0.564	**0.567**

IV. CONCLUSION

In this paper, we proposed a novel approach that can be used for automatic identification of the amino acid residues of proteins that could be part of binding regions where interactions occur. This approach has three steps. The first step is to extract the features of the amino acid residues. Then, in the second step feature selection is performed by applying the fuzzy decision reducts technique. Finally, the prediction model is generated in the third step by using the FPT classifier.

In the evaluation, we made comparison of the models obtained by using the set that contains the total ASA, average DPX, average CX and hydrophobicity features, against the sets obtained by making feature selection with the technique based on information gain and the fuzzy decision reducts technique. According to the experimental results, generally best prediction models are obtained by using the fuzzy decision reducts technique.

Further, we plan to apply some other feature selection techniques based on the fuzzy set theory, in order to make more appropriate selection of the features. We will also continue our study by considering some other more appropriate features of the residues of protein structures.

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, April 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, April 2004.

[3] M. Kirac, G. Ozsoyoglul, and J. Yang, "Annotating proteins by mining protein interaction networks," Bioinformatics, vol. 22, no. 14, pp. e260–e270, July 2006.

[4] 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, May 2009.

[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, September 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, January 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, July 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, May 1982.

[10] C.Z. Janikow, "Fuzzy decision trees: issues and methods," IEEE Transactions on Systems, Man, and Cybernetics, vol. 28, no. 1, pp. 1–14, February 1998.

[11] C. Olaru and L. Wehenkel, "A complete fuzzy decision tree technique," Fuzzy Sets and Systems, vol. 138, no. 2, pp. 221–254, September 2003.

[12] A. Suárez and J.F. Lutsko, "Globally optimal fuzzy decision trees for classification and regression," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 21, no. 12, pp. 1297–1311, December 1999.

[13] X. Wang, B. Chen, G. Olan, and F. Ye, "On the optimization of fuzzy decision trees," Fuzzy Sets and Systems, vol. 112, no. 1, pp. 117–125, May 2000.

[14] Y.L. Chen, T. Wang, B.S. Wang, and Z.J. Li, "A Survey of Fuzzy Decision Tree Classifier," Fuzzy Information and Engineering, vol. 1, no. 2, pp. 149–159, June 2009.

[15] Z.H. Huang, T.D. Gedeon, and M. Nikravesh, "Pattern trees induction: a new machine learning method," IEEE Transaction on Fuzzy Systems, vol. 16, no. 3, pp. 958–970, August 2008.

[16] R. Senge and E. Hüllermeier, "Top-Down Induction of Fuzzy Pattern Trees," IEEE Transactions on fuzzy systems, vol. 19, no. 2, pp. 241–252, April 2011.

[17] G. Mirceva and A. Kulakov, "Improvement of protein binding sites prediction by selecting amino acid residues' features," J. Struct. Biol., vol. 189, no. 1, pp. 9–19, January 2015.

[18] R. Jensen and Q. Shen, "New approaches to fuzzy-rough feature selection," IEEE Transactions on Fuzzy Systems, vol 17, no. 4, pp. 824–838, August 2009.

[19] C. Cornelis, R. Jensen, G. Hurtado, D. Ślezak, "Attribute selection with fuzzy decision reducts," Information Sciences, vol. 180, no. 2, pp. 209–224, January 2010.

[20] C. Chothia, "The Nature of the Accessible and Buried Surfaces in Proteins," J. Mol. Biol., vol. 105, no. 1, pp. 1–12, July 1976.

[21] 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, January 2001.

Analysis of Medical and Health-Related Data about Adult Obesity using Supervised and Unsupervised Learning

Gordana Ispirova
Jožef Stefan International Postgraduate School
Computer Systems Department, Jožef Stefan Institute
Ljubljana, Slovenia
gordana.ispirova@ijs.si

Tome Eftimov
Computer Ssystems Department, Jožef Stefan Institute
Ljubljana, Slovenia
tome.eftimov@ijs.si

Barbara Koroušić Seljak
Computer Systems Department, Jožef Stefan Institute
Ljubljana, Slovenia
barabara.korousic@ijs.si

***Abstract*— Obesity is a growing problem in most developed countries and it is responsible for a significant degree of morbidity and mortality. Overweight and obesity are linked to more deaths worldwide than underweight, and globally there are more people who are obese than underweight. This paper focuses on working with medical and health-related data concerning the problem of adult obesity, in particular using several machine learning algorithms on this kind of data to gain information from the collected data. In order to get the most out of the data, two machine learning techniques – supervised and unsupervised learning are used. Adequate labeling of the data is introduced and classification is selected as a supervised learning technique. Various classification algorithms are applied, using 10–fold cross validation, and a comparison of the results is presented. Clustering is used as an unsupervised technique, with the goal of identifying groups of participants that exhibit similar behavior in terms of the results from the programs. For this purpose, we applied a k-medoids clustering algorithm known as *PAM* (Partitioning Around Medoids) and a visualization technique to represent the detected clusters. The paper covers: data extraction, data preprocessing, classification, clustering and summary of the results.**

Keywords—supervised learning, unsupervised learning, data mining, classification, clustering, mixed data types

I. Introduction

Healthy eating and proper nutrition is a hot topic these days and researchers from this area tend to incorporate technology and automation as much as possible. One of the most common modern world problem is adult obesity, which is taking over in most developed countries. There are a lot of projects around the globe with the purpose of resolving this problem. This research is based on such program, which is part of a project from the department of Computer Systems at Jožef Stefan Institute, and is aimed towards obese adults who are trying to lose weight in a healthy way with the help from qualified medical specialists. This program was active all across Slovenia, in many cities, and the data is relatively new and has not been explored and analyzed previously. Health – related data of this type is very valuable and we chose to analyze it in order to help improve the program. After data exploration, summary and statistics based on all collected data an adequate feature extraction is selected and the relevant features are derived. As a supervised learning method we chose classification with previous adequate labeling of the data, and as an unsupervised method we chose clustering. Results gained from this research give a representative of how the level of expertise of the trainer - based on their education, profession and years of experience, and the quality of the program provided by the medical centers contributes to the participant's success in losing weight.

II. Data

There is an existing database with data which was collected from a group of obese adults who followed the program. The program included a user interface where each user (participant in the program) provided the necessary data: personal data, gender, age, anthropometry, dietary habits, demography, medical data and participation and attendance at the courses, progress in losing weight and similar. Also, data about the educators of the courses is available and which program (old or new) the participant entered.

A. Data preprocessing

The data is contained in multiple data tables, for the needs of this project a single table is constructed. First, the data base is stored locally and the extraction and derivation of the features is made mainly with *SQL* queries and *R* code. Filling the newly constructed table is made with extracting features (columns) from joining two, three or more tables and *SQL* functions for calculating some relevant features. The end result is one base table with the possibility of adding new features. After obtaining the finished table, the data is exported into a *".csv"* file for working with the data in *Weka* and *R Studio* (Table I).

TABLE I. REPRESENTATION OF THE DATA

participant_id	35	36
age	53	33
gender	M	F
body_height	188	161
center_id	2a	3a
workshop_id	2a	3a
workshop_new	y	y
performer_id	7a	8a
performer_education	52a	57a
performer_proffesion	1a	5a
performer_years_of_experience	24	9
num_of_meals	4	5
weight_v1	112	102
weight_v17	108	101
waistc_v1	114	112
waistc_v17	107	108
cholesterol_v1	5	4
cholesterol_v17	5	4

B. Data description and understanding

The data consists of 275 instances which is the number of participants in the program, and 17 attributes – the first one being the identification number – *'participant_id'* of the participant. Therefore, for every participant there are 16 attributes that describe him/her and his/her participation in the program. From these 16 which are of relevance 10 are numerical:

- *'age'*
- *'body_height'*
- '*performer_years_of_experience'*
- *'num_of_meals'*
- *'weight_v1'* – weight in first session
- *'weight_v7'* – weight in last session
- *'waistc_v1'* – waist circumference
- *'waistc__v17'*
- *'cholesterol_v1'*
- *'cholesterol_v17'*

And 6 are nominal:

- *'gender_id'*
- *'city_id'* – city of the health-care center
- *'workshop_id'*
- *'workshop_new'* – old or new program
- *'preformer_id'* – id of the trainer
- *'performer_education'*
- *'perfromer_profession'*

Thus, we are dealing with attributes from mixed data types. The algorithm that is chosen for the purpose of clustering the data does not handle missing values. After reading the data in *R Studio* and observing that there are missing values for some attributes we further preprocess the data with *SQL* and *R Studio*. For missing values in the column *'weight_v17'* we went back to the database and filled the fields with the last weight the users entered in their progress. The same is made for the missing values for the columns *'waist_v17'* and *'cholesterol_v17'* . However, for the missing values in the columns 'weig*ht_v1'* and *'age'* there are no such options, the only option is to enter an average value, but because the number of instances missing these values is not big and usually both of the values are missing at the same time, the whole instances missing these values are deleted. Thus, we are left with 239 instances to work with.

III. CLASSIFICATION

In order to perform classification to this data the first step is to label the data. Choosing the proper way to label the data is a major problem when dealing with unlabeled data, like in this case. The first approach is constructing a success function – the amount of kilograms lost during the program (Fig. 1).

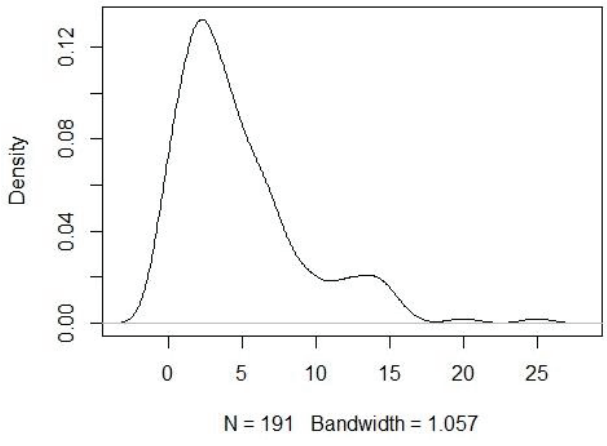


Fig. 1. Distribution of the success function

For this purpose, a new feature, *'weight_lost'* is introduced to the table. The value of this new attribute is calculated by

subtracting the first measured weight of the participant – *'weight_v1'* and the last measured weight – *'weight_v17'*.

Based on this attribute the instances are labeled with two labels, thus, binary classification is introduced and this attribute is removed:

- *'s'* – Indicating that the participant lost a small amount of weight.
- *'l'* – Indicating that the participant lost a large amount of weight.

Looking into the data, outliers, like instances representing people who gained weight, who clearly did not follow the program, are detected. These instances are deleted from the data set.

To better estimate the splitting point the distribution of the success function is plotted in *R Studio*. From the distribution of this function, and further looking into the data, 4 kg is selected as a splitting point for the binary classification. With this type of labeling we get a balanced classification: 99 instances from class 's' and 92 instances from class 'l'. The labeled data is then exported as a '$.csv$' file and imported in *Weka.* Several algorithms are then applied on this data, with previously removing the attributes: *'participant_id', 'weight_v1', 'weight_v17', 'waistc_v17', 'cholesterol_v17'* and *'weight_lost'*. The results obtained show that this type of labeling is not adequate. The highest accuracy is gained from the $FilteredClassifier$ algorithm – 62.3037% and recall of 0,737.

The second approach is deriving new attributes, indicating the *BMI* (Body Mass Index) of the participant at the beginning of the program and at the end, and their difference. Thus, three new attributes are introduced: *'BMI1', 'BMI2', 'BMI_lost'* and the previously added attribute – *'weight_lost'* is removed from the data.

The body mass index is a value derived from the mass (weight) and height of an individual. It is defined as the body mass divided by the square of the body height, and is universally expressed in units of kg/m^2 [1].

With further exploration about how the BMI is connected with the person's health and the relation of this measure and weight lost the instances are split in two classes:

- *'g'* – Indicating that the participant has a lot of improvement from the program.
- *'b'* – Indicating that the participant did not do so well and did not get good results from the program.

An instance is labeled with one of the two classes based on a Boolean expression.

- An instance is labeled with class *'b'* if the following Boolean expression is true:

$$if\ (BMI2 \geq 35\ AND\ BMI\_LOST \leq 2)\ OR\ (BMI2 \geq 40\ AND\ BMI\_LOST \leq 3.5)\ OR\ (BMI2 < 35\ AND\ BMI\_LOST = 0) \quad (1)$$

- An instance is labeled with class *'g'* if it does not satisfy (1).

A person is obese if his/her BMI is over 30. There are three categories of obesity based on the BMI (Table II). Loosing 10 kg is equal to lowering your BMI by less than 3.5 and loosing 5 kg is equal to lowering your BMI less than 2. Research [1] show that overweight people lose weight faster. The condition in the Boolean expression (1) is adjusted based on the class of obesity to which the participant belongs and BMI lost during the program. If a person is severely obese (Obese Class II) and his/her BMI hasn't dropped more than 2, implying that the person hasn't lost more than 5 kg for the 17 weeks that the program was active, that means that this person did not do well in the program. The same goes for the next category – if a person is very severely obese (Obese Class III) and during the program he/she lost less than 10 kg (BMI is lowered by less than 3.5) then, this person did not get significant results from the program.

TABLE II. CATEGORIES ACCORDING TO BMI

Category	BMI (kg/m^2)	
	from	*to*
Very severely underweight		15
Severely underweight	15	16
Underweight	16	18.5
Normal (healthy weight)	18.5	25
Overweight	25	30
Obese Class I (moderately obese)	30	35
Obese Class II (severely obese)	35	40
Obese Class III (very severely obese)	40	

With this type of labeling we get 117 instances that belong to the class *'g'* and 74 instances that belong to the opposite class – class *'b'*. The attributes: *'participant_id', 'weight_v1', 'weight_v17', 'waistc_v17', 'cholesterol_v17', 'BMI2'* and *'BMI_lost'* are not taken into consideration when applying the algorithms. The results obtained from several algorithms are shown in Table III. The results presented in this table show that there is no practical significance between the last four algorithms, while there is a small practical significance between the last four algorithms and the first two.

TABLE III. RESULTS FROM SEVERAL CALSSIFICATION ALGORITHMS

Algorithm	Accuracy	TP Rate	FP Rate	Precision	Recall
J48	95.8115%	0.958	0.031	0.961	0.958
RandomForest	94.7644%	0.948	0.053	0.948	0.948
LogiBoost	96.3351%	0.963	0.028	0.965	0.963

Algorithm	Accuracy	TP Rate	FP Rate	Precision	Recall
FilteredClassifier	96.3351%	0.963	0.028	0.965	0.963
ClassificationViaRegression	96.3351%	0.963	0.028	0.965	0.963
Bagging	96.3351%	0.963	0.028	0.965	0.963

IV. CLUSTERING

Finding similarities and placing people in groups (clusters) is a task of the descriptive data mining method. Clustering is a commonly used technique of data mining under which patterns are discovered in the underlying data [2]. In this case, with this data, clustering data mining algorithms can be used to find the different groups of people in the program which have similar body characteristics and similar results from the programs. The main factor that determines which algorithm to use is the presence of mixed types of variables – continuous and categorical. For this reason, the most commonly used clustering algorithm: k-means can't be applied in this case, as an alternative we are going to use the k – medoids technique. In this technique, instead of taking the mean value of the objects in a cluster as a reference point, the clusters are represented by the most centrally located object which are called medoids. The k – medoids method is more robust than k – means in the presence of noise and outliers because a medoid is less influenced by outliers or other extreme values than a mean [3].

In our case a cluster is represented by a medoid – a participant and whether a participant is part of a cluster will be determined based on the similarity with the medoid participant. The selection of medoids starts with random choices and continues with replacements based on a cost function defined by the algorithm used until convergence criterion is satisfied [3].

There are several clustering algorithms and for the purpose of this project the decision is to use the *PAM* (Partitioning Around Medoids) algorithm [4].

The *PAM* algorithm is applied to the data in *R Studio Version 1.0.44* with including the required *cluster* package. The first step is reading the data set from the exported csv file in *R Studio*. Next, because of the fact that *R* chooses to take each variable (vector) of type factor, each column is explicitly defined how it should be considered – as numerical (continuous) or as factor (categorical). In total, there are three decisions that needed to be taken for this approach of clustering data of mixed types:

1. Calculating distance.
2. Choosing a clustering algorithm.
3. Selecting the number of clusters.

A. Calculating distance

A popular choice for calculating dis(similarity) for clustering is Euclidean distance. However, Euclidean distance is only valid for continuous variables, and thus is not applicable here. In order for a clustering algorithm to yield sensible results, we used a distance metric that can handle mixed data types – Gower distance.

The concept of Gower distance is actually quite simple. For each variable type, a particular distance metric that works well for that type is used and scaled to fall between 0 and 1. Then, a linear combination using user – specified weights (most simply an average) is calculated to create the final distance matrix.

The Gower distance can be calculated in one line using the *daisy* function, with parameters: the data set minus the first column – *'participant_id'* and the corresponding type of distance metric – gower. Numeric columns are recognized as interval scaled variables and columns of class factor are recognized as nominal variables. The values in the distance matrix are calculated as follows:

$$d_{i,j} = \frac{\sum_{k=1}^{p} w_k \times \delta_{ij,k} \times d_{ij,k}}{\sum_{k=1}^{p} w_k \times \delta_{ij,k}} \quad (2)$$

Where $d_{i\,j,k}$ is the dissimilarity coefficient:

- for nominal variables is 1 if objects i and j have the same "state" for the k-th variable and 0 if they are different.
- for numerical variables is the k-th variable contribution to the total distance – a distance between *x[i,k]* and *x[j,k]* where these numbers are calculated by dividing each entry by the range of the corresponding variable after subtracting the minimum value so we get a rescaled variable of range [0,1].

The weight $\delta_{i\,j,k}$ becomes 0 when one of the variables *x[i,k]* and *x[j,k]* is missing. In all other situations it is 1. This coefficient is always 1, concerning the data we are working with, because there are no missing values in the data set.

In order to check the results from the daisy function the most dissimilar and most similar pair are extracted (Table IV).

B. Choosing a clustering algorithm

The *PAM* algorithm, in our case, is applied in the following steps:

1. Choose k random entities to become the medoids.
2. Assign every entity to its closest medoid (using our custom distance matrix in this case).
3. For each cluster, identify the object that would yield the lowest average distance if it were to be re-assigned as the medoid. If so, make this object the new medoid.
4. If at least one medoid has changed, return to step 2. Otherwise, end the algorithm [3, 4, 6].

TABLE IV. MOST SIMILAR AND MOST DISSIMILAR PAIR OF PARTICIPANTS

	Most simililar		Most dissimilar	
participant_id	228	236	138	248
age	55	60	86	37
gender	F	F	F	M
body_height	163	164	160	173
center_id	15	15	10	14
workshop_id	20	20	15	23
workshop_new	y	y	y	y
performer_id	49	49	24	42
performer_education	52	52	50	52
performer_proffesion	1	1	2	1
performer_years_of_experience	24	24	1	14
num_of_meals	5	5	3	5
weight_v1	79	86	70	204
weight_v17	76	79	65	190
waistc_v1	105	102	86	164
waistc_v17	100	98	83	160
cholesterol_v1	5	5	7.2	4.7
cholesterol_v17	4	5	5.9	4.7

C. *Selecting the number of clusters*

Selecting the number of clusters is the most important step in the clustering algorithm. There are algorithms that do that automatically, but that is not the case with the *PAM* algorithm. In order to manually choose the number of clusters, an internal validation metric, called silhouette width is used. Silhouette width is an aggregated measure of how similar an object is to its own cluster compared to its closest neighboring cluster. The metric can range from -1 to 1, where higher values are better. After calculating silhouette width for a number of clusters ranging from 2 to 30 for the *PAM* algorithm, we can see that 12 and 13 clusters yield the highest value (Fig. 2), thus 13 is selected as the number of clusters [6]. When the number of clusters is known the next step is to use the function *pam* on our data. The parameters passed on this function are: the distance (dissimilarity) matrix with dimensions [239:239], the number of clusters – previously discovered, and the argument *'diss'* was set to *'TRUE'* to explicitly say that we are working with dissimilarity of the objects.

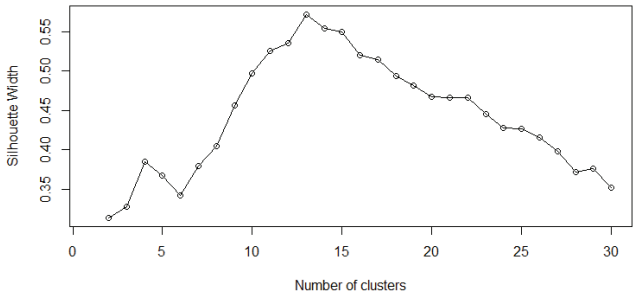


Fig. 2. Deciding for the number of clusters

This functions returns an object which among other things contains a matrix where each row gives numerical information for one cluster: size – number of objects (entities) in one cluster, the maximal and average dissimilarity between the objects in the cluster and the cluster's medoid, the diameter of the cluster (maximal dissimilarity between two objects of the cluster), and the separation of the cluster (minimal dissimilarity between an object of the cluster and an object of another cluster). The intra-cluster similarity is the diameter and inter-cluster similarity is the separation.

After running the algorithm and selecting 13 clusters, an interpretation of the clusters is made by running summary on each cluster (Table V).

TABLE V. NUMERICAL INFORMATION ABOUT EACH CLUSTER

cluster	size	max_diss	av_diss	diameter	separation
1	10	0.1715	0.0759	0.2501	0.1527
2	25	0.2254	0.0929	0.3062	0.1832
3	10	0.1117	0.0650	0.1736	0.2629
4	10	0.1711	0.0932	0.2016	0.1573
5	26	0.2118	0.1000	0.3619	0.1124
6	19	0.3207	0.1237	0.3619	0.1124
7	7	0.0863	0.0498	0.1187	0.1128
8	25	0.1877	0.0892	0.2862	0.1473
9	25	0.2336	0.1190	0.3274	0.2629
10	12	0.1921	0.0964	0.2710	0.2870
11	20	0.1632	0.0798	0.2476	0.1527
12	25	0.1783	0.0810	0.2778	0.1855
13	25	0.3209	0.1125	0.3457	0.1855

Fig. 3. Visualization of the clusters

D. *Visualization*

One way to visualize many variables in a lower dimensional space is with t – distributed stochastic neighborhood embedding, or t – SNE. This method is a dimension reduction technique that tries to preserve local structure so as to make clusters visible in a 2D or 3D visualization. While it typically utilizes Euclidean distance, it has the ability to handle a custom distance metric like the one we created above. In this case, the plot shows the thirteen well separated clusters that *PAM* is able to detect (Fig. 3) [5].

The experiment is repeated after removing: *'center_id', 'workshop_id', 'workshop_new', 'performer_id', 'performer_education', 'performer_profession', 'performer_years_of_experience'*. This results in two clusters where the participants are divided by gender. After removing the 'gender' column we end up with two very mixed clusters, indicating bad clustering (Fig. 4).

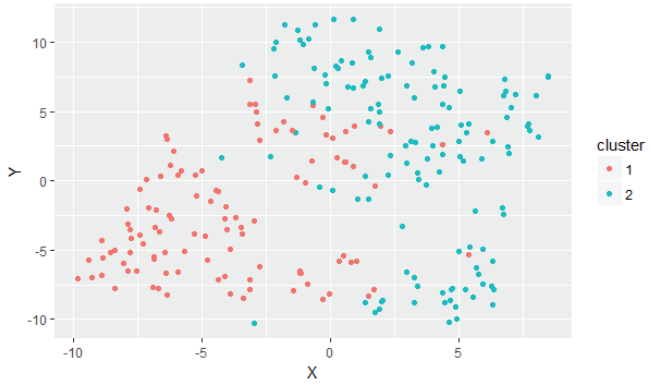


Fig. 4. Visualization of the clusters on the reduced dataset

V. Disscusion

Applying machine learning algorithms to health data or medical data is not a novelty, it is something that is really necessary and it provides valuable information. Similar studies have been done before. A study [7] uses massive multivariate analysis to characterize different types of healthcare users, both in terms of resource utilization and socio-demographic variables. Another study [8] shows that the k-means clustering algorithm appeared to be the most appropriate in healthcare claims data with highly skewed cost information when taking into account both change of cost patterns and sample size in the smallest cluster. Also, there is a simpler study [9] with very good outcome about clustering medical data to predict the likelihood of diseases.

VI. Conclusion

This paper gives a detailed analysis and a total overview of medical and health – related data by applying algorithms both of supervised and unsupervised learning. By labeling the data properly we were able to apply several classification algorithms compare the results, and therefore construct models for predicting the success of the participant in the program. Despite classification, by applying the *PAM* algorithm the participants were divided into 13 groups – clusters and we were able to see the common features in the clusters. These results can be used by health specialists to determine ways to improve the program or give advice to new obese patients.

REFERENCES

[1] Madden and D. Patrick, "Body mass index and the measurement of obesity", in A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

[2] N. K. Sidhu, R. Kaur, "Clustering in data mining," in International Journal of Computer Trends and Technology (IJCTT), 2013.

[3] J. Han, J. Pei, M. Kamber, Data Mining: Concepts and Techniques 3, 2011.

[4] L. Kaufman, P. J. Rousseeuw, Finding Groups in Data: An Introduction to Cluster Analysis, 1990.

[5] L. Van Der Maaten, G. Hinton, "Visualizing data using t-SNE," in Journal of Machine Learning Research 1, 2008, pp.1-48

[6] D. P. Martin, Clustering Mixed Data Types in R, 2016.

[7] T. Lefèvre, C. Rondet, I. Parizot, and Pierre Chauvin , "Applying Multivariate Clustering Techniques to Health Data: The 4 Types of Healthcare Utilization in the Paris Metropolitan Area" , in Plos One, 2014.

[8] M. Liao, Y. Li, F. Kianifard, E. Obi, and S. Arcona," Cluster analysis and its application to healthcare claims data: a study of end-stage renal disease patients who initiated hemodialysis", in BioMed Central Nephrol, 2016.

[9] R. Paul, and A. S. Md. L.Hoque, "Clustering medical data to predict the likelihood of diseases" , Digital Information Management (ICDIM), 2010.

Re-ranking candidates using fairness criteria

Ivan Kitanovski, Katarina Trojacanec, Ivica Dimitrovski and Suzana Loskovska
Faculty of Computer Science and Engineering
University Ss. Cyril and Methodius
Skopje, Macedonia

***Abstract*—In this paper we present a strategy for fair ranking of candidates during retrieval. The strategy is based on the standard notion of protected groups and it aims to ensure that the proportion of protected candidates in the top rankings remains above or is approximately at a specified minimum criteria. Utility is maintained by, first, making sure that every candidate included in the top ranks is more qualified that every candidate not included, and, second, for every pair of candidates in the top rankings, the more qualified should be ranked above. An efficient algorithm is presented for producing fair ranking and is experimentally tested on the COMPAS dataset. The results show that our approach produces small distortions in utility with respect to rankings that do not use fairness criteria.**

***Keywords*—ranking, nias in computer systems, top-k selection, retrieval**

I. Introduction

Search engines are increasingly used for people finding, whether it is for recruiting for a particular position or for finding friendship and companionship. Retrieval algorithms usually ranks the items by certain criteria and returns them ordered based on that criteria. In the case of people finding, these items are persons. If there are number of persons that match a given criteria is very large than, most users will not scan the entire list of results. That is why these search engines will usually return the top k persons (*top-k* ranking) from the list of all matched persons, where they were ranked in a descending order of some criteria of relative quality.

The main interest in this paper is the a biased search engine or machine learning model, in general, that is used in producing ranked lists can systematically further reduce the visibility or access to a already disadvantaged group [1], [2]. This disadvantaged group can correspond to a legally protected category of people such as ethic or racial minorities, people with disabilities or underrepresentation of a gender in a specific industry or position.

According to the authors of [3] a computer system is considered *biased* if *it systematically and unfairly discriminate(s) against certain individuals or groups of individuals in favor of others. A system discriminates unfairly if it denies an opportunity or a good or if it assigns an undesirable outcome to an individual or a group of individuals on grounds that are unreasonable or inappropriate.*

Still, what is important to note here is that *unfair discrimination alone does not give rise to bias unless it occurs systematically* and *systematic discrimination does not establish bias unless it is joined with an unfair outcome*. Individually, a rank is considered *good* if a person appears in the results and is present in the top k positions. The outcome is considered unfair if members of a protected group are systematically ranked lower than those of a privileged group. The search engine or ranking algorithm discriminates unfairly if the ranking decision is fully or heavily based on a feature that is found only in the privileged group of individuals. Previous research shows [4] that a machine learning model which is trained on a datasets that already incorporate a certain *preexisting bias* will include this bias and therefore produce biased results, which in turn can increase the disadvantage even further and reinforce existing bias.

In this paper we want to investigate whether re-ranking can increase the fairness in datasets which are known to be biased by using a single attribute for the protected group. Our goal is to reduce bias, while maintaining the utility of the retrieval. We propose a post-processing method that removes systematic bias by making sure that the representation of a given protected group cannot fall on n at top k rankings of the retrieval. This method should be efficient and should be easy to plug-in any existing search engine.

The paper is organized as follows: Section two presents the background of the field of anti-discrimination from the ranking perspective. Section three contains the details of our re-ranking mechanism. Section four presents the experimental setup and design. The results and discussion are presented in section five. The conclusion and final remarks are presented in the section six.

II. Background

Research towards analyzing anti-discrimination has been conducted only recently from the aspect of algorithms [5]. Some of the research is focused on discovering and measuring discrimination [6], [7], [2], while other attempts are directed at eliminating discrimination [1], [5], [8].

A. Fairness frameworks

In general, there are two basic frameworks that have been classified in recent studies on machine (algorithm based) discrimination:

- *individual fairness*, which is a requirement that individuals should be treated fairly [1]
- *group fairness*, which is a requirement that certain protected groups should be treated similarly to the population as whole [9]

This paper is focused on addressing the second framework.

B. Methods for fair ranking

There are a variety of fairness-aware algorithms that have been proposed to achieve group and/or individual fairness. Fairness constraints for several classification methods have been described in [8]. Disparate impact in data has been analyzed in [10], which explains how a certain group receives less benefits in ranking than a non-protected group. The solution that is offered there is by *tweaking* the ranking scores of members of the protected group to be more in line with the distribution of the non-protected group.

In [11], the authors consider three different approaches to deal with Naive Bayes models by changing the learning algorithm.

A statistical parity measure based on comparing the distributions of the members of the protected and non-protected groups in the top k rankings, for different values for k (e.g. 10, 20, 30), and then averaging these differences in some manner. The proposed averaging method is logarithmic, similarly to normalized discounted cumulative gain (NDCG), frequently used in information retrieval [12].

An interesting approach in [13] is aimed at detecting the source of bias in a given search engine. The proposed method tries to methodologically breakdown the points of bias in a given system, which can ultimately be addressed and modified.

In [14] a constraint ranking method is proposed, where constrain is a $k \times l$ matrix, where k is the length of the ranking and l is the number of groups (classes), where it is indicated the maximum elements of each group that can appear at any given position in the ranking. The goal of this approach is assign members with better qualifications in a higher position for each group separately.

Most recently, fairness-aware algorithms have been proposed for mostly supervised learning algorithms [15], [16].

C. Avoiding monotony

The general idea is that a given search engine needs to provide relevant, yet diverse results i.e. members from different classes (groups). This has been a widely studied in the field of information retrieval. The most common approach to this problem [17] is to consider the distances between elements and add penalty to the elements that are too similar to an element already appearing in a higher position. A similar approach is used in recommender systems for providing diversified results [18], [19].

Another approach in presented in [20], where a per-intent NDCG is evaluated for diversity. But this approach only evaluates a ranking rather then creating it.

The background on this field lead us to use a proportional re-ranking method, which will both increase the chances for inclusion of members of protected groups and will be efficient enough to be used in live production-grade information retrieval systems.

III. Proposed approach

Based on the background in the area we focused on creating a re-ranking algorithm that from a given ranked set O, will create a new ranked set of items R that should satisfy the following criteria:

1) The resulting set should fairly represent the protected group
2) The resulting set should contain the most qualified candidates

Our approach is presented as Algorithm 1. The algorithm, as input, takes the size k of the ranking that needs to be returned, the minimum proportion p of the member of the protected group in the ranking and the ranked set O, which is a ordered list and where for each item, there is an indicator of whether or not he/she is a member of the protected group and the initial ranking (quality) score.

The algorithm, first, initializes the result R as a empty list and two priority queues P_0 and P_1, where the members of the non-protected and protected groups are added, respectively. The list $rank$ is a utility list, which will contain the number of protected members at any given ranking. The algorithm then, greedily constructs the resulting list. While the result is not filled, the algorithm, checks for the given position $rank$, whether a protected member should be present at this position and takes the most *qualified* (according to the initial ranking, the member with the highest score) member of protected group (removing the member from P_1 in the process). In the case, where there is no protected member required at the current position, the algorithm, looks for the more qualified member between both groups and picks the most qualified one (removing it from the originating group P_0 or P_1).

IV. Experimental setup and design

In this section of we give the details of how the experimental evaluation was performed during our research.

A. Dataset

We conducted the experiments on the publicly available COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) dataset [6]. The dataset is an assessment tool for predicting recidivism based on a questionnaire of 137 questions. It is used in some parts of the judicial system in the United States. COMPAS has been accused of racial discrimination by producing a higher likelihood that African Americans will return to criminal behavior [6]. In the experiments we performed, we test a case in which we want to create a fair ranking for the top k people who are less likely to return to criminal behavior. Maybe the ranking, could be used, for instance, to consider people for pardoning or reduced sentences. We note that the African Americans and males are usually given a higher probability that they will return to committing criminal activities after release, compared to other groups, which is why we use those two categories as protected groups.

B. Baseline

In our experiments, for the dataset we generate a top k ranking with varying targets of minimum proportion of members of a protected group. To compare the results we use a

Algorithm 1 Re-ranking algorithm

```
function RE-RANK-MEMBERS(k, p, O)
    R ← empty list                                  ▷ init result
    P_0 ← empty queue                               ▷ init non-protected group
    P_1 ← empty queue                               ▷ init protected group
    rank ← empty list                               ▷ init ranks
    for i ← 1 to len(O) do
        if isprotected(O[i]) then
            add O[i] to P_1
        else
            add O[i] to P_0
        P_0 ← 1
    for i ← 1 to k do
        rank ← (i/(k/n))
    k_p ← 0                                         ▷ init number of protected
    k_n ← 0                                         ▷ init number of non-protected
    while k_p + k_n < k do
        if k_p < rank[k_p + k_n + 1] then
            k_p ← k_p + 1
            R[k_p + k_n] ← pop(P_1)
        else
            if q(peek(P_1)) >= q(peek(P_0)) then    ▷ qualities
                k_p ← k_p + 1
                R[k_p + k_n] ← pop(P_1)
            else
                k_n ← k_n + 1
                R[k_p + k_n] ← pop(P_0)
        return R
```

color-blind ranking. The color-blind ranking takes into account only the qualifications of the members, without considering group fairness as previously described. We use this as our baseline.

C. Metrics

In terms of metrics that we use to quantify the performance of our method, we use the Normalized Discounted Cumulative Gain (*NDCG*).

NDCG is a standardized measure in information retrieval [12], where a normalized weighted summation of the qualify of the elements in the ranking is considered, $\sum_{i=1}^{k} w_i q_i$ in which the weights have a logarithmic relation to the position i and it is represented in the following manner: $w_i = 1/log_2(i+1)$. The calculated value is normalized with *min-max* normalization, so that the maximum is 1.0.

V. Results and discussion

The results form our experiments are presented on Table I and Table II, which present the values of the re-ranking algorithm when the attribute on how the protected group is formed is *race* and *gender*, respectively.

The reported values presents the NDCG measure in case of the color-blind ranking i.e. a ranking using only the original qualities of the candidates and the re-ranked output, which is the the original ranking with *more* protected elements in the top k positions.

The experiments were conducted on multiple values for k as can be noted in the tables. The goal of this output is to see the difference in the NDCG measure at variable number of top k positions and see how the our re-ranking mechanism is effecting that measure on different scales.

The output from both NDCG measure for the color-blind and re-ranking algorithm are normalized with min-max normalization, so that we can compare the difference between the two.

As for the number of protected candidates in top rankings, according to the experiments performed in [21] on the same datasets we use the threshold at 50%, in the case of race and gender, which means that at least half of the candidates in the top rankings must be from the protected group. The protected group in terms of race is *African American* and in terms of gender is *male*.

In the case when the race being used an attribute to form the protected groups, from the experimental results at all scales, we can note that there is a small drop in the NDCG for lower numbers of k, which is very interesting, because in the color-blind ranking for the top k candidates the representation of the protected group is 12%. Our results mean, that by increasing the representation of the protected group in the top rankings, we only narrowly getting a reduction in NDCG (difference of 0.12).

But, that difference progresses as we increase k. We can note that as we increase k, the quality of the protected candidates in the top k rankings decreases compared to the non-protected ones.

In the case when the gender being used an attribute to form the protected groups, we can note that the drop is significantly sharper in at the lower ranks, but remains the same for all other values. This is because the original qualities of the candidates related to gender are more evenly distributed than that of race.

TABLE I
EXPERIMENTAL RESULTS FROM THE RE-RANKING ALGORITHM WHERE THE PROTECTED ATTRIBUTE IS *race*

k	*color-blind*	*with re-ranking*
1	1.0	1.0
5	1.0	0.99829
10	1.0	0.99187
50	1.0	0.98706
100	1.0	0.98401
200	1.0	0.97923
500	1.0	0.97743
1000	1.0	0.97992

VI. Conclusion

In this paper, we proposed a novel way of re-ranking candidates/elements based on an attribute which defines them as protected or not. This approach tries to address fairness in ranking, so that candidates that belong in certain disadvantaged groups can have a better chance of appearing in the top rankings. The logic behind the idea was that

TABLE II
EXPERIMENTAL RESULTS FROM THE RE-RANKING ALGORITHM WHERE THE PROTECTED ATTRIBUTE IS *gender*

k	*color-blind*	*with re-ranking*
1	1.0	1.0
5	1.0	0.98871
10	1.0	0.98231
50	1.0	0.98650
100	1.0	0.98757
200	1.0	0.98566
500	1.0	0.98532
1000	1.0	0.98249

some of the seemingly objective ranking methods discriminate certain types of groups. Hence, our approach tried to offer a solution to the problem. We conducted the experiments on the COMPAS dataset and it can be noted that our re-ranking method introduces as *small* reduction in NDCG, but increases the representativeness of the protected groups for 4 folds from the original rankings.

In future, we would like to also experiment with other different datasets and evaluate our algorithm there as well. We would also like include utility metrics to better calculate the loss of *quality* in the rankings and tuning the by applying machine learning methods.

ACKNOWLEDGMENT

The authors would like to thank the support of the Faculty of Computer Science and Engineering through the project DLMIA - Deep learning for medical image analysis.

REFERENCES

[1] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel, "Fairness through awareness," in *Proceedings of the 3rd innovations in theoretical computer science conference*. ACM, 2012, pp. 214–226.

[2] D. Pedreshi, S. Ruggieri, and F. Turini, "Discrimination-aware data mining," in *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2008, pp. 560–568.

[3] B. Friedman and H. Nissenbaum, "Bias in computer systems," *ACM Transactions on Information Systems (TOIS)*, vol. 14, no. 3, pp. 330–347, 1996.

[4] J. Bradley, "Weapons of math destruction: How big data increases inequality and threatens democracy," *Perspectives on Science and Christian Faith*, vol. 69, no. 1, pp. 54–56, 2017.

[5] S. Hajian, F. Bonchi, and C. Castillo, "Algorithmic bias: From discrimination discovery to fairness-aware data mining," in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. ACM, 2016, pp. 2125–2126.

[6] J. Angwin, J. Larson, S. Mattu, and L. Kirchner, "Machine bias," *ProPublica, May*, vol. 23, 2016.

[7] F. Bonchi, S. Hajian, B. Mishra, and D. Ramazzotti, "Exposing the probabilistic causal structure of discrimination," *International Journal of Data Science and Analytics*, vol. 3, no. 1, pp. 1–21, 2017.

[8] M. B. Zafar, I. Valera, M. Gomez Rodriguez, and K. P. Gummadi, "Fairness constraints: Mechanisms for fair classification," *arXiv preprint arXiv:1507.05259*, 2017.

[9] D. Pedreschi, S. Ruggieri, and F. Turini, "Measuring discrimination in socially-sensitive decision records," in *Proceedings of the 2009 SIAM International Conference on Data Mining*. SIAM, 2009, pp. 581–592.

[10] M. Feldman, S. A. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian, "Certifying and removing disparate impact," in *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2015, pp. 259–268.

[11] T. Calders and S. Verwer, "Three naive bayes approaches for discrimination-free classification," *Data Mining and Knowledge Discovery*, vol. 21, no. 2, pp. 277–292, 2010.

[12] K. Järvelin and J. Kekäläinen, "Cumulated gain-based evaluation of ir techniques," *ACM Transactions on Information Systems (TOIS)*, vol. 20, no. 4, pp. 422–446, 2002.

[13] J. Kulshrestha, M. Eslami, J. Messias, M. B. Zafar, S. Ghosh, K. P. Gummadi, and K. Karahalios, "Quantifying search bias: Investigating sources of bias for political searches in social media," in *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*. ACM, 2017, pp. 417–432.

[14] L. E. Celis, D. Straszak, and N. K. Vishnoi, "Ranking with fairness constraints," *arXiv preprint arXiv:1704.06840*, 2017.

[15] L. E. Celis, A. Deshpande, T. Kathuria, and N. K. Vishnoi, "How to be fair and diverse?" *arXiv preprint arXiv:1610.07183*, 2016.

[16] S. Corbett-Davies, E. Pierson, A. Feller, S. Goel, and A. Huq, "Algorithmic decision making and the cost of fairness," in *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2017, pp. 797–806.

[17] J. Carbonell and J. Goldstein, "The use of mmr, diversity-based reranking for reordering documents and producing summaries," in *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 1998, pp. 335–336.

[18] S. Channamsetty and M. D. Ekstrand, "Recommender response to diversity and popularity bias in user profiles," 2017.

[19] M. Kunaver and T. Požrl, "Diversity in recommender systems–a survey," *Knowledge-Based Systems*, vol. 123, pp. 154–162, 2017.

[20] T. Sakai and R. Song, "Evaluating diversified search results using per-intent graded relevance," in *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*. ACM, 2011, pp. 1043–1052.

[21] M. Zehlike, F. Bonchi, C. Castillo, S. Hajian, M. Megahed, and R. Baeza-Yates, "Fa* ir: A fair top-k ranking algorithm," in *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*. ACM, 2017, pp. 1569–1578.

An improved elephant herding optimization by balancing local and global search for continuous optimization

Huseyin Hakli
Department of Computer Engineering
Necmettin Erbakan University
Konya, Turkey
e-mail: hhakli@konya.edu.tr

***Abstract*— Elephant herding optimization algorithm (EHO), one of the recent bio-inspired optimization technique, mimics the herding behavior of elephants. In the EHO algorithm, the whole population is divided the some clans and also has two main process: clan updating and separating. The elephants are updated using its current position and matriarch which is the best elephant in the clan updating operator. Due to this, EHO is good at local search and rapid convergence, but it has insufficient on the global search. To overcome this problem, EHO is improved by balancing global and local search (GL-EHO) and inspiring the artificial bee colony and particle swarm optimization. While the search equation of ABC is used for the global search, the search mechanism of PSO is originated to strengthen the local search. Firstly these two mechanisms are implemented to basic EHO separately (L-EHO and G-EHO) and then both are used (GL-EHO). The proposed method is compared with basic EHO on the fourteen benchmark functions to investigate the success of the GL-EHO. Experimental results show that the proposed method is more successful and robust than basic EHO by means of balancing between global and local search.**

Keywords- elephant herding optimization; continuous optimization; local search; global search

I. INTRODUCTION

Many bio-inspired algorithms which simulate the behavior of the living things were proposed in last two decades. Due to fact that the conventional and mathematical methods cannot be solved the real world optimization problems effectively, bio-inspired algorithms which obtain the satisfactory solutions within a reasonable time become more widespread in rapidly. Some of the mostly known bio-inspired algorithms are artificial bee colony (ABC) [1], particle swarm optimization (PSO) [2] and ant colony optimization [3]. In recent years, many algorithms have been developed such as lion optimization algorithm [4], grey wolf optimizer [5], artificial algae algorithm [6], elephant herding optimization algorithm (EHO) [7].

A swarm based technique, EHO, was proposed by inspiration of herding behavior of elephant groups for solving continuous optimization problems. The elephants in nature have some social properties. Two of these properties are the elephants belonging to different clans live together under the leadership of a matriarch, and male elephants tend to leave their family group [8]. EHO has two main operator which mimic these two behavior: clan updating operator and separating operator. The best elephant in the clan named matriarch. The elephants are updated using its current position and matriarch in the clan updating operator. Afterward, the worst elephant individual is replaced by separating operator to improve the population diversity.

Due to the fact that EHO is newly-emerging algorithm, there are several studies in the literature. The performance of basic EHO was analyzed on the CEC2013 benchmark functions by Tuba et. al [9] and the results show that EHO has good characteristics. An elephant herding optimization based tuning approach was proposed for obtaining desired parameters of PID controller [10]. Tuba et. al. implemented the elephant herding optimization to solve the multilevel thresholding problem [11]. In EHO algorithm, the fittest elephant in the clan is directly updated the position of center of the clan. Instead of this, Parashar et al. proposed a modified elephant herding optimization to add the influence of center of clan to the previous position of the fittest elephant to find its new position [12]. In other study [13], two improvements were added the elephant herding optimization. One of them is updating the position of matriarch elephants around the current best position as same in the Parashar et al' study. The other one is the new babies are generated the position near to matriarch of the respective clan in the separating operator instead of the randomly determining in the search range.

Due to the positions of the elephants are updated in accordance with the best elephants in the clan, the basic EHO is good in terms of search of exploitation but it has a weakness of the exploration ability. To overcome this problem and compromise between local and global search, the elephant herding optimization was improved by inspiring the ABC and PSO algorithm. While the ABC algorithm is superior on the global search [14], PSO performs local search very well [15]. Therefore, the search mechanism of basic EHO was modified in three different forms to examine local and global search ability of proposed techniques: local search EHO (L-EHO), global search EHO (G-EHO), global and local search EHO (GL-EHO). The basic EHO and proposed approaches were compared on the 14 benchmark functions which have different characteristics.

The following sections of the study are as follows: Section 2 covers the basic EHO algorithm. The proposed approaches are explained in the Section 3. Experimental results are examined in Section 4, and conclusions are presented in Section 5.

II. BASIC ELEPHANT HERDING OPTIMIZATION

Being inspired from the elephant clans, EHO was proposed by Wang et al. in 2015 [7]. In the EHO algorithm, each elephant is belonging to clan and represents the candidate solution. There are two main process in EHO algorithm: clan updating and separating. Each clan has a matriarch which is fittest elephant in the clan. In the clan updating operator, the positions of the elephants are updated with its current position and the matriarch as seen in (1).

$$X^{j}_{new,ci} = X^{j}_{ci} + \alpha \times (X_{best,ci} - X^{j}_{ci}) \times r \quad (1)$$

where X^{j}_{ci} is the position of elephant j in clan ci, $X_{best,ci}$ is the position of the matriarch in clan ci and $X^{j}_{new,ci}$ represents the new position of X^{j}_{ci}. α is a scale factor which determines the influence of matriarch and r is a random number generating uniform distribution in the range [0,1].

When the position of matriarch is updated by (1), there will be no change in its position. Therefore, Equation (2) is used for updating the position of matriarchs in the clans.

$$X^{j}_{new,ci} = \beta \times X_{center,ci} \quad (2)$$

where $X_{center,ci}$ represents mean/average position of clan ci and β is a factor in the range [0,1].

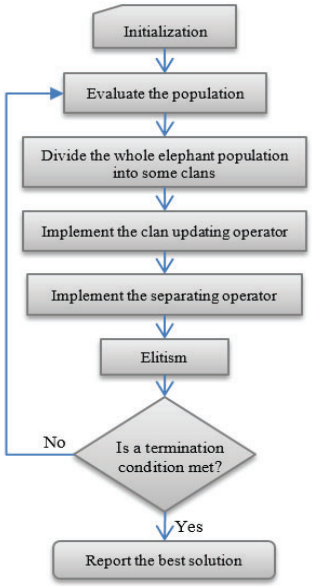


Fig. 1. The flowchart of basic EHO algorithm.

Male elephants will leave their clans, when they grow up and reach puberty. This behavior named as separating operator in the EHO algorithm. The elephant which has a worst fitness value in the clan is randomly generated in the search space with (3).

$$X_{worst,ci} = X_{\min} + (X_{\max} - X_{\min} + 1) \times rand \quad (3)$$

where $X_{worst,ci}$ is the worst individual in the clan ci. $X_{\min}$ and $X_{\max}$ are lower and upper bound in the search space.

The flowchart of basic EHO is given in Fig.1. Detailed information about the EHO, please refer to [7, 8].

III. THE PROPOSED APPROACHES FOR EHO

The most important problem of the optimization techniques is providing the compromise between global and local search. The basic EHO algorithm has an ability of local search owing to its solution search equation. Due to rapidly convergence, the EHO gets stuck the local minima and loses the diversity of the population.

To strengthen the performance of basic EHO, its solution search equation is improved with effective search mechanisms. While the search phenomenon of ABC is used for the global search, the search equation of PSO is originated to enhance the local search. Firstly these two mechanisms are implemented to basic EHO separately (L-EHO and G-EHO) and then both are used (GL-EHO). In addition, Equation 2 is ignored for the proposed approaches based on EHO, due to inconsistency of mean position in the clan.

In the solution search equation of ABC algorithm, one dimension is randomly selected and this dimension of new individual is updated with its current position and neighbor individual as seen in (4).

$$X^{j,RD}_{new,ci} = X^{j,RD}_{ci} + \Phi \times (X^{j,RD}_{ci} - X^{k,RD}_{ci}) \quad (4)$$

where RD is a randomly selected dimension, $X^{k,RD}_{ci}$ represents RD^{th} dimension of elephant k in clan ci. k and j are not equal to each other. Φ is a random number and scaling factor in the range of [-1,1]. In the proposed approach named G-EHO, Equation (4) is implemented instead of Equation (1).

```
for i=1:population size
  if rand < 0.5
        Generate new elephant with Equation (4)
  else
        Generate new elephant with Equation (5)
  end
end
```

Fig. 2. The search mechanism of proposed GL-EHO algorithm.

Equation (5) is adapted to in the place of Equation (1) by inspired the search mechanism of the PSO algorithm. This approach named L-EHO which strengthens the ability of the local search of basic EHO.

$$X_{new,ci}^{j} = X_{ci}^{j} + c_1 \times rand_1 \times (X_{best,ci} - X_{ci}^{j}) + c_2 \times rand_2 \times (X_{best} - X_{ci}^{j}) \quad (5)$$

where X_{best} is the fittest individual in the whole population and $X_{best,ci}$ represents the elephant which has best fitness value in the clan *ci*. In addition, while c_1 and c_2 are acceleration coefficients, $rand_1$ and $rand_2$ are random number in the range [0,1].

In the last proposed approach GL-EHO, Equation (4) and Equation (5) are applied to basic EHO by a selection parameter which is 0.5. If the random number is smaller than this parameter Equation (4) is selected, in the other situation the new individual is generated with the Equation (5). The search mechanism of GL-EHO algorithm which is using instead of Equation (1) in the basic EHO is seen Fig. 2.

TABLE I. BENCHMARK FUNCTIONS(C: CHARACTERISTIC, U: UNIMODAL, M: MULTIMODAL, S: SHIFTED)

Range	C	Function	Formulation
[-100, 100]	U	Sphere	$f_1 = \sum_{i=1}^{n} x_i^2$
[-100, 100]	U	Elliptic	$f_2 = \sum_{i=1}^{n} (10^6)^{(i-1)/(n-1)} x_i^2$
[-10, 10]	U	Rosenbrock	$f_3 = \sum_{i=1}^{n-1} \left[100\left(x_{i+1} - x_i^2\right)^2 + \left(x_i - 1\right)^2 \right]$
[-5.12, 5.12]	M	Rastrigin	$f_4 = \sum_{i=1}^{n} \left[x_i^2 - 10\cos(2\pi x_i) + 10 \right]$
[-600, 600]	M	Griewank	$f_5 = \frac{1}{4000} \sum_{i=1}^{n} x_i^2 - \prod_{i=1}^{n} \cos\left(\frac{x_i}{\sqrt{i}} \right) + 1$
[-32, 32]	M	Ackley	$f_6 = -20\exp\left\{ -0.2\sqrt{\frac{1}{D}\sum_{i=1}^{D} x_i^2} \right\} - \exp\left\{ \frac{1}{D}\sum_{i=1}^{D} \cos(2\pi x_i) \right\} + 20 + e$
[-10, 10]	U	SumSquare	$f_7 = \sum_{i=1}^{n} i x_i^2$
[-10, 10]	U	Schwefel 2.22	$f_8 = \sum_{i=1}^{n} \lvert x_i \rvert + \prod_{i=1}^{n} \lvert x_i \rvert$
[-10, 10]	M	Alpine	$f_9 = \sum_{i=1}^{n} \lvert x_i \cdot \sin(x_i) + 0.1 \cdot x_i \rvert$
[-100, 100]	S	Shifted Sphere	$f_{10} = \sum_{i=1}^{n} z_i^2 \quad z = x - o$
[-5.12, 5.12]	S	Shifted Rastrigin	$f_{11}(\vec{X}) = \sum_{i=1}^{n} \left[z_i^2 - 10\cos(2\pi z_i) + 10 \right] \quad z = x - o$
[-600, 600]	S	Shifted Griewank	$f_{12}(\vec{X}) = \frac{1}{4000} \sum_{i=1}^{n} z_i^2 - \prod_{i=1}^{n} \cos\left(\frac{z_i}{\sqrt{i}} \right) + 1 \quad z = x - o$
[-32, 32]	S	Shifted Ackley	$f_{13}(\vec{X}) = -20\exp\left\{ -0.2\sqrt{\frac{1}{n}\sum_{i=1}^{n} z_i^2} \right\} - \exp\left\{ \frac{1}{n}\sum_{i=1}^{n} \cos(2\pi z_i) \right\}$ $z = x - o$
[-10, 10]	S	Shifted Alpine	$f_{14}(\vec{X}) = \sum_{i=1}^{n} \lvert z_i \cdot \sin(z_i) + 0.1 \cdot z_i \rvert \quad z = x - o$

IV. EXPERIMENTAL RESULTS

To investigate the performance of proposed approaches, the basic EHO, L-EHO, G-EHO and GL-EHO are tested on the fourteen benchmark functions which have different characteristics. The range, characteristic, name and formulation of these functions are given in Table 1. In Table 1, there are five unimodal, four multimodal and five shifted functions.

TABLE II. A COMPARISON OF THE EHO AND PROPOSED APPROACHES

Func. No. (Benchmark)		Basic EHO	L-EHO	G-EHO	GL_EHO
F1	Mean	9.53E-06	3.72E-29	2.47E-03	**3.05E-51**
	Std.Dev.	9.26E-07	2.04E-28	2.78E-03	**1.57E-50**
F2	Mean	2.08E-01	2.83E-08	1.64E+03	**1.24E-46**
	Std.Dev.	2.94E-02	1.44E-07	3.93E+03	**6.79E-46**
F3	Mean	2.86E+01	1.04E+03	6.27E+01	**4.03E+00**
	Std.Dev.	8.77E-03	3.04E+03	2.91E+01	**4.98E+00**
F4	Mean	9.59E-06	1.13E+01	**4.27E-09**	9.95E-02
	Std.Dev.	1.06E-06	9.69E+00	**2.34E-08**	3.04E-01
F5	Mean	**1.68E-05**	3.19E-02	3.67E-02	4.34E-02
	Std.Dev.	**2.68E-06**	2.95E-02	3.71E-02	3.58E-02
F6	Mean	7.47E-04	7.53E+00	1.27E-02	**4.40E-10**
	Std.Dev.	4.06E-05	5.65E+00	9.11E-03	**2.01E-09**
F7	Mean	1.65E-06	1.43E-14	1.25E-04	**8.78E-52**
	Std.Dev.	2.77E-07	4.08E-14	2.18E-04	**4.79E-51**
F8	Mean	1.57E-03	**3.54E-12**	1.30E-03	9.38E-06
	Std.Dev.	1.19E-04	**1.36E-11**	1.46E-03	4.36E-05
F9	Mean	1.53E-04	**5.30E-09**	6.41E-04	9.11E-07
	Std.Dev.	7.34E-06	**2.78E-08**	1.17E-03	2.61E-06
F10	Mean	3.77E+04	**0.00E+00**	4.46E-03	1.33E-27
	Std.Dev.	2.93E+03	**0.00E+00**	4.12E-03	6.21E-27
F11	Mean	3.16E+02	1.56E+00	9.47E-01	**1.66E-01**
	Std.Dev.	1.87E+01	1.37E+00	9.83E-01	**3.77E-01**
F12	Mean	2.56E+02	9.45E-02	**5.42E-02**	9.36E-02
	Std.Dev.	3.95E+01	5.71E-02	**3.42E-02**	5.10E-02
F13	Mean	1.97E+01	5.11E+00	2.00E-02	**2.19E-11**
	Std.Dev.	2.85E-01	6.35E+00	1.02E-02	**7.41E-11**
F14	Mean	4.81E+01	2.50E-03	1.78E-02	**6.26E-04**
	Std.Dev.	5.96E+00	6.50E-03	2.27E-02	**3.43E-03**
Friedman Rank		3.07	2.42	2.85	**1.64**
Corrected Rank		4	2	3	**1**

In the experiments are performed in this study, the population size is 50 and the number of elephant in each clan is set 10. So, the number of clan is 5 and α is set 0.5 for all methods. For the basic EHO, β is determined as 0.1. The acceleration coefficients c_1 and c_2 are equal to each other and are 1.5 for the L-EHO and GL-EHO.

The dimension of functions is determined as 30 and termination condition is used the maximum number of function evaluations which is 1.5E+05. All methods are run 30 times for the benchmark functions under the mentioned

conditions. Table 2 contains a comparison of the basic EHO and the proposed approaches. The mean result and standard deviation values of 30 runs for benchmark functions listed in Table 1 are given in Table 2. The best mean result and its standard deviations of the benchmark functions obtained by the algorithms are shown in bold. In addition, Table 2 shows the result of the Friedman rank test for the methods.

The basic EHO has a better performance than the other algorithms for the only Griewank function when the examining experimental results given Table 2. L-EHO has a good ability on the local search and, so it is more successful than the basic EHO on the unimodal functions (except Rosenbrock). Also, L-EHO obtains the optimal solution for the Shifted Sphere function. For the Rastrigin function, the best mean solution is obtained by the G-EHO, but it has an underachieving performance on the multimodal functions. One of these reasons is inconsistency of the individuals owing to make only global search. GL-EHO has a higher performance than the other methods on eight functions. For the shifted functions, while the GL-EHO obtains the better results, the basic EHO drops behind much due to losing diversification of the population rapidly.

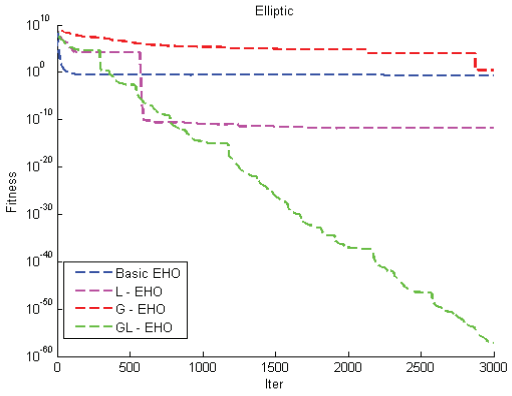


Fig. 3. The convergence graph of methods for the Elliptic function.

To evaluate the methods clearly, the Friedman rank test are implemented to all results in Table 2. With respect to the Friedman test, GL-EHO approach is first rank between the methods. In addition, the basic EHO is last rank and L-EHO is located in the second rank.

In order to investigate convergence performances of basic EHO and proposed approaches, convergence graphs of the methods on the Elliptic (unimodal), Ackley (multimodal) and Shifted Sphere functions are respectively given in the Fig. 3, 4 and 5. When examining the convergence graph of Elliptic function in Fig. 3, GL-EHO has a superior performance and continues the improving solution until the end of iteration. Although the basic EHO converges quickly, it has a stagnation in a short time due to the decreasing the diversification of the population.

For the Ackley function, the basic EHO has early convergence and then gets stuck the local minima in Fig. 4. L-EHO converges very slowly, it gets rid of the local minimum towards the end of the iteration and improves the solution. G-EHO has a better convergence performance on the Ackley function than the unimodal Elliptic function but this performance is not enough to pass even the basic EHO. GL-EHO has an outperformance on the convergence of Elliptic and Ackley functions thanks to balancing between global and local search.

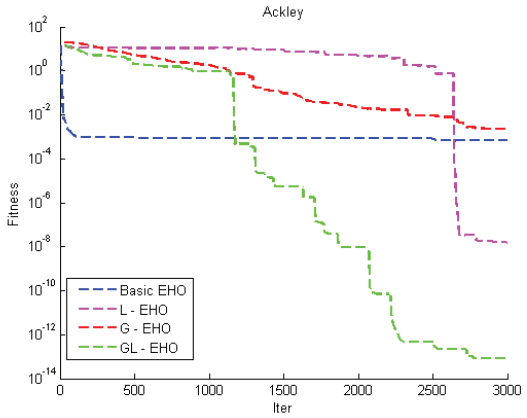


Fig. 4. The convergence graph of methods for the Ackley function.

As seen in the Fig. 5, the basic EHO's poor convergence performance on the Shifted Sphere function attracts the attention. GL-EHO and L-EHO has good convergence performance and they obtain the optimum solution for this function.

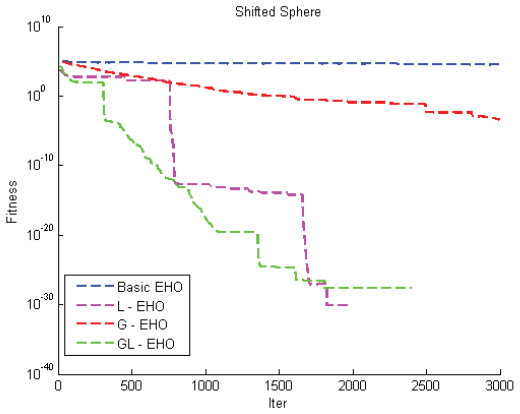


Fig. 5. The convergence graph of methods for Shifted Sphere function.

When the all experiments are evaluating, GL-EHO outperforms the basic EHO and other proposed approaches of EHO in terms of robustness and solution quality. In addition it has a good compromise between local and global search.

V. CONCLUSION

To solve the continuous optimization problem effectively, the methods need to balancing on the exploration and exploitation. The basic EHO, newly-emerging optimization technique, is good at local search. Despite an ability of local search, EHO has a weakness on the exploration. To overcome this problem and show the effect of the global and local search on the methods, three new approaches are proposed. While the search mechanism of ABC is used for the exploration, the search equation of PSO is originated to enhance the exploitation. The proposed approaches and EHO were compared on the fourteen benchmark functions have different characteristics. The GL-EHO is more successful and robust than the other algorithms in accordance with the experimental results. The usage of only local search (L-EHO) or only global search (G-EHO) for the optimization methods will reduce the success on the problems of different characteristics and will not be enough to solve the problems effectively. Therefore, the GL-EHO ranks first in the Friedman test made for the experimental results of methods by means of compromise between global and local search.

ACKNOWLEDGMENT

This study has been supported by Scientific Research Project of Necmettin Erbakan University.

REFERENCES

[1] D. Karaboga, An idea based on honey bee swarm for numerical optimization., in, Erciyes University, Kayseri, Turkey, Tech. Rep., TR06, 2005.

[2] J. Kennedy, R. Eberhart, Particle Swarm Optimization, in: Proceedings of the Sixth International Symposium on Micro Machine and Human Science, Nagoya, Japan, 1995, pp. 39-43.

[3] M. Dorigo, G.D. Caro, Ant colony optimization: a new meta-heuristic, in: Proceedings of the 1999 Congress on Evolutionary Computation, Washington, DC, 1999, pp. 1470–1477.

[4] M. Yazdani, F. Jolai, Lion Optimization Algorithm (LOA): A nature-inspired metaheuristic algorithm, J Comput Des Eng, 3 (2016) 24-36.

[5] S. Mirjalili, S.M. Mirjalili, A. Lewis, Grey Wolf Optimizer, Adv Eng Softw, 69 (2014) 46-61.

[6] S.A. Uymaz, G. Tezel, E. Yel, Artificial algae algorithm (AAA) for nonlinear global optimization, Appl Soft Comput, 31 (2015) 153-171.

[7] G.G. Wang, S. Deb, L.D. Coelho, Elephant Herding Optimization, 2015 3rd International Symposium on Computational and Business Intelligence (Iscbi 2015), (2015) 1-5.

[8] G.G. Wang, S. Deb, X.Z. Gao, L.D. Coelho, A new metaheuristic optimisation algorithm motivated by elephant herding behaviour, Int J Bio-Inspir Com, 8 (2016) 394-409.

[9] V. Tuba, M. Beko, M. Tuba, Performance of Elephant Herding Optimization Algorithm on CEC 2013 real parameter single objective optimization, WSEAS TRANSACTIONS on SYSTEMS, 16 (2017) 100-105.

[10] S. Gupta, V.P. Singh, S.P. Singh, T. Prakash, N.S. Rathore, Elephant herding optimization based PID controller tuning, International Journal of Advanced Technology and Engineering Exploration, 3(24) (2016) 194-198.

[11] E. Tuba, A. Alihodzic, M. Tuba, Multilevel Image Thresholding Using Elephant Herding Optimization Algorithm, in: 14th International Conference on Engineering of Modern Electric Systems (EMES), Oradea, Romania, 2017.

[12] S. Parashar, A. Swarnkar, K.R. Niazi, N. Gupta, A modified elephant herding optimization for economic generation co-ordination of DERs and BESS in grid connected microgrid, J Eng-Joe, (2017).

[13] N.K. Meena, S. Parashar, A. Swarnkar, N. Gupta, K.R. Niazi, Improved Elephant Herding Optimization for Multiobjective DER Accommodation in Distribution Systems, IEEE Transactions on Industrial Informatics, PP (2017).

[14] M.S. Kiran, H. Hakli, M. Gunduz, H. Uguz, Artificial bee colony algorithm with variable search strategy for continuous optimization, Inform Sciences, 300 (2015) 140-157.

[15] S. Devi, D.G. Jadhav, S.S. Pattnaik, PSO Based Memetic Algorithm for Unimodal and Multimodal Function Optimization, Lect Notes Comput Sc, 7076 (2011) 127-134.

Microservice based architecture for the genetic algorithm

Evgenija Stevanoska, Kristijan Spirovski, Goran Petkovski,Boro Jakimovski, Goran Velinov
Faculty of Computer Science and Engineering
University Sts Cyril and Methodius
Skopje, Macedonia
{evgenija.stevanoska, kristijan.spirovski, goran.petkovski}@students.finki.ukim.mk, {boro.jakimovski, goran.velinov}@finki.ukim.mk

***Abstract*—Microservice architecture is becoming more popular and more frequently used, mainly because of its numerous advantages over monolith approach. Namely, the developed systems nowadays need more agile distribution of the processing power, and due to their size, a way to deploy, maintain and test individual components separately. Each algorithm/software composed of individual and independently executable parts that do not share many parameters are good candidates for a solution based on a microservice architecture. This paper presents a microservice approach when building an architecture for the genetic algorithm. We identified eight independent parts of the genetic algorithm, and each one is represented as a microservice. This design leads to a solution that has low coupling and high cohesion. The advantages of this approach include distributing the computations on more physical locations, and furthermore, scaling only the parts of the system which require more performance (which means need large processing power or are frequently executed).**

***Keywords*—microservice, genetic algorithm, Spring Boot, Spring Cloud**

I. Introduction

The large number of NP-hard problems that can't be solved in polynomial time imposes the need to use an alternate approach for finding near optimal solutions to the optimization problems in efficient time. Popular choices for that task are meta-heuristic optimization algorithms. One of the oldest and still very popular meta-heuristic algorithm is the genetic algorithm (GA), which models the evolution of the population by applying mutation and crossover to the individuals in the population. The nature of the genetic algorithm offers easy identification of the independent parts and their parallel and distributed execution. These characteristics make the genetic algorithm suitable for a microservice based implementation.

A microservice is an independent service which is responsible for one functionality and it collaborates with other microservices using a strongly defined interface, often using messages with a predefined structure. Microservice architecture is an approach for developing applications using small, independent services. Each microservice is executed as a small, independent process. Nowadays, the microservice architecture becomes more popular, and the advantages over the monolith approach are obvious.

Microservice based implementation of the genetic algorithm allows individual components to be executed on more physical locations. Furthermore, not all components have the same processing needs. This approach allows us to identify the most computationally extensive components and assign them more processing power, which leads to easy scalability, where each component has only the resources needed for optimal execution. The components of the algorithm can be independently executed. The nature of the microservice architecture allows parallel execution of the independent parts.

This paper presents a microservice based implementation of the genetic algorithm. The goal is to show that GA is suitable for a microservice approach and to give a prototype of the implementation (which is defined by the choice of independent components). We implemented eight microservices that communicate through messages (REST API). The microservices are implemented in Spring Framework using Spring Boot and Spring Cloud libraries, and they are deployed to Pivotal.

The paper is organized as follows: Section II gives an explanation and a definition to microservoces and microservice architecture, and it also lists the advantages over the monolith approach. Section III contains a description and pseudo-code of the genetic algorithm. The proposed implementation is given in Section IV, which contains information about the micriservices, and also a list of the used technologies. Finally, the last section gives the conclusion.

II. Microservices

The term microservices was first introduced in 2011 at an architecture workshop, as a way of describing the common idea of the members of the workshop for a new software architecture. Since then, they are implemented in many popular and commonly used software solutions. For example, Netflix uses microservice architecture known as Grained SOA [1].

Microservices are small processes that can be independently deployed, scaled and tested. Each microservice has only one functionality and responsibility, which means it can be easily updated and understood by a programmer [2]. Functionalities that are not strongly connected are modeled using different microservices, thus, the microservice architecture supports the high cohesion-loose coupling paradigm.

Microservice architecture is an application whose components (modules) are implemented using microservices. The

microservices communicate using predefined interface (often using messaging concept). The microservice approach has numerous advantages: [3]

- Each microservice implements a limited number of functionalities, leading to a small codebase, and thus low probability of a bug in the code. Furthermore, if a bug exists, its scope is limited to the microservice, which leads to an easy identification of the part of the code causing the bug. The fact that microservices are independent means that they can be easily tested, and their scope easily understandable even if they are isolated from the system, leading to a higher code reusability. Usually, each microservice is developed and deployed by one team, and the team is responsible for maintaining it.
- Each microservice can be easily replaced while other services work normally. If an application based on a microservice architecture needs to be updated, the process can be performed by gradually replacing each microservice (or furthermore, uploading the new version alongside with the old one), and then updating all the microservices that are dependent of the replaced microservice.
- Consequence of the previous point is the fact that the replacement of one microservice doesn't cause down-time to the whole system. A reboot is needed only on the microservices of the replaced module.
- Microservice architecture scaling doesn't mean doubling all components' instances. Namely, it offers the opportunity to the developers to monitor the load to each microservice, and to scale only the microservices that need higher processing power.
- The only limitation that microservices have is the technology used for communication. Other than that, the developers are free to choose the optimal resources for a microservice, which includes the framework, the implementation language etc.

The microservice approach has few disadvantages, caused mainly by the message-based communication between the microservices. They include:

- Sending messages through the network is slower than in-memory calls. To achieve comparable performances with monolith architecture, the number of calls through the network should be limited.
- Special attention should be paid on security of the communication. The messages are usually send in json or xml format, which can be easily intercepted. To maximize the security, the communication should be encrypted.

These facts are highlighting the advantages of the microservice architecture over a monolith one, making the former one more popular and commonly used. Namely, monolith systems are often enormous, which makes them hard to maintain. Finding a bug in such system takes longer because of the inability to easily identify the part of the code causing the bug. Attention should be paid on dependencies between different libraries, because adding or updating a library to a newer version could cause inconsistency in the system. A change in one part of the system causes reboot on the whole application, which means greater downtime. Furthermore, the scalability of a monolith application is limited to making more instances of the whole application, and balancing the load between them. Obviously, this approach is not efficient when the traffic is increased for one part of the application. Also, monolith architecture uses one language for the whole application, which eliminates the opportunity to choose the most convenient language and framework for each functionality.

Compared to the standard service oriented architectures, Microservices show a large number of common parts. Namely, they are both relying on as the main component, but they vary in terms of service characteristics (for example, service size and functionalities each service models).

The advantages of microservice architecture over the SOA include:

- Microservices can operate and be deployed independently (unlike SOA), which makes it easier to deploy a new version of the application and scale only the necessary parts.
- they are better at fault tolerance. If a microservice fails, it only affects the part where the microservice is used, all other micriservices function independently. In SOA, Enterprise Service Bus (ESB) becomes a single point of failure.
- In SOA the data is shared and accessible from all services, making the services tightly coupled.
- Containers can be easily used with microservices (and not with SOA).

These facts just confirm that microservice architecture is gaining its rightful popularity, and is a good and commonly used architectural choice in newly developed systems.

III. Genetic Algorithm

GA is a common heuristic optimization method based on the principle of natural evolution. In nature, the individuals evolve following the principles of natural selection and survival of the fittest. Theoretical basis of the genetic algorithm are introduced by Holland [4]. GA follows this principles of evolution in order to find a near optimal solution to an optimization problem.

A. Description of the Algorithm

Genetic algorithms encodes a potential solution to a specific problem on a simple chromosome-like data structure and applies recombination operators to these structures so as to preserve critical information. An implementation of a GA begins with an initial population of (typically random) chromosomes. Each chromosome is evaluated using a fitness function that is specific to the problem being solved. Then, based on the fitness values, the GA allocates reproductive opportunities in a way that the ones that represent a better solution are given more chances to reproduce than the chromosomes that represent poorer solutions. On the chosen subset of chromosomes (to be part of the next generation) following operators for simulating the evolution process through generations are used in order to create the next generation:

- **Selection**. Every GA uses a selection mechanism to decide which individuals will comprise the mating pool which will form the basis of the next generation and will be used to generate new offspring. When dealing with a problem of minimization, the individuals with lower fitness values will have a better chance to be selected for the mating pool [5].
- **Crossover**. Crossover is the process of randomly selecting two parent chromosomes from the mating pool, exchanging genetic material between them and producing new offspring chromosomes. This process tends to increase the quality of the populations and force convergence.
- **Mutation**. This operator represents a small change on the genetic material of a chromosome. It is performed by changing one or more components of a chromosome. Mutation is used to improve the diversity of the chromosomes in the population. It lowers the probability of the algorithm being trapped in local optima, hence it plays an important role in any GA.

B. Parameters of the Genetic Algorithm

As previously mentioned, the genetic algorithm takes few input parameters which have a direct impact of the GA's behavior and the quality of the found solutions.

The population size and the number of generations. These two parameters influence the trade-off between the quality of the found solution and the execution time. Increasing the number of chromosomes in the population is proportionally increasing the size of the search space that the algorithm covers, and thus it increases the probability of finding a better solution.

Crossover probability. This parameter controls the portion of the population which is directly inherited from the last generation, (and thus the portion of the newly formed individuals). Usually, 20% of the individuals go to the next generation unchanged, and the remaining 80% of the population are filled by applying the crossover operators on the best individuals from the previous generation.

Mutation probability. As previously mentioned, mutation is a mechanism for avoiding local optima traps. With certain probability, each newly created individual is subject to mutation. Usually, this parameter is in the scale of 10%.

Dimension of the search space. This parameter depends only on the concrete application of the algorithm (the optimization problem that it tries to solve) and the defined encoding of the solutions as vectors.

C. Characteristics of the Genetic Algorithm

The genetic algorithm is easy to implement and to understand. The nature of the algorithm allows simple parallelism and distributed execution. The best found solution is always part of the last generation, and no additional memory for storing is needed. The number of values exchanged between the microservices are relatively small. Mutation and crossover allow the algorithm to escape a local optima and to find a good solution. The algorithm includes a probability model, so it is recommended to execute the algorithm more times from different initial populations. One of the main disadvantages of the algorithm is the execution time.

These characteristics are the reason why genetic algorithm is one of the most commonly used meta-heuristic optimization algorithms today, 25 years after its introduction. Especially the easy distribution and parallelization make the algorithm suitable for a microservice implementation, which is the main motivation for this paper.

IV. Implementation

In the proposed implementation, GA is divided into seven independent components, each implemented as a different microservice. Also, another microservice is added, which is visible to the user, and its goal is to catch the calls to the genetic algorithm and sends a request to the corresponding services. The complete architecture is given on Fig. 1. Each microservice is represented with a rectangle. The arrows are showing service calls (the service which is on the side of the arrow is called by the service on the other end). The green parts contain information about the input parameters and the output value of each microservice. The next paragraphs contain a detailed description of each microservice.

GA Service. This is the main service visible to the user. As an input it takes a parameter object, and is responsible for calling and coordinating of the other services. At the end of the optimization process it returns the best solution to the user.

Initialization. This microservice takes a parameter object as an input (which contains the parameters described above), and based on the values of the size of the population and the dimension of the chromosomes randomly initializes the first generation. The result of this microservice is an initial population, which is a set of chromosomes randomly placed in the search space. Furthermore, this service stores the parameter object which is used by all other microservices. This approach achieves greater speed by sending fewer parameters between service calls. Namely, this service implements REST API which is used to return the parameter object or to change some value of the parameters.

Evolve. The service is responsible for evolution of the generations. As input it takes the initial population, evolves it through more iterations, and returns a list which contains the best solution of each generation.

Select best. This microservice implements the selection mechanism of the best chromosomes in the population. As input it takes the current generation and outputs a subset of chromosomes which will be part of the next generation. This set forms the basis for the next generation, namely, the rest of the next population will be formed by applying crossover and mutation on the individuals contained in the chosen subset. In the literature many selection operators are introduced, which are suitable for different applications, by making a trade-off between the execution time and the probability of returning exactly the best k solutions. For example, tournament selection [6] (with complexity $O(k)$) makes k iterations, and in each

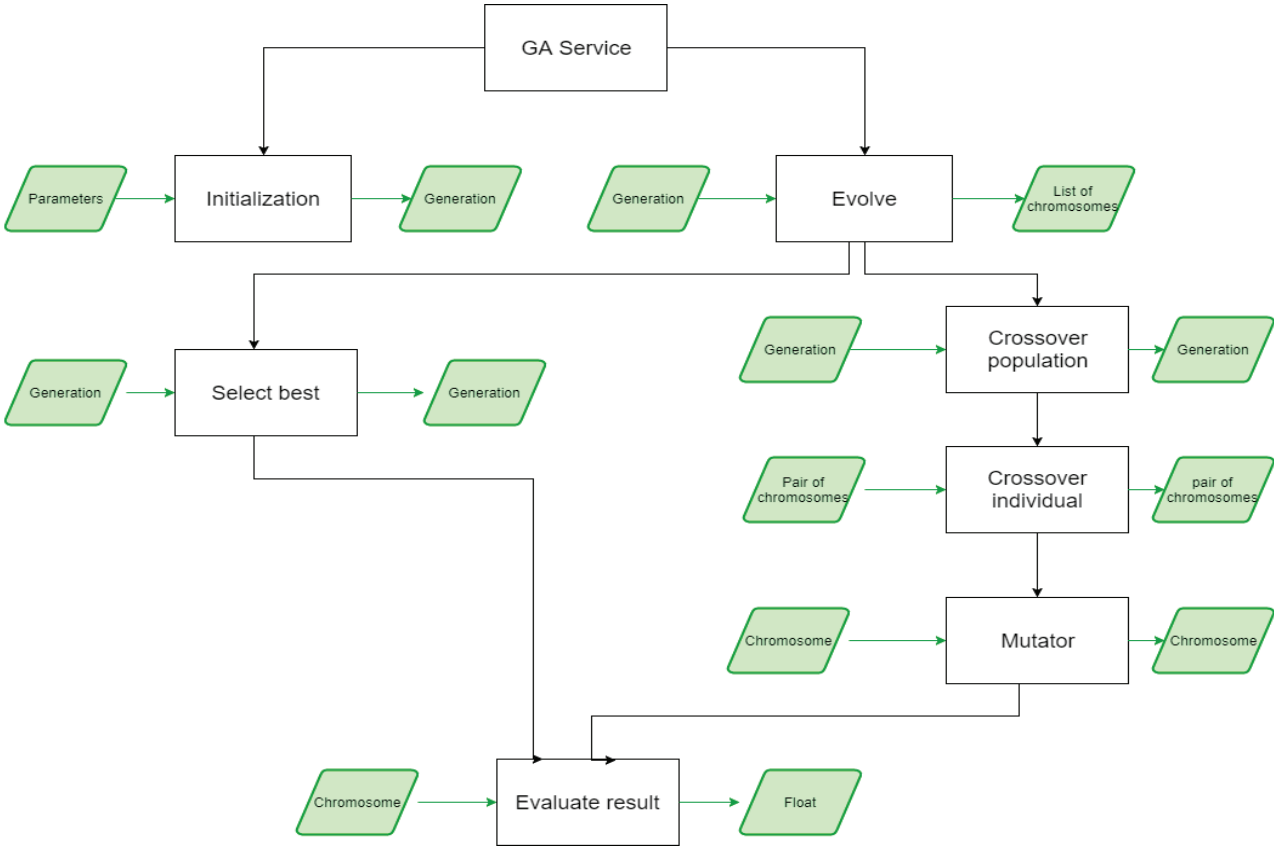


Fig. 1. Implemented microservices and their communication. The green parts present data flow (Input and output parameters of each service)

iteration randomly chooses two chromosomes in constant time, evaluates them and chooses the better one with greater probability to be part of the next generation. On the other end is the selection method which operates with $O(nlog(n))$ complexity, where n is the population size. It sorts all the individuals in the populations, and chooses the best k to continue to the next generation.

Evaluate result. This microservice is responsible for evaluating the quality of each solution, and it is specific for each optimization problem. It takes chromosome as an input, and returns a float value representing the quality of the solution.

Crossover population. After it is decided which individual will continue to the next generation, this microservice is responsible for completing the next generation by performing a crossover on chromosomes of the chosen subset. Namely, cp times it chooses a pair of individuals, and puts their offspring in the next generation. Its input is a generation, it evolves it and returns the new generation.

Crossover individual. The input of this service is a pair of chromosomes, and it implements a crossover technique, which is used to generate offspring to the input chromosomes. The return value is the generated offspring. There are many possible off-the-shelf choices for crossover operator, suitable for different applications. One of the commonly used are Cycle Crossover (CX), Order Crossover (OX) and Partially Matched Crossover (PMX) [7] [8].

Mutator. This service performs a mutation changes on one chromosome. The mutation process usually makes small changes to one or more components of the chromosome. Its input parameter is the original chromosome, and the modified one is returned after the mutation is performed. The most commonly used mutation operators are Insertion mutation, Simple Inversion Mutator and Swap mutator. [7] [8].

A. Used Technologies

All previously described microservices are implemented using Spring framework (Spring cloud part [9]), which contains a large number of libraries used in building a distributed system. This platform offers a simple way to deploy the application to the Cloud. Spring Cloud has powerful libraries which can be used in discovering microservices, managing the configuration, distributed messaging, load balancing etc. The mentioned libraries are developed and used by industry leaders, such as Netflix and other companies whose solutions are implemented using the microservice architecture. The libraries used in the proposed implementation are:

Eureka [10] is a part of Spring Cloud which offers discovering of the microservices, in order to balance the workload between them and to successfully manage the application

when some of the microservices are down. Eureka can be seen as a service registry.

To use this library, a service which works as a monitoring service for all other microservices needs to be created. Namely, when a new service is started (introduced to the system), it tries to register to the main Eureka service. The main task of the monitoring service is to check if each registered microservice is active and to show this information to administrator. Each registered microservice is called an Eureka client. Each Eureka client is registered with its host and port, as well as additional meta-data. The meta-data can contain service name, which can be used for robust code on the client side and easier communication between the microservices. Additionaly, when a service is re-deployed, and the service is given a new IP address, the only change happens in the Eureka service. The other services still can call the new service by its unique name, and their code base remains unchanged.

RabbitMQ library is used for secure and reliable communication between the microservices. Namely, RabbitMQ [11] is a message broker, which means that it is a module which translates the messages from the senders format to the format used by the receiver. Furthermore, the messages are exchanged in the following way: RabbitMQ builds a message queue, and then each service which needs to send a message connects to the queue and uploads the message. The message is saved until the recipient service is connected to the queue and receives the message. Further processing of the message is left to the service that received it. RabbitMQ implements many message protocols, but the most popular and commonly used is Advanced Message Queuing Protocol (AMQP).

Zuul [12] is a service used for dynamically routing, monitoring and improving security of the microservices. This service allows the inner ports of the services to be hidden, and it is used as a ”front door” (known as ”gate keeper”) for all the client requests for other services. This means that its role is to accept the incoming requests and to redirect them to the right services. With this controlling strategy of the requests to the inner microservices it is possible to have further understanding about the health of the implemented system, as well as its protection from various malicious attacks. Zuul offers dynamically reading, compiling and executing of series of filters which are used to control the HTTP requests and responses. Zuul allows the client not to know the specific ports, but to focus on the provided API, and relies on Zuul to make the call to the corresponding microservice.

Pivotal Cloud Foundry [13] is used as a platform for deploying of the microservices. This platform is chosen because it offers many implemented solutions for microservice management, for example on demand scaling, dailover/resilience, road balancing, monitoring etc.

Fig. 2 illustrates the communication between the used technologies. Namely, the client requests first are coming to the Zuul service, because the ports and addresses of the individual microservices are not known to the user. This service distributes the request to the corresponding microservices and returns the results to the user. To achieve that, it relies on Eureka, which has a completely registry of the services. The communication between the services (including the Eureka service) is performed using RabbitMQ and its implemented message queue.

Note. All microservices communicate using the Rest Template. This allows simple request building and parsing of the answers. In the proposed implementation all the messages are in json format.

The performances of a microservice approach on the genetic algorithm implementation depends heavily on the computational performance of the evaluation function, and other specific components of GA. Namely, the network latency is a big problem, and can overpower the save in execution time by the distributed execution. That means, in order to obtain an improvement with a microservice approach, we need to test on a very complex evaluation function. Unfortunately, we were not able to do this, due to the high cost of the used platform for large scaling. Our services were deployed on the free version, with 1GB of RAM and single core CPU. Of course, for some applications, this approach will have poorer performances, but the point is that the presented approach has different advantages, for example flexibility and easy scaling up and down.

V. Conclusion

This paper shows that Genetic Algorithm is suitable for a microservice based implementation, because it consists of many independent components with different processing power needs. Advantages of this approach are numerous: for each component, information about its workload is available, leading to easily scaling of the components that need more processing power. The main disadvantage is the message based communication between the microservices, which is significantly slower compared to the in-memory calls. Furthermore, each new implementation of the crossover and mutation operators needs to be implemented as a new microservice and deployed, which is more complex and less generic that simply adding a new function in the monolith architecture.

As further work is left to develop the prototype into a more complex and fully implemented application, which offers many possibilities for different operators used by the genetic algorithm. In addition, the scaling of the components of the proposed implementation (at the moment) is done by hand, namely the administrator observes which components have high workload, and by hand initializes more processing units with this components. A way should be found to automate this process. Additionally, experiments should be introduced to examine the needed processing power, in order to achieve performance gain by the presented approach.

References

[1] A. Wang and S. Tonse, ”Announcing Ribbon: Tying the Netflix Mid-Tier Services Together”, Medium, 2018. [Online]. Available: http://techblog.netflix.com/2013/01/announcing-ribbon-tying-netflix-mid.html. [Accessed: 25- Mar- 2018].

[2] J.Thnes, ”Microservices.” IEEE Software vol 32 no. 1 pp. 116-116, 2015

[3] N. Dragoni, ”Microservices: yesterday, today, and tomorrow.” Present and Ulterior Software Engineering. Springer, Cham, pp.195-216. 2017.

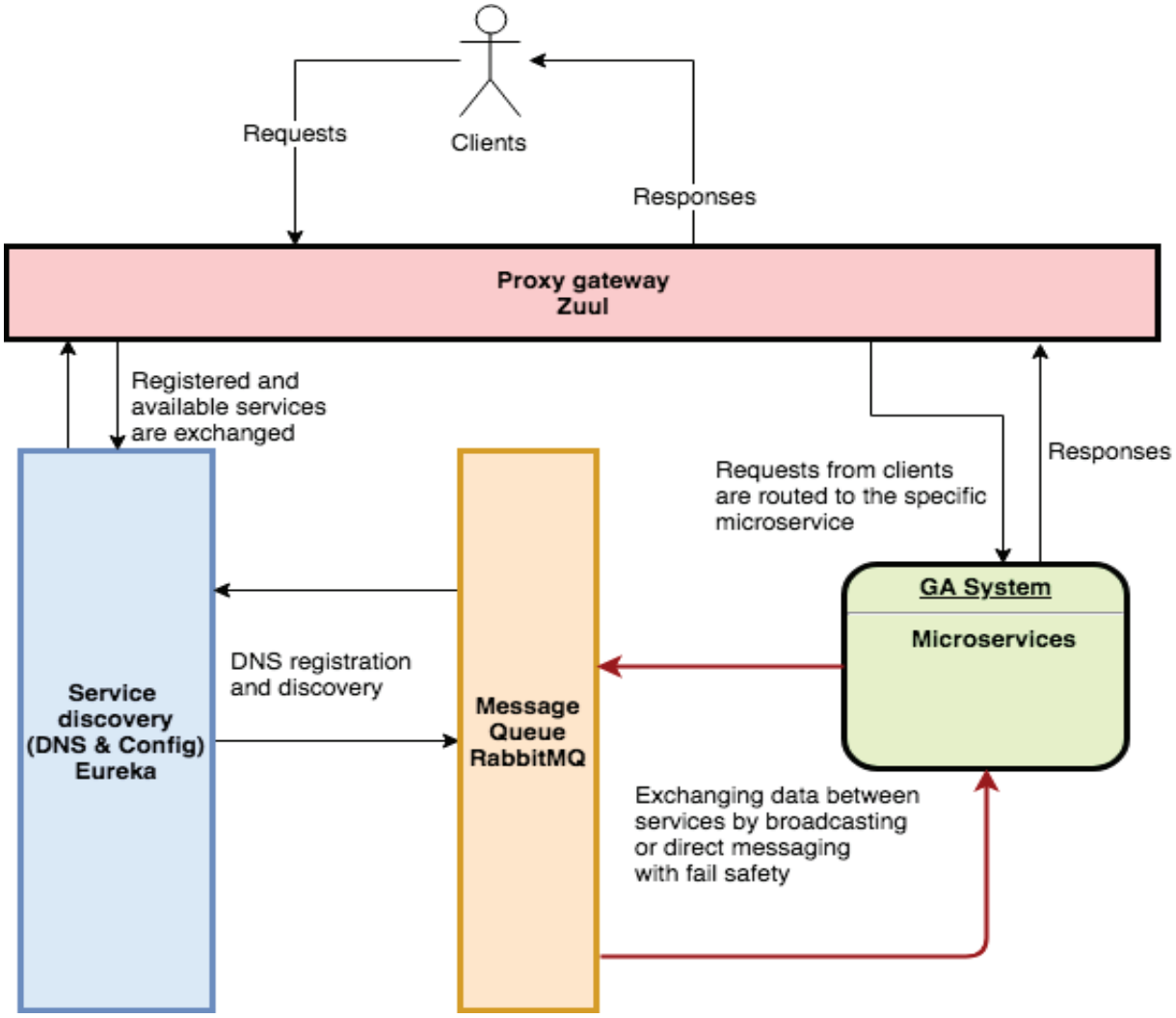


Fig. 2. Overview of the interactions between the used technologies.

[4] J. Holland, ”Genetic algorithms.” Scientific american vol 267 no. 1 pp. 66-73, 1992
[5] L, Kwang, and M. A. El-Sharkawi, ”Modern heuristic optimization techniques: theory and applications to power systems”. John Wiley and Sons vol 39, 2008.
[6] B. Miller and D. E. Goldberg. ”Genetic algorithms, tournament selection, and the effects of noise.” Complex systems vol 9 no. 3 pp. 193-212 1995
[7] S.N. Sivanandam and S. N. Deepa. ”Genetic algorithm optimization problems.” Introduction to Genetic Algorithms. Springer Berlin Heidelberg pp. 165-209, 2008.
[8] L.Y. Kwang, and M. A. El-Sharkawi, ”Modern heuristic optimization techniques: theory and applications to power systems.” vol. 39. John Wiley and Sons, 2008.
[9] ”Spring Cloud”, Projects.spring.io, 2018. [Online]. Available: http://projects.spring.io/spring-cloud/. [Accessed: 25- Mar- 2018].
[10] ”Introduction to Spring Cloud Netflix - Eureka — Baeldung”, Baeldung, 2018. [Online]. Available: http://www.baeldung.com/spring-cloud-netflix-eureka. [Accessed: 25- Mar- 2018].
[11] ”RabbitMQ - Messaging that just works”, Rabbitmq.com, 2018. [Online]. Available: https://www.rabbitmq.com/. [Accessed: 25- Mar- 2018].
[12] ”Netflix/zuul”, GitHub, 2018. [Online]. Available: https://github.com/Netflix/zuul. [Accessed: 25- Mar- 2018].
[13] ”Pivotal Cloud Foundry (PCF)”, Pivotal.io, 2018. [Online]. Available: https://pivotal.io/platform. [Accessed: 25- Mar- 2018].

Overview of Creativity Assessment Framework for a Computer Program

Gordana Ispirova
Jožef Stefan International Postgraduate School
Computer Systems Department, Jožef Stefan Institute
Ljubljana, Slovenia
gordana.ispirova@ijs.si

Tome Eftimov
Computer Systems Department, Jožef Stefan Institute
Ljubljana, Slovenia
tome.eftimov@ijs.si

Barbara Koroušić Seljak
Computer Systems Department, Jožef Stefan Institute
Ljubljana, Slovenia
barabara.korousic@ijs.si

***Abstract*— Judging computational creativity is not a simple task. In order to assess if something behaved creatively first we need a formal and precise definition of the meaning of creativity in the certain area. The next thing would be formalized criteria whose application rates the level of creativity. In the area of computer creativity there is not a lot of work on this topic. The first framework for computational creativity assessment is very well backed up by all the previous and following research in the area of computer creativity and it has been proven on many examples. In this paper an overview of the proposed framework for assessing the creativity of a computer program and a few points on the overall task for creativity assessment are presented.**

Keywords—creativity assessment, computer creativity, assessing computer creativity

I. Introduction

One of the major questions in computational creativity is how to assess the creativity of a system/process. Such high-level assessment of computation creativity is presented in [1]. In this paper, a detailed mathematically supported framework for accessing the creativity of a computer program is presented. The main goal of the paper is to set out the relevant matters for determining if a particular computer program has behaved, or is behaving "creative", and if it is creative enough to value the level of creativity. In order to assess something, there is a need of a formal and explicit definition for the subject of assessment, or in this case – creativity.

Referencing previous definitions of creativity, it is more than obvious to start from [2]:

"True creativity results from the transformation of conceptual space."

The evidence for this is that a creative process results with an artefact, which not only is significantly different from its predecessors, but also establishes new norms by which followers may be judged. Boden goes on to use this analysis in terms of conceptual space as a criterion for assessing the creativity of programs, by applying it not merely to the final artefact but also to the manner in which the result is produced. Implying that she regards an explicit transforming of the space as a necessary condition for a program to be creative.

Other definitions of creativity include [3] and its updated version [4] which states that:

"The philosophy, science and engineering of computational systems which, by taking on particular responsibilities, exhibit behaviors that unbiased observers would deem to be creative."

It is in our nature to judge creativity found on human creativity. Needless to say, we, as humans, are extremely bias when it comes to evaluating activities as creative. For example: writing poetry, painting, or composing music are almost always considered as creative, even when done in an ordinary way, whereas, it is very uncommon to class scientific, mathematic or engineering tasks as creative, unless they are done exceptionally well.

In this context, we should mention the difference between H-creativity (producing an idea/artefact which is wholly novel within the culture, not just new to its creator) and P-creativity (producing an idea/artefact which is original as far as the creator is concerned, even though it might have been proposed or generated elsewhere in the culture, perhaps much earlier in history) which is defined in [2]. Boden states that the P-creativity is of importance when we overlook the process of being creative, even though isn't very useful to society because of its repeating nature.

II. Framework for Assessing Creativity

Two main criteria for assessing creativity are:

Novelty: *To what extent is the produced item dissimilar to existing examples of that genre?*

Quality: *To what extent is the produced item a high-quality example of that genre?*

The main aim of this framework is to define as precisely and formally as possible the attributes needed to decide if a particular computer program had behaved or is behaving creatively. This is done based on a reference architecture [5] where a protocol/pattern of a creating program is presented.

A computer program that produces artefacts is a creating program. The produced artefacts are the set of output data of the program or the set of basic items (which can be infinite, but most commonly is a finite set). All of the items in this set are not necessarily "successful" or "valid" items. For example: if we are dealing with a program that generates jokes the set of basic items is sequences of strings, but not every sequence is a joke, and even more, not a funny joke.

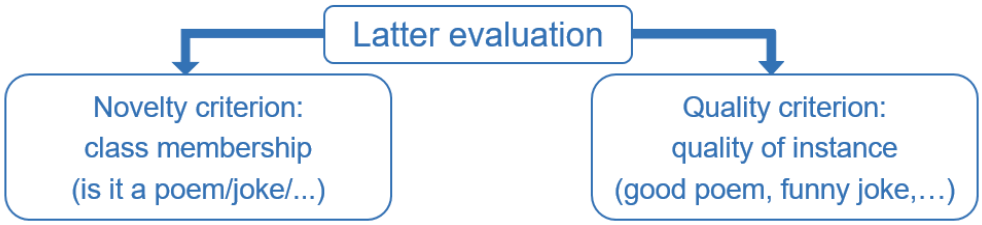


Fig. 1. Two step evaluation of an instance from the set of basic items

In order to evaluate an item of a creating program first we need to determine if the item is indeed a member of the so-called target class (the class of "valid" items). The basic items are possible instances of the intended class of artefacts – the target class, which is represented as a mapping from the basic items to the interval [0,1]. Here, it is necessary to point out that computational generation of artefacts which resemble some kind of art and are valid (valid poem/joke etc.) is a very requiring and demanding task.

To assess an item from the set of basic items according to the two criterions – Novelty and Quality there is a need of a latter evaluation (Fig. 1). This evaluation contains two separate mappings. The first one is, as mention, before mapping for class membership in order to determine to what extent an item meets the criteria for membership of the intended artefact class. The items that gain high scores on the mapping are part of the class, and the low-scoring items are implicitly dissimilar to the class. The second mapping is for the properties that indicate that the item (part of the target class) is a good instance of the corresponding type.

These two mappings are based on a rating scheme. The set of possible rating schemes for a set A are written as '$\mathcal{RAT}(A)$'.

Given the set of basic items, there are properties that these items may or may not have, i.e. may have to varying degrees. For example, a poem may have properties like: rhyme, rhythm, imagery, etc. This is given in the following definition:

*Given some set A a **property rating scheme** for A is a tuple* $\langle f_1, \dots, f_n \rangle$ *of functions from A to the closed interval* [0,1].

Where, each f_i represents a property that an element of A may have to some degree. This definition also has an enhanced version for weighted property rating scheme, where for each property we have a pair of the property function and a corresponding weight w_i:

*Given some set A a **weighted property rating scheme** for A is a tuple* $\langle (f_1, w_1), \dots, (f_n, w_n) \rangle$ *where each* f_i *is a function from A to the closed interval* [0,1], *and each* w_i *is a numerical **weight**, with* $\sum_{i=1}^{n} w_i = 1$.

As stated previously, we have a rating scheme and a set of basic items which consist the artefact class. The rating scheme consists of two types of ratings: typical ratings (for class membership) and value ratings (for rating the quality of an item):

*A **value-based artefact class** consists of a triple* $(\mathcal{B}, typ, val)$, *where* $\mathcal{B}$ *is a set (the basic items) and* $typ, val \in \mathcal{RAT}(\mathcal{B})$ *(the **typical ratings** and the **value ratings** respectively).*

Referencing Fig. 2, where the overall scheme is shown, in order to define the criteria for creativity set from this framework we shall define a few more general terms:

- Selection – the process of separating items from the set of basic items, which guide the construction of the creative program.
- Inspiring set – the set (usually finite) of items selected in the selection process.
- Program construction – the process of mapping from the inspiring set (and the relevant rating schemes) to a program:

  Given a value-based artefact class $(\mathcal{B}, typ, val)$, *a **program construction scheme** consists of a pair* $(S_{\mathcal{B}}, C_{\mathcal{B}})$ *where* $S_{\mathcal{B}}$, *the **selection** process, is a mapping from* $\mathcal{RAT}(\mathcal{B}) \times \mathcal{RAT}(\mathcal{B})$ *to* $\mathcal{P}(\mathcal{B})$ *(the powerset of* $\mathcal{B}$*), and* $C_{\mathcal{B}}$, *the **construction** process, is a mapping from* $\mathcal{P}(\mathcal{B}) \times \mathcal{RAT}(\mathcal{B}) \times \mathcal{RAT}(\mathcal{B})$ *to the set of generating programs for* $\mathcal{B}$.

- The program – consists of: generating procedure (the main algorithm) and a definition of the ranges of initial data values (the possible parameters for the algorithm):

  Given a set of basic items $\mathcal{B}$, *a **generating program** for* $\mathcal{B}$ *consists of a pair* $(\langle D_1, \dots, D_k \rangle, G)$ *where* $\langle D_1, \dots, D_k \rangle$ *is a k-tuple of sets, and* G *is a mapping from* $D_1 \times \dots \times D_k$ *(*D_i *is the domain for some parameter for the procedure G) to* $\mathcal{P}(\mathcal{B})$.
  ***An initialization** is a mapping* $I_{\mathcal{B}}$ *from* $\mathcal{P}(\mathcal{B}) \times \mathcal{RAT}(\mathcal{B}) \times \mathcal{RAT}(\mathcal{B})$ *to* $\langle D_1, \dots, D_k \rangle$. *That is, it is a choice of initial parameters, based on the inspiring set* $(S_{\mathcal{B}}, typ, val)$, *and the ratings schemes* typ, val.

The criteria of this framework are defined for a single run of a program, but the author assumes generalization to a set of runs: *For a program* $(\langle D_1, \dots, D_k \rangle, G)$, *a **run** is a pair* $(\langle d_1, \dots, d_k \rangle, G(d_1, \dots, d_k))$ *where* $\langle d_1, \dots, d_k \rangle \in D_1 \times \dots D_k$.

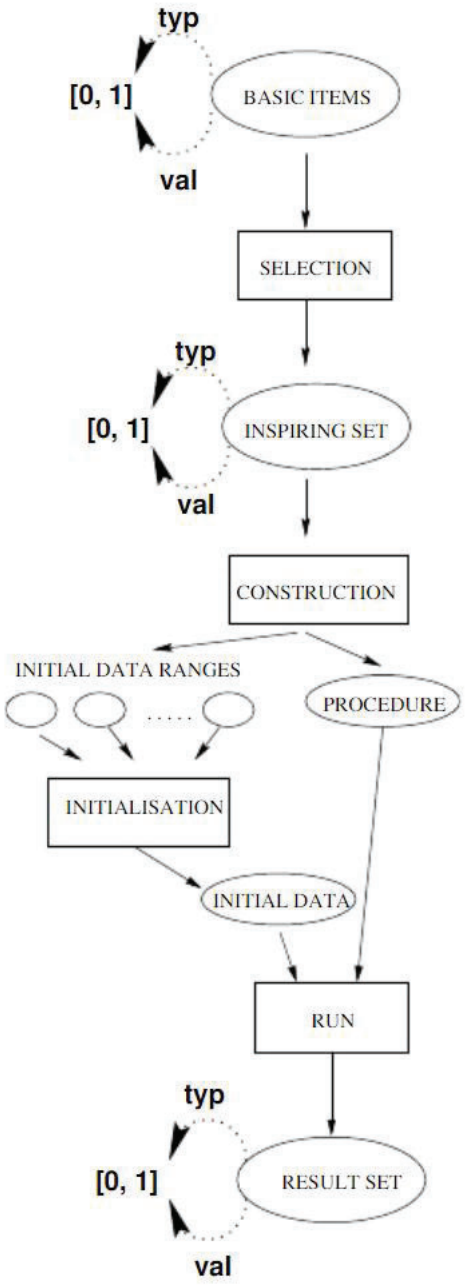


Fig. 2. The framework for accessing creativity [1]

The criteria are the following:

1. The average rating should be high.
2. Highly rated (very typical) items should be a big portion of the results.
3. Generated items should have intrinsic merit, on average.
4. High quality items should make up a significant proportion of the results.
5. If a high proportion of the program's "normal" output is scoring well, the program can be successful.

Untypical high-valued items are judged by comparing them with (6-8):

6. The entire set of outputs.
7. The whole set of untypical items.
8. The set of typical high-valued items.
9. A creative program might well replicate most of the inspiring set.
10. Producing more than just the inspiring set is a symptom of creativity.
11. Results which are not in the inspiring set should be at least typical of the genre, and better still highly-valued.
12. Well rated items not in the inspiring set should be a significant proportion of the results.

III. Practical Use of the Framework

The framework described is based on metatheoretical ideas so when it comes to using the framework in practice, to real programs, the whole range of criteria is unlikely to be suitable for detailed assessment of all generating programs. The first 8 criteria are more tractable, and in many cases, for suitable values of the parameters in these criteria, it is possible to compute precise values.

A simple example for this can be taken from a program called JAPE, which is a joke-generator. Its output was evaluated by human judges against two standards: 'is this item a joke?' which corresponds directly to typical rating and 'how funny is this item?' corresponding to value rating of the framework [6]. So, if the complete raw data from the evaluation is given, JAPE can be assessed for the criteria 1-8.

However, when it comes to the rest of the criteria, this does not apply, simply because of the fact that the inspiring set is usually not documented precisely. Nevertheless, there are cases when these criteria can be computed with great precision. For example, when the construction process is carried out by machine learning techniques, the training set acts as the inspiring set, hence the inspiring set is very well documented.

After the text edit has been completed, the paper is ready for the template. Duplicate the template file by using the Save As command, and use the naming convention prescribed by your conference for the name of your paper. In this newly created file, highlight all of the contents and import your prepared text file. You are now ready to style your paper; use the scroll down window on the left of the MS Word Formatting toolbar.

IV. Discussion

Assessing creativity is a demanding task, not easily approachable, merely because of the sole perception of creativity itself as a process. The affinity towards a produce from a human creative process is highly personal, what some may appreciate, others may despise and there even may be some who are completely indifferent to it. In society this is accepted as difference of taste, meaning each individual has personal and cultural patterns of choice and preference.

When it comes to assessing creativity of a computer program the task becomes even more requiring. Relativistic judgement enters into the assessment of creativity of a computer program in two aspects. The first one being the previously mentioned personal appraisal of a program's output, in this framework, this is modeled by the class-rating and the value-rating. The two rating schemes (typical and value) act as a formalization of the

judge's viewpoint, i.e. any formal definitions of the quality of the program results can be stated to be relative to the rating schemes involved, and hence to the subjective view of a particular judge.

The second aspect is the lack of general agreement on what counts as creative, particularly when considering computer programs. The framework outlined here allows for multiple definitions of creativity. As mentioned before, it is a highly demanding task for a computer to manage even to produce "normal" or typical" items of a genre. The fact that different criteria seem to lead in different directions with respect to the underlying intuition about creativity is not a problem. Rather, one can define different variants by suitable choices and combinations of criteria.

They say that a house must be built on solid foundations if it is to last. Having a strong and good quality base for something is crucial, this framework is exactly that for the field of assessment of computer creativity. A framework like this one deserves a lot of attention, not just because of the fact that is the first and only one to tackle this type of problem, but because of its many possible extensions. There are a number of ways in which the ideas presented in the framework could be elaborated to capture subtler aspects of creative processing.

One of the possible extensions is allowing for a creating program to somehow rate its output by assigning some sort of rating to each item that it produces. This can be done by modifying the formal definitions above in order to accommodate this form of data. Of course, it is inevitable that the program's rating of its own output would be a produce or at least in high correlation with the program's typical and value ratings for the corresponding output. This type of extension would make the framework more appealing and more understandable.

Another possible extension is random generation of the set of basic items. This can be done by having a detailed definition of the available space of basic items and how this space could be randomly explored. This requires making the internal structure of an item explicit (e.g. as an array of pixels, or as a sequence of words) and having a suitable definition for random combination of the atomic parts from the internal structure.

Having its strong sides, the framework has its weaknesses too. An important weakness is that it does not handle the notion of similarity very cleanly. In particular, the present set of criteria compare the result set with the inspiring set only in terms of overlap in membership, but make no allowance for the idea that output items could vary in the extent to which they are similar to those in the inspiring set. This can be partially handled by building some notion of multi-dimensional space into the assessment framework. Certain forms of rating schemes would allocate the basic items to points in a multi-dimensional space, which in turn would lead naturally to a measure of distance between items. In this way the overlapping and the generation of similar items can be minimized if not totally taken care of.

Another topic that is open for discussion is the fact that the criteria set out by this framework are defined on a single run of a creating program and a generalization to a set of runs is assumed. This seems to be pretty straightforward and self-explanatory when it comes to many things but not very desirable from a statistical point of view. If the process of random generation of the inspiring set is applied, then the whole process loses its deterministic nature and it becomes stochastic. In this case a single run of a program does not provide enough insight into the algorithm or the kind of data itself, let alone enough data to conduct a proper statistical or any kind of analysis. Anyone who has dealt at some point with generating or any kind of output computer programs will state that a single run does not provide enough information and a program can be tested only from various runs. Therefore, this is also considered as a drawback to this framework.

After the publication of this framework there have been several newer approaches [7-9] regarding the subject of creativity assessment, but none of them concerning about assessing computer creativity. For example, the study [9] is on an assessment framework for creativity called RADSE (Research, Analysis, Development, Solution, Evaluation), which is a method currently used in computer animation that could be employed as a general framework for assessment in art, media and design (ADM), as it allows for the assessment of the creativity of a person's creations in this field. However, assessing computational creativity still is a topic open for discussion.

V. CONCLUSION

This framework proposes an approach to the assessment of creativity in computer programs which provides a general sketch of what might constitute a creative program, highlights various factors which are relevant to an empirical judgement of creativity (about a program) and is explicit and formal about these factors. As the first of its kind it is a well-rounded base for further improvement or other versions based on it.

[1] G. D. Ritchie, "Assessing Creativity," in Proceedings of AISB Symposium on AI and Creativity in Art and Science Society for the Study of Artificial Intelligence and Simulation of Behaviour, 2001.

[2] M. Boden, The Creative Mind, Abacus, London, 1992.

[3] G. A. Wiggins, "Searching for computational creativity," New Generation Computing, 2006, pp. 209-222.

[4] S. Colton, and G. A. Wiggins, "Computational Creativity: The Final Frontier", in Proceedings of 20th European Conference on Artificial Intellegence, 2012.

[5] L. Cahill, C. Doran, R. Evans, D. Paiva, C. Mellish, M. Reape, D. Scott, and N. Tipper, "In search of a reference architecture for nlg systems," in European Workshop on Natrual Language Generation, Toulouse, 1999.

[6] K. Binsted, H. Pain, and G. D. Ritchie, "Children's evaluation of computer-generated punning riddles," in Pragmatics and Cognition, 1997, pp. 309-258.

[7] A.Loveless, J. Burton, and K. Turvey, "Developing conceptual frameworks for creativity, ICT and teacher education", in Thinking Skills and Creativity, 2006.

[8] M. Blamires, and A. Peterson, "Can creativity be assessed? Towards an evidence informed framework for assessing and planning progress in creativity", in Cambridge Journal of Education, 2016.

[9] M. Salamon, "Developing a Strategy for Assessing Creativity: the Creative Spiral", in Investigations in university teaching and learning, 2008.

A Note On an Error-Detecting Code

Nataša Ilievska
Faculty of Computer Science and Engineering
"Ss.Cyril and Methodius" University
Skopje, Republic of Macedonia
natasa.ilievska@finki.ukim.mk

***Abstract*— In this paper we consider an error-detecting code based on linear quasigroups of order 4. We are focused on the number of errors that the code surely detects. Namely, for each quasigroup of order 4 from the best class of quasigroups of order 4, using simulations we obtain the number of errors that the code detects for sure. At the end we conclude whether this number is the same for all quasigroups from the best class of quasigroups of order 4 for coding.**

Keywords*— *error-detecting code, linear quasigroup, code word, binary symmetric channel

I. Introduction

In this paper we are focused on the number of errors that an error-detecting code based on linear quasigroups detects for sure. Namely, the considered error-detecting code is completely analyzed from the aspect of the probability of undetected errors, i.e., the probability that there will be errors in transmission that the code will not detect. The probability of undetected errors depends on the quasigroup used for coding ([1]). In [2] is given the best class of quasigroups of order 4 for coding, i.e., the class of quasigroups of order 4 that gives smallest probability of undetected errors. Naturally arises a question about the number of errors that the code surely detects. Namely, first question is which is the number of incorrectly transmitted bits that the code surely detects when for coding is used quasigroup from the best class of quasigroups of order 4. Second question is whether these numbers are equal for all quasigroups from the best class of quasigroups of order 4. In order to answer these questions, we use a simulation. We make some conclusions based on the obtained experimental results.

The paper is organized in the following way. In Section II are given the basic mathematical definitions, in Section III is given the definition of the error-detecting code. The main results are presented in Section IV, where in Subsection *A* are given the results from the simulation procedure for obtaining the number of incorrectly transmitted bits that the code detects for sure, when for coding is used each of the quasigroups of order 4 from the best class of quasigroups of order 4.

II. Mathematical Preliminaries

The code that we consider in the paper is based on linear quasigroups. For that reason, in this section we give the definition of this algebraic structure.

Definition 1: Quasigroup is algebraic structure $(Q,*)$ such that

$$(\forall\, u, v\in Q)(\exists!\, x, y\in Q)(x*u=v\wedge u*y=v) \quad (1)$$

Definition 2: 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

$$(\forall x, y\in Q)\ x*y=z\Leftrightarrow \boldsymbol{z}=\boldsymbol{x}A+\boldsymbol{y}B+C \quad (2)$$

where $\boldsymbol{x}$, $\boldsymbol{y}$ and $\boldsymbol{z}$ are the binary representations of x, y and z as vectors of order $1\times q$ and + is a binary addition.

When we say that $(Q,*)$ is a quasigroup of order 2^q, then we take $Q=\{0, 1, 2,\ldots, 2^q-1\}$.

From the above definition follows that linear quasigroups instead to be presented with the quasigroup operation * can be presented with a binary matrices A, B and C. Namely, if the linear quasigroup is given by the operation *, then using (2) one can obtain the corresponding binary matrices A, B and C. Conversely, if the linear quasigroup is given by the binary matrices A, B and C, again using (2) one can obtain the quasigroup given by the operation * (Example 1).

Example 1: One can very easy check that the quasigroup

*	0	1	2	3
0	0	3	2	1
1	1	2	3	0
2	3	0	1	2
3	2	1	0	3

Fig.1 Linear quasigroup of order 4

is linear, i.e., there are non-singular binary matrices A and B and a binary matrix C, such that (2) is satisfied. Namely, let we represent $\boldsymbol{x}$, $\boldsymbol{y}$ and $\boldsymbol{z}$ in (2) as 1×2 binary vectors $[x_1\ x_2]$, $[y_1\ y_2]$ and $[z_1\ z_2]$ respectively and let the matrices A, B and C are

$$A=\begin{bmatrix}a_{11} & a_{12}\\ a_{21} & a_{22}\end{bmatrix}, B=\begin{bmatrix}b_{11} & b_{12}\\ b_{21} & b_{22}\end{bmatrix}, C=[c_1\ \ c_2] \quad (3)$$

Now, in order to check whether the quasigroup is linear, one should check whether there are binary values a_{ij}, b_{ij} and c_i, i, $j\in\{1,2\}$ such that A and B are non-singular matrices and the following equation is satisfied for all x and y from the quasigroup Q:

$$[z_1\ z_2]=[x_1\ x_2]\begin{bmatrix}a_{11} & a_{12}\\ a_{21} & a_{22}\end{bmatrix}+[y_1\ y_2]\begin{bmatrix}b_{11} & b_{12}\\ b_{21} & b_{22}\end{bmatrix}+[c_1\ \ c_2] \quad (4)$$

where $z=x*y$.

Equation (4) is equivalent to the following system of equations

$$\begin{cases} z_1 = a_{11}x_1 + a_{21}x_2 + b_{11}y_1 + b_{21}y_2 + c_1 \\ z_2 = a_{12}x_1 + a_{22}x_2 + b_{12}y_1 + b_{22}y_2 + c_2 \end{cases} \quad (5)$$

By substituting the binary representations of all possible x, $y \in \{0, 1, 2, 3\}$ and the corresponding binary representation of $z=x*y$ in (5), we obtain a system of 32 equations with 10 unknowns. The solution of this system of equations is $a_{11}=1$, $a_{12}=1$, $a_{21}=0$, $a_{22}=1$, $b_{11}=1$, $b_{12}=0$, $b_{21}=1$, $b_{22}=1$, $c_1=0$ and $c_2=0$, from where follows that the quasigroup given in Fig. 1 is linear and the corresponding binary matrices are:

$$A = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}, B = \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix}, C = [0\ 0] \quad (6)$$

Conversely, if the quasigroup is given with the matrices (6) then the matrix in Fig.1 can be obtained using (2). Namely, in order to obtain $x*y$, we substitute the binary representations of x and y in (2). For example, in order to obtain 3*2, since the binary representation of $x=3$ is $\boldsymbol{x}=[1\ 1]$ and the binary representation of $y=2$ is $\boldsymbol{y}=[1\ 0]$, we compute

$$\boldsymbol{x}A+\boldsymbol{y}B+C=[1\ 1]\begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}+[1\ 0]\begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix}+[0\ 0]=[0\ 0]$$

which means that 3*2=0. In this manner we compute $x*y$ for all $x, y \in \{0, 1, 2, 3\}$. Whit this we obtain the quasigroup given in Fig. 1.

III. Definition of The Error-Detecting Code

Let for coding be used the quasigroup $(Q, *)$ of order 4. Let the input block be $a_0a_1\ldots a_{n-1}$, where $a_i \in Q$, for all $i \in \{0, 1, 2, \ldots, n-1\}$. The redundant characters are calculated using the following equation:

$$d_i=a_i*a_{i+1},\ i=0, 1, \ldots, n-1 \quad (7)$$

where all operations in the indexes are modulo n. The extended block $a_0a_1\ldots a_{n-1}d_0d_1\ldots d_{n-1}$, previously converted into binary form is transmitted through the binary symmetric channel. But, since there are noises in the channel some of the characters may not be correctly transmitted. To check whether the code word is correctly transmitted, the receiver checks if the output block satisfies (7) for all $i \in \{0, 1, 2, \ldots, n-1\}$. If for some $i \in \{0, 1, 2, \ldots, n-1\}$, (7) is not satisfied, the receiver concludes that there are errors in transmission and asks the sender to send the block once again. But, since the redundant characters $d_0, d_1, \ldots, d_{n-1}$ are also transmitted through the binary symmetric channel, noises affect them also. For this reason, it is possible some of the redundant characters to be incorrectly transmitted in such a way that (7) is satisfied for all $i \in \{0, 1, 2, \ldots, n-1\}$, although some of the information characters are incorrectly transmitted. Therefore, it is possible to have undetected errors in transmission (Example 2). For this reason, two parameters are important: the probability of undetected errors (obtained in [1] and [2]) and the number of errors that the code surely detects. In the next section we will obtain experimental results for the second parameter.

Example 2: Let we use the quasigroup from example 1 for coding. Let the input block be: 230121. This means that the information characters are $a_0=2$, $a_1=3$, $a_2=0$, $a_3=1$, $a_4=2$ and $a_5=1$. We compute the redundant characters:

$$d_0=a_0*a_1=2*3=2$$

$$d_1=a_1*a_2=3*0=2$$

$$d_2=a_2*a_3=0*1=3$$

$$d_3=a_3*a_4=1*2=3$$

$$d_4=a_4*a_5=2*1=0$$

$$d_5=a_5*a_0=1*2=3$$

Now, the coded block is 230121223303. This coded block turned into binary form 101100011001101011110011 is transmitted through the binary symmetric channel. Let suppose that the output block is 101**0**0001101**1**1101011110011, i.e., the fourth and the eleventh bits of the code word are incorrectly transmitted. The output block turned into string over the alphabet $Q=\{0,1,2,3\}$ is 2**2**012**3**223303. The information block is 2**2**012**3** and the redundant block is 223303. In order to check if there are any errors in transmission, the receiver checks whether the output block satisfies (7). The first redundant character should be 2*2=1, but we see that the first redundant character is 2, from where the receiver concludes that there are errors in transmission.

But, let suppose that also some of the redundant characters are incorrectly transmitted and the output block is 101**0**0001101**1**0**1**1**1**1111**1**0**0**. The bolded bits are the ones that are incorrectly transmitted. This output block turned into string over the alphabet Q is 2**2**012**3**1**3**33**2**0. Now, when the receiver checks whether the code word is correctly transmitted, obtains:

2*2=1=1

2*0=3=3

0*1=3=3

1*2=3=3

2*3=2=2

3*2=0=0

which means that (7) is satisfied. In this case, although the code word is incorrectly transmitted, the receiver does not detect the errors in transmission and accepts the code word as correctly transmitted.

IV. Simulational Results For The Number of Errors That The Error-Detecting Code Detects For Sure

The best class of quasigroups of order 4 contains four pairs of non-singular binary matrices A and B, given in Fig. 2.

$$A_1 = \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix}, B_1 = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}$$

$$A_2 = \begin{bmatrix} 0 & 1 \\ 1 & 1 \end{bmatrix}, B_2 = \begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix}$$

$$A_3 = \begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix}, B_3 = \begin{bmatrix} 0 & 1 \\ 1 & 1 \end{bmatrix}$$

$$A_4 = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}, B_4 = \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix}$$

Fig. 2 The best class of quasigroups of order 4 for coding

Although the matrix C can be arbitrary binary matrix of order 1×2, we will take the zero matrix of order 1×2 as matrix C.

Note that the linear quasigroup in example 1 is the quasigroup represented with the binary non-singular matrices A_4 and B_4 in Fig. 2 and C is zero matrix.

In what follows, using simulations we will obtain experimental results for the number of errors that the code surely detects when for coding is used a linear quasigroup of order 4 from Fig. 2.

In the simulation procedure, we transmit large number of code words of length n bits. These code words are transmitted through a simulated binary symmetric channel with a probability of bit-error p. For each code word we found the number of incorrectly transmitted bits and for each incorrectly transmitted code word using (2) and (7) we determine whether the error is detected. In this way, for every j from 1 to the length of the code words n, we obtain the number of code words that have j incorrectly transmitted bits and the number of code words with j incorrectly transmitted bits in which the error in transmission is not detected.

We obtain these results for different values of the length of the code words n, i.e., we run the simulation process when the length of the code words run from 8 to 64 bits with step 4 bits. Also, each of the four quasigroups given in Fig. 2 is used for coding. Using these results, we make a conclusion about the number of errors that the code surely detects.

A. The experimental results

Since we want in the results from the simulation process to have large number of incorrectly transmitted code words with small number of incorrectly transmitted bits, we chose the probability of bit-error to be small number, i.e., we chose p=0.06. Note that the number of incorrectly transmitted bits that the code surely detects does not depend on the probability of bit-error p in the binary symmetric channel.

When the input block has length n characters from the quasigroup used for coding, the code word will have length $2n$ characters. Every character from the quasigroup of order 4 is represented with 2 bits, from where follows that the code words have length $4n$ bits. For this reason, the length of the code words in the case when quasigroup of order 4 is used for coding is multiple of 4.

The number of code words with j incorrectly transmitted bits $e(j)$ and the number of code words with j incorrectly transmitted bits in which the error in transmission is not detected $ue(j)$ when for coding is used the linear quasigroup of order 4 represented with the pair of non-singular binary matrices A_1 and B_1, A_2 and B_2, A_3 and B_3, A_4 and B_4, are given in Table I, Table II, Table III and Table IV respectively. Due to the space limitation, in the tables are given results for code words with up to 7 incorrectly transmitted bits. In the results from the simulation, the number of code words with more than 7 incorrectly transmitted bits is relatively small and the error-detecting code detected all of them, i.e., the number of such code words in which the error is not detected is 0.

As we can see from Table I, the number of incorrectly transmitted code words with one, two or three incorrectly transmitted bits is large for all considered values of the length of the code words. The simulation showed that despite the number of incorrectly transmitted code words with one, two or three incorrectly bits is large, the code detects all such incorrectly transmitted code words, i.e., the number of incorrectly transmitted code words with j incorrectly transmitted bits in which the error is not detected $ue(j)$ is 0 for all values of the length of the code words n when $j \leq 3$. But, although the number of incorrectly transmitted code words with 4 incorrectly transmitted bits $e(4)$ is smaller than the number of incorrectly transmitted code words with 3 incorrectly transmitted bits $e(3)$ for all values of the length of the code words n, there are code words in which the error in transmission is not detected, i.e., $ue(4)$ is number greater than 0, for all values of n. From Table I – Table IV, we can notice that this statement is valid regardless which of the four linear quasigroups given in Fig. 2 is used for coding.

TABLE I. NUMBER OF INCORRECTLY TRANSMITTED CODE WORDS OF LENGTH N BITS WITH J INCORRECTLY TRANSMITTED BITS $E[J]$ AND NUMBER OF INCORRECTLY TRANSMITTED CODE WORDS OF LENGTH N BITS WITH J INCORRECTLY TRANSMITTED BITS IN WHICH THE ERROR IS NOT DETECTED $UE[J]$ WHEN LINEAR QUASIGROUP OF ORDER 4 REPRESENTED WITH A_1 AND B_1 IS USED FOR CODING

n	$e(1)$	$ue(1)$	$e(2)$	$ue(2)$	$e(3)$	$ue(3)$	$e(4)$	$ue(4)$	$e(5)$	$ue(5)$	$e(6)$	$ue(6)$	$e(7)$	$ue(7)$
8	311267	0	69390	0	8851	0	747	156	33	0	4	0	0	0
12	242794	0	85002	0	18252	0	2577	75	269	0	20	0	1	0
16	189122	0	91422	0	26777	0	5557	39	854	0	81	2	12	0
20	148449	0	89405	0	34484	0	9419	36	1876	0	314	0	34	0
24	115659	0	84897	0	39858	0	13200	29	3438	0	683	1	111	0
28	90428	0	78811	0	42353	0	16870	16	5233	0	1322	0	276	0
32	71053	0	70200	0	44288	0	20091	20	7354	0	2238	1	571	0
36	55297	0	62120	0	44143	0	23172	8	9519	0	3181	0	918	0
40	43179	0	53762	0	43522	0	25178	12	11579	0	4339	0	1338	0
44	33904	0	46899	0	41185	0	26684	2	13388	0	5417	1	1945	0
48	26324	0	39439	0	38824	0	27521	5	15246	0	6917	0	2790	0
52	20285	0	33823	0	35602	0	27642	6	16865	0	8261	0	3546	0
56	15856	0	28606	0	32491	0	27367	5	17564	0	9682	0	4372	0
60	12385	0	23847	0	29204	0	26452	1	18716	0	10796	0	5231	0
64	9372	0	19807	0	26411	0	25357	1	19080	0	11951	0	6225	0

TABLE II. NUMBER OF INCORRECTLY TRANSMITTED CODE WORDS OF LENGTH N BITS WITH J INCORRECTLY TRANSMITTED BITS $E[J]$ AND NUMBER OF INCORRECTLY TRANSMITTED CODE WORDS OF LENGTH N BITS WITH J INCORRECTLY TRANSMITTED BITS IN WHICH THE ERROR IS NOT DETECTED $UE[J]$ WHEN LINEAR QUASIGROUP OF ORDER 4 REPRESENTED WITH A_2 AND B_2 IS USED FOR CODING

n	$e(1)$	$ue(1)$	$e(2)$	$ue(2)$	$e(3)$	$ue(3)$	$e(4)$	$ue(4)$	$e(5)$	$ue(5)$	$e(6)$	$ue(6)$	$e(7)$	$ue(7)$
8	310830	0	69263	0	8965	0	703	136	33	0	1	0	0	0
12	243374	0	85279	0	18124	0	2643	71	240	0	31	1	2	0
16	189713	0	91286	0	26793	0	5591	70	826	0	91	0	6	0
20	147502	0	90323	0	34141	0	9138	38	1896	0	285	1	35	0
24	116088	0	84892	0	39931	0	13285	30	3440	0	696	0	118	0
28	90723	0	78372	0	42336	0	16773	16	5311	0	1412	0	297	0
32	70973	0	70372	0	44218	0	20212	27	7347	0	2172	0	552	0
36	55576	0	61975	0	44270	0	23076	17	9304	0	3239	0	932	0
40	43002	0	54513	0	43042	0	25229	10	11556	0	4308	0	1387	0
44	33741	0	46601	0	41382	0	26625	8	13348	0	5656	1	2011	0
48	26289	0	39790	0	38743	0	27527	5	15144	0	6936	1	2699	0
52	20274	0	33747	0	35950	0	27924	8	16446	0	8183	0	3469	0
56	15890	0	28610	0	32231	0	27080	7	17917	0	17917	0	4492	0
60	12310	0	23993	0	29393	0	26250	2	18793	0	10677	0	5325	0
64	9662	0	19705	0	26038	0	25457	5	18990	0	11938	0	6287	0

TABLE III. NUMBER OF INCORRECTLY TRANSMITTED CODE WORDS OF LENGTH N BITS WITH J INCORRECTLY TRANSMITTED BITS $E[J]$ AND NUMBER OF INCORRECTLY TRANSMITTED CODE WORDS OF LENGTH N BITS WITH J INCORRECTLY TRANSMITTED BITS IN WHICH THE ERROR IS NOT DETECTED $UE[J]$ WHEN LINEAR QUASIGROUP OF ORDER 4 REPRESENTED WITH A_3 AND B_3 IS USED FOR CODING

n	$e(1)$	$ue(1)$	$e(2)$	$ue(2)$	$e(3)$	$ue(3)$	$e(4)$	$ue(4)$	$e(5)$	$ue(5)$	$e(6)$	$ue(6)$	$e(7)$	$ue(7)$
8	311376	0	69258	0	8894	0	699	139	30	0	1	0	0	0
12	243207	0	85365	0	18165	0	2688	84	276	0	19	3	0	0
16	189436	0	90899	0	27073	0	5640	46	860	0	92	0	9	0
20	147741	0	90074	0	34241	0	9417	36	1887	0	298	0	36	0
24	116000	0	84931	0	39700	0	13356	34	3409	0	696	0	109	0
28	90742	0	78143	0	42641	0	17084	17	5165	0	1344	0	274	0
32	71209	0	70229	0	44559	0	20147	20	7086	0	2169	0	572	0
36	55267	0	62403	0	43888	0	23208	10	9444	0	3211	0	927	0
40	43342	0	54203	0	43367	0	25064	8	11463	0	4295	0	1366	0
44	33805	0	46783	0	41265	0	26482	8	13483	0	5683	0	2046	0
48	26224	0	39816	0	38999	0	27474	5	15002	0	6810	0	2828	0
52	20372	0	33918	0	35661	0	27623	4	16365	0	8328	0	3614	0
56	16104	0	28244	0	32840	0	27125	7	17640	0	9409	0	4465	0
60	12422	0	23851	0	29400	0	26295	3	18365	0	10939	0	5321	0
64	9623	0	19826	0	26042	0	25083	4	19287	0	11938	0	6202	0

TABLE IV. NUMBER OF INCORRECTLY TRANSMITTED CODE WORDS OF LENGTH N BITS WITH J INCORRECTLY TRANSMITTED BITS $E[J]$ AND NUMBER OF INCORRECTLY TRANSMITTED CODE WORDS OF LENGTH N BITS WITH J INCORRECTLY TRANSMITTED BITS IN WHICH THE ERROR IS NOT DETECTED $UE[J]$ WHEN LINEAR QUASIGROUP OF ORDER 4 REPRESENTED WITH A_4 AND B_4 IS USED FOR CODING

n	$e(1)$	$ue(1)$	$e(2)$	$ue(2)$	$e(3)$	$ue(3)$	$e(4)$	$ue(4)$	$e(5)$	$ue(5)$	$e(6)$	$ue(6)$	$e(7)$	$ue(7)$
8	311390	0	69406	0	8680	0	733	128	41	0	0	0	0	0
12	242695	0	85603	0	18483	0	2594	82	266	0	25	1	2	0
16	189352	0	91100	0	26962	0	5582	45	829	0	97	1	6	0
20	147875	0	90019	0	34315	0	9138	44	2027	0	275	0	47	0
24	116080	0	84705	0	39707	0	13255	29	3414	0	697	0	122	0
28	90761	0	77779	0	42856	0	17234	18	5357	0	1297	0	287	0
32	70995	0	70231	0	44288	0	20512	26	7206	0	2096	0	551	0
36	55358	0	62099	0	44350	0	23047	19	9304	0	3258	0	907	0
40	43180	0	54014	0	43254	0	25115	9	11560	0	4386	0	1413	0
44	33694	0	46677	0	41429	0	26453	7	13380	0	5670	0	1951	0
48	25990	0	40019	0	38799	0	27356	4	15228	0	7024	0	2859	0
52	20418	0	33652	0	35614	0	27549	2	16585	0	8388	0	3643	0
56	16017	0	28416	0	32356	0	27071	3	17712	0	9711	0	4418	0
60	12321	0	23707	0	29133	0	26604	5	18939	0	10758	0	5324	0
64	9581	0	19818	0	26160	0	25476	1	19020	0	11817	0	6200	0

From Table I - Table IV we see that although the number of incorrectly transmitted code words with less than 4 incorrectly transmitted bits is large, the code succeeds to detect all of them. But, some of the code words with 4 incorrectly transmitted bits pass undetected and the receiver accepts them as correctly transmitted code words. The above assertion holds for all four linear quasigroups of order 4. The obtained experimental results show that regardless which linear quasigroup of order 4 from the best class of quasigroups of order 4 for which the matrix C in the representation (2) is zero matrix is used for coding, the code surely detects up to 3 incorrectly transmitted bits.

V. CONCLUTION

The experimental results obtained using simulations show that the number of errors that the code considered in this paper surely detects is equal for all linear quasigroups of order 4 from the best class of quasigroups of order 4 for coding for which the matrix C in the linear representation (2) is zero matrix. When linear quasigroup of order 4 from the best class of quasigroups of order 4 for which the matrix C in the representation (2) is zero matrix is used for coding, the code surely detects up to 3 incorrectly transmitted bits, regardless of the length of the code words.

ACKNOWLEDGMENT

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

REFERENCES

[1] N. Ilievska and V. Bakeva, "A model of error-detecting codes based on quasigroups of order 4," Proc. 6th Int. Conf. on Informatics and Information Technologies, R. Macedonia, pp. 7-11, 2008.

[2] V. Bakeva and N. Ilievska, "A probabilistic model of error-detecting codes based on quasigroups," Quasigroups and Related Systems, vol. 17, no. 2, pp. 135-148, 2009.

[3] N. Ilievska and D. Gligoroski, "An Error-Detecting Code based on Quasigroups," Proceedings of 11th International Conference for Informatics and Information Technology, Bitola, Republic of Macedonia, pp. 285-290, 2014.

[4] S. A. Vanstone and P. S. van Ooschot, "An Introduction to Error Correcting Codes with Applications," Kluwer Academic Publishers, Boston, 1989.

[5] T. Klove, "Codes for Error Detection," World scientific, 2007.

On the Complexity of Computing Hamming Weight for Greedy Constructions

Dejan Spasov
Faculty of Computer Science and Engineering
Skopje, Macedonia
dejan.spasov@finki.ukim.mk

***Abstract*— We are interested in generating linear codes. Greedy algorithms are one of the oldest known methods for code construction. They are simple to define and easy to implement, but require exponential running time. Codes obtained with greedy algorithms have very good parameters. Greedy algorithms rely on efficient algorithm for computing the Hamming weight of a vector. In this paper we give an overview of the algorithms for computing the Hamming weight.**

Keywords—error correcting codes, Hamming weight, covering radius, minimum distance, linear codes

I. INTRODUCTION

Let F_q be a finite field of q elements and let F_q^n be n-dimensional vector space over F_q. Then a *code* C is subset of F_q^n of M elements. Elements of the code $c_i \in C$ are called *codewords*.

Let $d(x, y)$ denote the *Hamming distance*, i.e. the number of coordinates in which two vectors x and y differ, and let $wt(x)$ denote the (*Hamming*) *weight*, i.e. the number of nonzero coordinates of x. We say that a code C has *(minimum) distance* d if

$$d = \min\{d(c_i, c_j)\}, \forall c_i, c_j \in C, i \neq j \quad (1)$$

A code C is linear if its codewords form k- dimensional linear subspace over F_q^n. We will write $[n, k, d]$ to denote a linear code over the field F_q^n. For linear codes there exist k basis vectors that are kept as rows in a matrix G, called the *generator matrix*. For each linear code there is a generator matrix of type $G = [I \quad A]$ for which we say that is in *standard form*. It is well-known that for linear codes there exist a so-called *parity check matrix* H, such that $\forall c_j \in C, Hc_j = 0$. Let $G = [I \quad A]$ be the generator matrix, then $H = [-A^T \quad I]$ is the parity check matrix of the same code.

The following theorem is a fundamental result in coding theory:

Theorem 1 [1]: A code C with parameters $[n, k, ?]$ and parity check matrix H has minimal distance d if every linear combination of $d - 1$ columns of H is linearly independent, and there exist a linearly dependent combination of d columns of H.

The *covering radius* ρ of a code C is the largest possible distance between the code C and a vector from F_q^n, i.e.

$$\rho = \max_{x \in F_q^n} \min_{\forall c \in C} d(x, c) \quad (2)$$

We will use $(x|y)$ to denote concatenation of two strings, and x^k to denote a string of k symbols x, namely $\underbrace{x \cdots x}_{k}$.

Fundamental problem in coding theory is finding *optimal codes*. A code (n, M, d) is optimal if it has maximal number of codewords M for a given n and d. In general, finding an optimal code is a difficult problem. Trivial way to do this is by super-exponential search over all possible orderings of the field F_q^n. For small fields $(q \leq 9, n \leq 256)$ there exist tables of best known (some of them optimal) codes [2], but for larger spaces optimal-code parameters can be estimated with the Gilbert-Varshamov bound and its asymptotical variant.

In this paper we are looking for best codes for a given minimum distance. We use greedy algorithms in our search. Bottleneck in greedy algorithms is computing the Hamming weight of a 64-bit word. In Section II we give an overview of greedy algorithms. In Section II we give an overview of algorithms for computing Hamming weight. In Section III we present performance of the greedy algorithms with various algorithms for computing the hamming weight.

II. GREEDY ALGORITHMS

It is well-known that a simple greedy procedure produces code family with parameters that follow the Gilbert-Varshamov bound

$$R \geq 1 - H(\delta). \quad (3)$$

In binary case no better code family to-date is known, but, on the other hand, greedy algorithms are considered impractical due to its exponential time complexity. In this section we give an overview of the greedy algorithms.

A. Gilbert's Construction

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

In general, Gilbert's algorithm produces a nonlinear (n, M, d) code. Given the code length n and its minimum distance d, the algorithm will search the entire space F_q^n and greedily will add to C the first vector x that is at distance d from C, i.e. $d(x, c) \geq d, \forall c \in C$.

B. Jenkins' Construction

The idea in this construction is to build the A matrix of a systematic code $G = [I \quad A]$ [3]. Each $x \in F_q^{n-k}$ is considered as a row in A. Then for each i-linear combinations of rows of A, where $i \leq (d-2)$, we check if it has Hamming weight at least $d - i$.

C. Lexicographic Construction

Given a code $[n, k, d]$ with generator matrix G_k. Let for each syndrome s we denote with $w(s)$ the Hamming weight of the coset leader $e(s)$. Let assume that the pairs $(s, s(s))$ for the code $[n, k, d]$ are kept in a look-up table. The Lexicographic Construction is iterative algorithm that can be described in 3 steps. In the first step, using linear search over $(s, w(s))$, the algorithm finds the covering radius ρ. In the second step it picks arbitrary syndrome s with weight $w(s) = \rho$ and forms a new codeword with the c onstruction $c_{k+1} = (1^{d-\rho}|0^k|s)$. In the third step the algorithm builds the table $(s_{k+1}, w(s_{k+1}))$ from $(s, w(s))$.

Given two syndromes s_k and s. The *companion set* of the syndrome s_k with respect to the syndrome s is the set:

$$K_{s_k} = \{y_i | y_i = s_k + i \cdot s, i \in F_q\} \tag{5}$$

We use the concept of companion sets to easily explain the creation of the new table $(s_{k+1}, w(s_{k+1}))$:

Theorem 6 [4]: Given ρ, s_k, and $(s, w(s))$. The table $(s_{k+1}, w(s_{k+1}))$ can be constructed with the following minimization:

$$w(s_{k+1}) = \min_{\substack{y_i \in K_{s_{k+1}} \\ i \in F_q}} \left(wt(v + i^{d-\rho}) + w(y_i)\right) \tag{8}$$

for each syndrome $s_{k+1} = (v|s)$.

D. Hybrid greedy algorithm

The algorithm that we use for finding new best codes is a combination of the Jenkins Algorithm and the Lexicographic Construction (Fig. 1). The hybrid algorithm is initialized with the lexicographic construction and the lexicographic construction runs until half of the RAM memory is used. The algorithm stores weights of the coset leaders in the RAM memory. After entire RAM has been used, and a code C_L is generated, the hybrid algorithm switches to Jenkins algorithm. However, coset leaders are kept in the RAM memory and the distance between the subcode C_L and a vector x is computed from the content in the RAM.

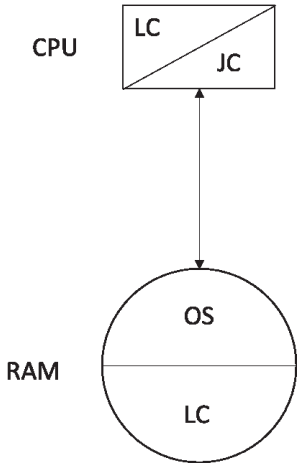


Figure 1. Hybrid Greedy Algorithm

III. Algorithms For Computing Hamming Weight

Hamming weight of a string is the number of symbols that are different from the zero symbol. The problem of computing the Hamming weight has been widely studied. The Hamming weight of a vector of length n may be computed sequentially with n additions or in parallel with $\log n$ additions. Fig. 2, demonstrates computing the weight of the binary string 00011011.

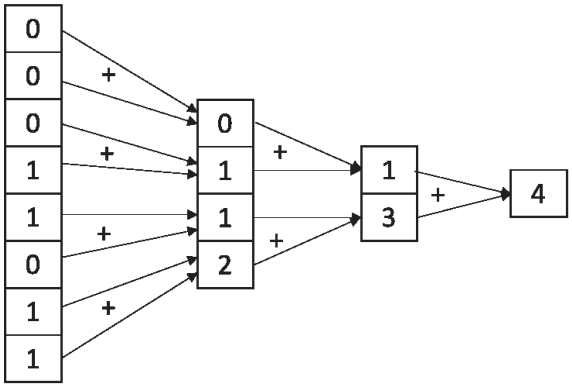


Figure 2. Parallel Computing of the Hamming weight

Bellow we list several parallel algorithms [5]. Let for simplicity we denote

```
m1  = 0x5555555555555555;
m2  = 0x3333333333333333;
m4  = 0x0f0f0f0f0f0f0f0f;
m8  = 0x00ff00ff00ff00ff;
m16 = 0x0000ffff0000ffff;
m32 = 0x00000000ffffffff;
hff = 0xffffffffffffffff;
h01 = 0x0101010101010101;
```

then, an algorithm may be devised as follows:

```
int Algorithm1(uint64_t x){
    x = (x&m1 ) + ((x>> 1)&m1);
    x = (x&m2 ) + ((x>> 2)&m2);
    x = (x&m4 ) + ((x>> 4)&m4);
    x = (x&m8 ) + ((x>> 8)&m8);
    x = (x&m16) + ((x>>16)&m16);
    x = (x&m32) + ((x>>32)&m32);
    return x;
}
```

The above algorithm uses 24 arithmetic operations. The following algorithm reduces the number of arithmetic operations:

```
int Algorithm2(uint64_t x){
    x -= (x >> 1) & m1;
    x = (x & m2) + ((x >> 2) & m2);
    x = (x + (x >> 4)) & m4;
    x += x >>  8;
    x += x >> 16;
    x += x >> 32;
    return x & 0x7f;
}
```

Algorithm2 is performed with 17 arithmetic operations. The following algorithm Computes Hamming weight in 12 arithmetic operations, but one of them is multiplication:

```
int Algorithm3(uint64_t x){
    x -= (x >> 1) & m1;
    x = (x & m2) + ((x >> 2) & m2);
    x = (x + (x >> 4)) & m4;
    return (x * h01) >> 56;
}
```

Given considerable amount of RAM memory, another way to compute the Hamming weight is to precompute Hamming weights up to a certain number and to store precomputed weights in a look-up table. Then, use the number, for which we want to compute Hamming weight, as index for the look up table. Since we are operating with 64-bit numbers, a look up table for 64-bit numbers is prohibitively large, i.e. 2^{64} bytes. Therefore, for practical implementations 64-bit number is split into two or more smaller numbers, e.g. two 32-bit numbers, four 16-bit numbers, or 8 8-bit numbers. Hamming weights of 32-bit numbers may be stored in a look-up table of size 1 GB. Hamming weights of 16-bit numbers may be stored in a look-up table of size 64 KB. Hamming weights of 8-bit numbers may be stored in a look-up table of size 256 Bytes. Hence, Hamming weight of a 64-bit number may be obtained as sum of Hamming weights of smaller numbers, where the Hamming weight of smaller numbers is obtained from a look-up table. The following code snippet demonstrates computation of Hamming weight with look-up table:

```
//initialization
char *lookup = new char[0xffffffff];
for (i = 0; i<=0xffffffff; ++i)
    hw_table[i] = Algorithm1(i);

//Algorithm 4
int *h1 = (int *)&x, *h2 = h1 + 1;

wt = hw_table[*h1]+hw_table[*h2];
```

In the initialization stage, a look up table of size 1 GB is created and for each 32-bit integer from 0 to 0xffffffff, the Hamming weight is computed with any of the prior algorithms, i.e. Algorithm1, Algorithm2, or Algorithm3. Computed Hamming weights are stored in the appropriate entry in the look up table. Next, we want to compute the Hamming weight of a 64-bit integer x. The integer x is considered as concatenation of two 32-bit numbers. The two pointers h1 and h2 point to the two 32-bit numbers. Hamming weights of the two 32-bit numbers are obtained from the lookup table. The Hamming weight of the 64-bit number x is obtained as sum of the weights of the two 32-bit numbers. This approach uses two memory reads and one arithmetic operation.

The following code snippet uses look-up table of size 64 KB to store Hamming weights of 16-bit numbers. A 64-bit number x is considered concatenation of 4 16-bit numbers.

```
//initialization
char *lookup = new char[0xffff];
for (i = 0; i<=0xffff; ++i)
    hw_table[i] = Algorithm1(i);

//Algorithm 5
short *h1=(int*)&x,*h2=h1+1,
      *h3=h2+1,*h4=h3+1;

wt = hw_table[*h1]+hw_table[*h2]
    +hw_table[*h3]+hw_table[*h4];
```

The above algorithm uses 4 memory reads and 3 additions. However, the advantage of Algorithm5 over Algorithm4 may be that the look-up table may be stored in L2 cache, thus providing faster computation of the Hamming weight.

The following code snippet uses look-up table of size 256 Bytes to store Hamming weights of 8-bit numbers. A 64-bit number x is considered concatenation of 8 8-bit numbers.

```
//initialization
char *lookup = new char[0xff];
for (i = 0; i<=0xff; ++i)
    hw_table[i] = Algorithm1(i);
```

```
//Algorithm 6
short *h1=(int*)&x,*h2=h1+1,
      *h3=h2+1,*h4=h3+1,*h5=h4+1,
      *h6=h5+1, *h7=h6+1,*h8=h7+1;

wt = hw_table[*h1]+hw_table[*h2]
    +hw_table[*h3]+hw_table[*h4]
    +hw_table[*h5]+hw_table[*h6]
    +hw_table[*h7]+hw_table[*h8];
```

The above algorithm uses 8 memory reads and 7 additions. However, the advantage of Algorithm6 over Algorithm5 may be that the look-up table may be stored in L1 cache, thus providing faster computation of the Hamming weight.

IV. CONCLUSION

Assuming that 64-bit arithmetic operation is performed in 3 to 4 cycles, we conclude that Algorithm1 may be performed in 75 to 100 cycles. In similar fashion, we may conclude that Algorithm2 may be performed in 50 to 70 cycles. If we assume that execution time of the multiplication is few cycles more than the execution time of addition, then we may assume that execution time of Algorithm3 is 40 to 55 cycles. Therefore, we conclude that Algorithm3 outperforms Algorithm1 and Algorithm2. Figure 3 confirms our reasoning.

Figure 4 shows performance comparisons of Algorithms 4, 5, and 6. Figure 4 shows that the best performance achieves algorithm 5, who splits the 64 bit word to 16-bit half words. Our explanation of this result is that the 64 KB look-up table for the 16-bit words is uploaded to L2 cache; thus, achieving best performance.

Analyzing figures 4 and 5 we can conclude that Algorithm 5 is fastest algorithm for computing the Hamming Weight.

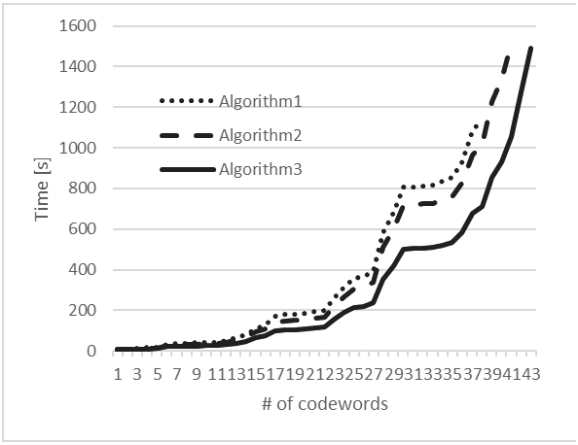


Figure 3. Performance comparison of Algorithm1, Algorithm2, and Algorithm3

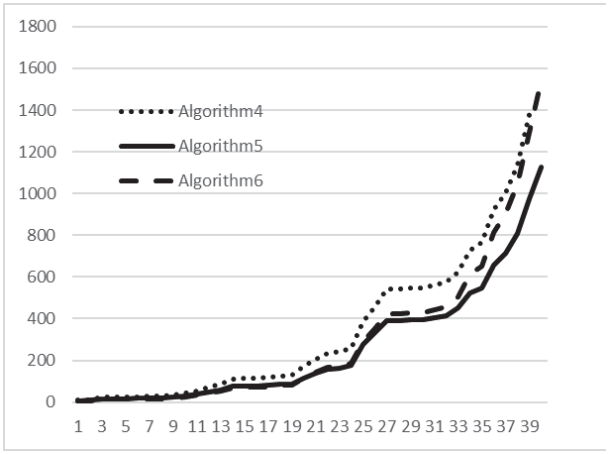


Figure 4. Performance comparison of Algorithm4, Algorithm5, and Algorithm6

REFERENCES

[1] F. J. MacWilliams, N.J.A. Sloane. The Theory of Error-Correcting Codes. North Holland, Amsterdam, 1977.

[2] M. Grassl "Bounds on the minimum distance of linear codes and quantum codes." Online available at: http://www.codetables.de.

[3] B. Jenkins B. "Tables of lexicodes." Online available at: http://burtleburtle.net/bob/math/lexicode.html..

[4] Spasov, D. “Some properties of good codes”. PhD Thesis, Ss. Cyril and Methodius University, Skopje, 2010.

[5] Wikipedia "Hamming Weight." Online available at: https://en.wikipedia.org/wiki/Hamming_weight

Enterprise Information Security and Risk Management

1st Ljubica Panova
Faculty of Computer Science and Engineering
Ss. Cyril and Methodius University
Skopje, Macedonia
ljubica.panova@students.finki.ukim.mk

2nd Vesna Dimitrova, PhD
Faculty of Computer Science and Engineering
Ss. Cyril and Methodius University
Skopje, Macedonia
vesna.dimitrova@finki.ukim.mk

***Abstract*—During the past decade, businesses have become increasingly dependent on technology. Technology trends indicate that IT environments will continue to develop and to become more complicated and connected, which also increases the risk of enterprise's data compromising. This paper discusses information security management including information security policy as its basis, the importance of information security risk management and the process itself, as well as the latest potential enterprise's information security threats.**

Keywords—risk management, information security management, threat, enterprise, policy

I. Introduction

The information is probably the most valuable asset of many companies and organizations. In a relatively short period of time the company's data has moved from paper files and filing cabinets to electronic files stored on powerful computers. When we talk about information security today we don't think about locking paper files in a file room within the office but securing data on computers accessed on networks or via the Internet.

Information security is a mainstream business issue that affects every stakeholder. As a result, the companies are taking measures to protect the information from the risk of being stolen, modified, destroyed or misused by increasing their investment in information security.

This paper explains the main components of Enterprise Information Security and Risk Management. Section II defines Enterprise Governance, Information Security Management and what an Information Security Policy includes. Section III explains what a risk is and what it consists of. Section IV defines Enterprise Risk Management, the elements of Information Risk Management Policy, the risk management process and we conclude by discussing the latest information security threats.

II. Enterprise Governance and Information Security Management

Enterprise governance is the structure and relationship that control, direct, or regulate the performance of an enterprise (and its) projects, portfolios, infrastructure and processes [1].

Information security management is one of the main components of enterprise governance. It includes all the processes that affect enterprise assets in terms of their confidentiality, availability and integrity, as well as their risks and treats. Information security management involves a risk management process, which includes IT risks, Human Resources (HR) risks, service risks etc. This means that information security management has a greater scope than IT management and it reports directly to the company board of directors.

A. Information Security Policy

As part of information security management, every company should have an Information Security Policy (ISP). The information security policy tells staff members what they can do, what they cannot do, what they must do and what their responsibilities are. Also, a difference should be made between a policy and a procedure. The policy states what should be done and the procedure explains how to implement the policy. Here comes the question, what exactly the policy has to include:[2]

- **Responsibilities –** Every employee is responsible for the information security in the company. The employees have to understand and comply with the company's security policy.
- **Access Control Standards -** Users need to have access only to the information and applications they need to perform their work tasks, and nothing more. They are responsible for managing the necessary changes to their passwords.
- **Accountability-** Users are accountable for all activities carried out under their user accounts.
- **Audit trails -** The following activities carried out by every user must be logged: At least 30 days of login details; All unauthorized attempts to read, write or delete data; Applications must provide detailed audit trails of changing data when required by the business.
- **Continuity plans -** Regarding the continuity of the work, there should be a contingency plan for all

computer services that support critical systems and that plan is designed, implemented and tested. The business continuity plan should identify the services that are critical to the operation of the business and ensure that there are plans to handle unpredictable situations.

- **Backups** - All software and user data have to be backed up regularly to alternative media and stored in a safe and remote place. The frequency of backups must be specified in the policy and has to be appropriate to the importance of the system.
- **E-mail –** Business emails cannot be used for private conversations. The employees couldn't have any expectations of privacy respect to any sent or received mail. It should also be considered that, legally, an email carries the same weight as a letter on company letterhead.
- **Downloading from the Internet -** Most companies allow their employees access to the Internet. The policy must clearly state that employees have access to the Internet for business use only.
- **Portable devices -** Laptop computers today are just as powerful as desktops and create new problems because they can easily be taken out of the office and can easily be stolen. Employees should be aware that they are responsible for the equipment issued to them. If employee's laptop is stolen it means that the information in it are stolen too. The information can have more value than the value of the laptop itself.
- **Media disposal -** The way storage media that contain confidential data are destroyed when they are no longer needed, must be carefully considered because there is a danger for an unauthorized person to retrieve data from the media if they are not properly deleted. The policy should determine the method of removal, depending on which type of media and what kind of data is involved.
- **Noncompliance –** The policy has to contain a reference to the consequences of noncompliance and disciplinary actions.
- **Other –** The policy may include other elements depending on the needs of the company. It is important to be simple and understandable, because the goal is for all employees to understand and apply it.

III. WHAT IS A RISK?

ISO 31000 defines risk as "effect of uncertainty on objectives". This effect can be both positive or negative. That means, according ISO 31000 standard risk is neutral. However, it is usually considered negative. [5]

Risk can be described by a mathematical formula: [3]

$$Risk = Threat * Vulnerability * Asset\ value \quad (1)$$

Every element of this formula could be defined as follows:

- ***Risk*** is any event that could impact a business and prevent it from reaching its corporate goals.
- ***Threat*** is the possibility that the enterprise will be exposed to an incident that has an impact on the business.
- ***A vulnerability*** is the point of weakness, it is the Achilles' heel where, despite the tough armor shielding the rest of the system, the attack is launched and may open the entire system or network to compromise. However, if risk is viewed as a potentially positive scenario, one should replace the term "vulnerability" with the term "opportunity". In this scenario, the key is to recognize and exploit the opportunity in a timely manner so that the maximum benefit of the risk is realized.
- ***The asset*** is the component that will be affected by the risk. Quantitative risk analysis attempts to describe risk from a purely mathematical viewpoint, fixing a numerical value to every risk and using that as a guideline for further risk management decisions. [3]

IV. ENTERPRISE RISK MANAGEMENT

According ISO 31000, Enterprise Risk Management (ERM) is a strategic organizational approach that supports the achievement of the institution's objectives by addressing the full spectrum (reputational, strategic, financial, operational and compliance) of its risks and managing the combined impact as an interrelated set of risks. [5]

On a practical level, ERM means managing all risks that the enterprise can be exposed to. There are five ways to deal with any given risk:

- Ignore it;
- Reduce it;
- Eliminate it;
- Transfer it;
- Accept it;

A. Information security risk management

Information Security Risk Management (ISRM) is an important element of the enterprise risk management in the companies today. It can be defined as follows:
Information security risk management is the process of managing risks associated with the use of information technology. It involves identifying, assessing, and treating risks to the confidentiality, integrity and availability of an organization's assets. The end goal of this process is to treat risks in accordance with an organization's overall risk tolerance. Businesses shouldn't expect to eliminate all risks; rather, they should seek to identify and achieve an acceptable risk level for their organization. [6]
Proper risk management requires a strong commitment and support from the top management/board of directors, a documented process that supports the organization's mission, an information risk management (IRM) policy and a delegated IRM team.

a) Information risk management policy

Information risk management (IRM) policy should begin with a high-level policy statement and support goals, scope, constraints, responsibilities, and approach. This policy should address the following:

- Objectives of the IRM team;
- Information security risks categorization;
- Clearly defined levels of information security risks that are acceptable for the enterprise;
- Defined methodologies for risk identification, accessing and monitoring;
- Roles and responsibilities for managing and reporting information security risks;
- Connection between the IRM policy and the company's strategic planning processes;

The IRM policy should be communicated effectively and enforced to all parties.

b) Risk management process

The processes of identifying, analyzing and assessing, mitigating, or transferring risk are generally characterized as Risk Management. [4]

Stages of ISRM:

- Identification – Identification process includes identification of assets (what data assets will be considered), identification of vulnerabilities (what vulnerabilities are putting the confidentiality, integrity and availability of the assets at risk or could result in information being compromised), identification of threats (what are the causes of information becoming compromised) and identification of controls (what has been already in place to protect identified assets).
- Assessment - This process includes combining and analyzing the information that have been gathered about assets, vulnerabilities, threats and controls to define a risk. There are many frameworks and approaches for making risk assessment.
- Treatment - Once a risk has been assessed and analyzed, the enterprise has to select treatment options:
    - Elimination – Implementing a control that fully or nearly fully fixes the underlying risk;
    - Mitigation – Reduction in the extent of exposure to a risk and/or the likelihood of its occurrence;
    - Transference – Transferring the risk to another party so the company can recover from incurred costs of the risk being realized;
    - Risk acceptance – If it is considered that the risk is low and the time and effort to fix the risk costs more than the costs that would be incurred if the risk were to be realized, then the company accepts the risk.
    - Risk avoidance – Removing all exposure to an identified risk
- Monitoring – It is an important part of the risk management process. When recommended risk mitigation measures have been acquired/developed and implemented it is time to begin and maintain a process of monitoring IRM performance. This can be done by periodically reassessing risks to ensure that there is sustained adherence to good control or that failure to do so is revealed, consequences considered, and improvement, as appropriate, duly implemented [4].
- Communication – And last but not least, communication plays a significant role in this process. Regardless of how the risk is treated, the decision needs to be communicated within the company. Stakeholders need to understand the costs of treating or not treating a risk. Responsibility and accountability needs to be clearly defined and associated with individuals and teams in the company to ensure the right people are engaged at the right times in the process. [6]

c) Establish Risk Acceptance Criteria

As mentioned above, if it has been considered that the risk is low, it can be accepted. But how can be a certain risk classified as low? There must be some criteria that will define which events are considered as acceptable risk and in which situations. The management and the IRM team should establish the maximum acceptable financial risk. For example: "Do not accept more than a 1 in 100 chances of losing $200,000 in a year".

B. Information security threats today

Every year, The Information Security Forum (ISF), a global, independent information security body considered the world's leading authority on cyber security and information risk management, releases its 'Threat Horizon' report to provide a forward-looking view of the biggest security threats over a two-year period. According to them, these are the top nine threats to watch through 2018: [7]

- **Theme 1: Technology adoption dramatically expands the threat landscape** – ISF predicts that personal and business technology dependence will increase in the next two years. Enterprises try to increase their effectiveness and efficiency through improved integration and connectivity, but they will also open themselves to associated threats.
    - ***The IoT leaks sensitive information*** - The Internet of Things (IoT) is growing very fast and many organizations are adopting IoT devices. But they can potentially create a backdoor into organizations compromising sensitive information.
    - ***Opaque algorithms compromise integrity*** – Companies are increasingly using algorithms to make decisions and operate critical systems. Without a human included in those decisions, the company has less visibility and cannot be sure how the system works and

interacts. Unintended interactions between algorithms can be a significant security risk that affects information integrity.

- ***Rogue governments use terrorist groups to lunch cyberattacks*** – According to ISF in the next two years the support of terrorist, that is already present, will expand to include cyberattack capabilities. This would result in more damaging cyberattacks than many organizations have ever experiences.

- **Theme 2: Ability to protect is progressively compromised** – The established methods of information risk management will be compromised by a variety of actors, usually non-malicious.
  - ***Unmet board expectations exposed by a major incident*** – In the past years, IFS stated that not appreciating the value of security by the board is a top threat. Today boards approve more money for information security budgets and want to see immediate results. They have great expectations because they see the security the same way they see any other business issue which cannot be compared.
  - ***Researches silenced to hide security vulnerabilities*** – Researchers regularly uncover information security vulnerabilities and make them public in order to improve security. But manufacturers, instead of fix the vulnerabilities working together with the researches, have started responding with legal actions. IFS predict that this trend will silence the researchers. That will leave the software with vulnerabilities which manufacturers prefer to hide rather than repair.
  - ***Cyber insurance safety net is pulled away*** – ISF believes that in the next two years several information security attacks will result in significant financial losses for insurance companies which offer cyber insurance. As a result, many insurance companies will withdraw from the market and the rest of them will set more stringent requirements and rules for their clients.

- **Theme 3: Governments become increasingly interventionist** – Governments around the world will become more interested in examination of new and existing technologies used by the citizens. The IFC believes that governments will start adoption of high-pressure in dealing with organizations and companies that handle personal information.
  - ***Disruptive companies provoke governments*** – Enterprises with aggressive commercial strategies probably will animate politicians and regulators to take a closer look at the domestic impact of new technologies. The bad thing is that such reactions will neither increase the protection of personal data of the population nor encourage economic growth.
  - ***Regulations fragment the cloud*** – Companies that use cloud services could suffer a heavy impact if legislative changes happen about the way how personal data are collected, stored and exchanged. According to the ISF the companies could face a serious problem to conduct their business as usual, while trying to remain compliant with the new data protection and data localization requirements, that could start implementing in the next two years.
  - ***Criminal capabilities expand gaps in international policing*** – Unfortunately, the cybercriminal today reach a level with governments and other organizations. That will affect the companies and organizations by decreasing the ability of the security mechanisms they are using today. The criminals usually attack companies outside their country, so they will use the lack of cross-border cooperation between low enforcement agencies.

V. Conclusion

The main thing that needs to be emphasized is that in the information security and risk management involves the whole enterprise with senior leadership involvement. There are many threads to an enterprise's data and if these issues are ignored, there could be unwanted consequences, including legal, financial and reputational harm. Due to increased complexity, uncertainty and threats from a wide range of sources, information security risk management within the enterprise has become a key challenge and core competence for enterprise's sustainable success. That's why the company must keep up with the new technologies, at the same time to be aware of the new potential information security threats and find appropriate protection mechanisms.

References

[1] Wilson, W.L, 2009, Conceptual Model for Enterprise Governance, Ground System Architectures Workshop.

[2] Brian Shorten, CISSP, CISA – Information Security Policies from the Ground Up (Information Security Management Handbook, Fifth Edition – Harold F. Tipton, Micki Krause);

[3] Kevin Henry, CISA, CISSP – Risk Management and Analysis (Information Security Management Handbook, Fifth Edition – Harold F. Tipton, Micki Krause);

[4] Chris Hare, CISSP, CISA – Policy Development (Information Security Management Handbook, Fifth Edition – Harold F. Tipton, Micki Krause);

[5] ISO 31000 Risk Management

[6] https://www.rapid7.com/fundamentals/information-security-risk-management/

[7] https://www.securityforum.org/

A Survey on Applications of Blockchain Technology

Daniela Mechkaroska Vesna Dimitrova Aleksandra Popovska-Mitrovikj
Faculty of Computer Science and Engineering,
Ss. Cyril and Methodius University, Skopje, Macedonia
Emails: daniela-mec@hotmail.com,{vesna.dimitrova, aleksandra.popovska.mitrovikj}@finki.ukim.mk

***Abstract*—Blockchain is a distributed database of records or public ledger of all timestamped transactions saved in all computers in one peer-to-peer network. It allows a secure and transparent transfer of digital goods including money and intellectual property.**

Bitcoin is the first major Blockchain innovation, a digital decentralized cryptocurrency. The public key cryptography provides the confidance of this currency. Exchanging a value or assets between two owners based on a set of conditions is included in an agreement called Smart contract.

This paper is a survey about Blockchain in a peer-to-peer network. Here we describe the basic principles of Blockchain technology and its major innovations and applications.

***Keywords*— Blockchain, Bitcoin, cryptocurrency, Smart contract, peer-to-peer network.**

I. Introduction

A Blockchain is a shared distributed database, which can be agreed upon a peer-to-peer network. It contains a connected sequence of blocks, holding timestamped transactions [3]. The confidence is based on the security provided by the public-key cryptography and verified by the network community. Once an element is appended to the Blockchain, it cannot be changed. This feature makes the Blockchain unchangeable.

In year 2008, an individual or group writing under the name of Satoshi Nakamoto published a paper entitled Bitcoin: A Peer-To-Peer Electronic Cash System. Bitcoin, which is the first major application of Blockchain, is a peer-to-peer version of the electronic cash that allows direct online payments from one party to another without the need of trusted third party. This cryptocurrency uses cryptography to secure transactions. Everybody who installs the open source program that implements this new protocol can become part of the bitcoin peer-to-peer network [3].

The other major application of Blockchain is called the Smart contract. Exchanging a value or assets between two owners based on a set of conditions is included in an agreement called Smart contract, which is controlled by decentralized consent on a Blockchain. Smart contract control the transactions between entities which can agree between themselves through the Blockchain and therefore there is no need of central authority [6].

II. How Blockchain works?

Blockchain is a decentralized ledger of transactions, distributed on all computers in one peer-to-peer network where all details of transactions are visible to everyone connected to the network. Namely, this is a growing list of linked blocks. The blocks consist of valid transactions, a timestamp and a hash pointer as a link to the previous block in the chain.

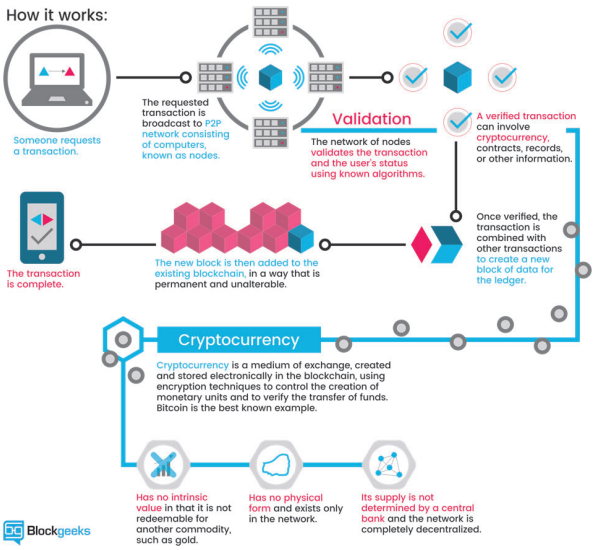


Fig. 1. Blockchain technology [1]

When someone requests a transaction or two parties exchange data, this could be money, contract or any other asset that can be digitally described, the requested transaction is broadcast to peer-to-peer network consisting of computers called nodes. This network of nodes validates the transaction and the status of the user using known algorithms. Depending on the network‘s parameters, the transaction is either verified immediately or transcribed into a secured record and placed in a list of pending transactions. Each block is identified by a hash and contains a header (a reference to the hash of the previous block) and a group of transactions. The sequence of linked blocks creates a secure, interdependent chain. Blocks must first be validated to be added to the Blockchain. When a block is verified, it is distributed through the network (added to the existing chain) and each node adds the block to the majority Blockchain. Then the transaction is completed.

When a malicious user tries to submit an altered block in the chain, the hash function of that block and all following

blocks would change. These changes will be detected by the other nodes and they will reject the block from the majority Blockchain, preventing corruption.

The Blockchain can be Public, Private or Permissioned. The most popular Public Blockchain innovation are Bitcoin and Ethereum, digital currencies [4]. In Private Blockchain the number of users is limited and the permissions to read and write data are controlled by one trusted organization. On the other hand, Permissioned Blockchain is a type of Private Blockchain where a few selected nodes control the permissions.

III. Applications

A. Bitcoin

Bitcoin is the first decentralized peer-to-peer cryptocurrency.

Decentralized means non-existence of a central bank. Emitters of this currency are the users or holders of ”mining” computers who verify the transactions.

The security provided by cryptography with public key gives **the confidence** of this currency.

Authentication of the transactions for one to another peer in Blockchain network is made with digital signature.

Digital signature of the message also provides **integrity** through the transfer.

Every transaction consists of:

- *input* (that shows transactions with which the sender received Bitcoins)
- *quantity* (the amount of Bitcoins that sender sends to the recipient)
- *output* (Bitcoin address of the recipient)

Bitcoin miners verify the transactions, put them into blocks of transactions and decide which block is the next one, i.e., *bitcoin system* groups the transactions into blocks and connects them in Blockchain. Since, many people can create blocks at the same time, which block will enter first in Blockchain is decided by using a hash function. The factor that determines the probability that the ”mining” computer will verify the block of transactions is known as *”difficulty” factor*.

Additionally, we will be describe some processes included in Bitcoin system.

Hashing in Bitcoin system

A hash function is a mathematical process that takes as input a string of arbitrary length, performs an operation on it, and returns output data of a fixed length.

Bitcoin uses both SHA-256 and RIPEMD-160 hashes. Usually, when a hash is used in Bitcoin system, it is computed twice. Most often SHA-256 hashes are used, while RIPEMD-160 is used for creating a shorter hash for a bitcoin address. On Table I we give an example of double hashing of string ”bitcoin”.

TABLE I
EXAMPLE OF DOUBLE HASHING OF STRING ”BITCOIN”

SHA256	
first round of SHA-256	**6b88c087247aa2f07ee1c5956b8e1a9f 4c7f892a70e324f1bb3d161e05ca107b**
second round of SHA-256	**a23b7f87e4250b3a64b737f349c06422 f752f419cbb25ae9169a6cf1e23f4462**
SHA-256 & RIPEMD-160	
first round with SHA-256	**6b88c087247aa2f07ee1c5956b8e1a9f 4c7f892a70e324f1bb3d161e05ca107b**
second round with RIPEMD-160	**b67f99610e811d5eba9e 337877a8f55f766d7401f**

Difficulty factor in Bitcoin system

The output of the hash function of a block has to be under specific value called **target**. **Difficulty** is a measure about how hard is to find a hash bellow this target. The value of the difficulty factor changes on every 2016 blocks. This will produce, on average, one block every ten minutes (a rate of the block). The current difficulty factor and **hash-rate** (the speed at which a computer is completing an operation in the Bitcoin code) determine the number of generated Bitcoins that one individual can achieve daily. We made analyses of how difficulty factor changes daily during August 2017 and annually in period from April 2017 to April 2018. In Fig. 2 and in Fig. 3 the results of those analyses are shown.

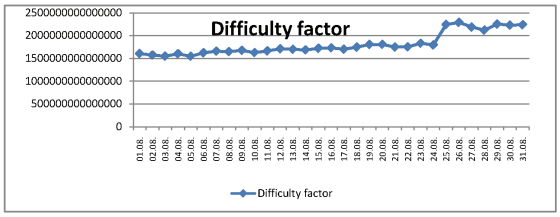


Fig. 2. Difficulty factor in August 2017

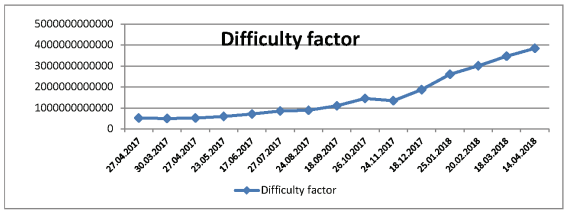


Fig. 3. Difficulty factor from April 2017 to April 2018

Countries where Bitcoin is legal

Bitcoin is not related to any government as it is an independant currency. Whether it is legal or not, it depends on the country [15].

- In **US** Bitcoin is used partially depending on the state. Every individual who mines Bitcoins is subject to taxation. Since November 2013, Bitcoin is regarded as a ”legal means of exchange”.

- In **Australia** it is legal to trade, mine and buy Bitcoins all over the country. Bitcoin is regarded as real money since 1 July 2017.
- Last year the **Japanese** goverment legalized Bitcoin as a method of payment. Now, the country is trying to protect the Bitcoin transactions.
- The Russian Deputy Finance Minister is thinking to legalize Bitcoin and other cryptocurrencies this year, although there is no offical report yet. **Russia** is also working on development of CryptoRuble - its own cryptocurrency.
- In **Europe**, each country has its own legal system when it comes to Bitcoin.
  - In **Belgium**: "The Minister of Finance indicated that government intervention with regard to the bitcoin system does not appear necessary at the present time" [17].
  - The use of Bitcoin in **UK** is not regulated and Bitcoin is treated as "private money" which has to be VAT taxed whwn trading.
  - **France** has certain regulations about Bitcoin use. Each exchange or Bitcoin wallet must contain user data, which is not a characteristic of Bitcoin.
  - Bitcoin in **Macedonia** is still illegal.
- **Argentina, Italy** and **the Netherlands** do not prohibit it, but they are skeptical towards Bitcoin. Argentinian government manifest sceptisism due to the impossibility to establish control over the new currency by the authorities, which is interpreted as a fertile soil for illegal transactions.

B. Smart contract

Exchanging a value, property, shares, or anything of value in a transparent, conflict-free ways, between two owners based on a set of conditions is included in an agreement called Smart contract, which is controlled by decentralized consent on a Blockchain. As a traditional contract, a Smart contract defines the rules and penalties around an agreement. Furthermore, the Smart contract automatically enforces those obligations [6].

A program code, a storage file and an account balance are parts of the Smart contract. A contract's storage file is stored on the Blockchain, while a contract's program logic is executed by the network of miners. Whenever a user or another contract sends a message the contract's code is executed. A contract can also receive and send money to other contracts or users into its account balance [5].

In Fig. 4 we give a simple model of a Smart contract.

In the following we will describe some applications of Smart Contracts.

- **Advantage of Blockchain technology in music industry**
  Blockchain can be also applied for a music distribution. Using this technology musicians and music companies that own music rights could receive appropriate funds anytime when their music is used for commercial purposes. Blockchain ledger can make a more direct connection between musicians and fans. A music with a unique ID and time stamp can be made public on the ledger. In this way, problems with downloading, coping and modifying of the digital content can be prevented. Every record on the Blockchain ledger has metadata with information for ownership and the rights, so everyone can see it and verify. The monetization of the music is made with Smart contract that enable payment transactions. In this way, consumers can choose the record and immediately pay to the owners using cryptocurrency [10].
  Benji Rogers' on-line music platform is one of the companies that have made a globally decentralized Blockchain ledger as a solution for the problems of ownership, payment and transparency. The records of this Blockchain ledger are "dotBlockChain" - a new dynamic file format (one type of zip file) that carries all the information for each record.

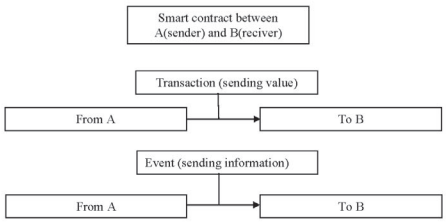


Fig. 4. A simple model for a Smart contract

- **Smart contracts in insurance policies**
  The process of requirments in insurance policies is still manuel with a big amount of human involvement and it can take a long time to be paid. Smart contracts can autiomate this process. The advantages of Smart contract are cost reduction, increased transparency and trust. The input conditions of the Smart contract are recorded onto the Blockchain. And for example, in case of natural disaster, the claim process is activated automatically without the need of human action [12].
  For example, this year Etherisc begins token sale for blockchain powered insurance.

- **How Blockchain enhances the supply chain**
  A supply chain is a system of organizations, people, activities, information, and resources involved in moving a product or service from supplier to customer. Supply chain activities involve the transformation of natural resources, raw materials, and components into a finished product that is delivered to the end customer [8].
  Everything that is included in creating and distributing goods is consisted in the supply chain, but it is very

difficult to trace back these events because of the wide span of the supply chains. There is a lack of transparency across supply chains which makes it hard for consumers to verify the real value of the products and they can not know the real costs. Also, the environmental damage from the production of goods is difficult to trace because of supply chains.
Blockchain repairs problems of the supply chain and in that way keeps record of the transfer of goods on the ledger as transactions. These transactions include price, date, location, quality and origin of the product, as well as the parties involved in the supply chain. The decentralized structure of the Blockchain provide transparency and security among all participants included in the supply chain of the products [9].
Sweetbridge is an example of a project that proposes a layered blockchain-based protocol stack in the service of a broad ecosystem of supply chain. Also, Provenance is another example of a traceability system for materials and products using blockchain technology.

- **How Blockchain could help the patent system**
  The patent system is an instrument that helps innovators to utilize their products and protect their rights over the products. Because the patent system has still some weaknesses, using Blockchain could help to correct some of its disadvantages [13]. This can be done with the following two characteristics of Blockchain:

  – *Hashing*. All hashes are unique and even a small change in the input will result in a different hash. Same hash is produced only by hashing the identical input.
  – *Proof of existence*. All hashes are recorded on the Blockchain by creating a record that the hash existed at a given time. Everyone can verify the record, but nobody can see its content. In order the patent owners to prove the existence of that document at a specific time, they have to hash the identical copy of it again. This process can help innovators to protect their work by storing a hash of their patent description on the Blockchain.

IV. Conclusion

In this paper we give a survey about Blockchain technology and its major innovations and applications. We describe the basic principles of Blockchain and how it works. Some known applications of Blockchain are considered. The major application is the Bitcoin the first decentralized peer-to-peer cryptocurrency. Additionally, we described some processes included in Bitcoin system. We made an analysis of how difficulty factor changes daily during August 2017. Also, we considered the second major application of Blockchain called Smart contract and its application in different areas. Our ongoing research is about the Blockchain scaling.

Acknowledgment

This research was partially supported by Faculty of Computer Science and Engineering at "Ss Cyril and Methodius" University in Skopje.

References

[1] http://www.iamwire.com/wp-content/uploads/2016/12/how-blockchain-works.png
[2] Nomura Research Institute, "Survey on blockchain technologies and related services", March 2016.
[3] M. Crosby, Nachiappan, P. Pattanayak, S. Verma,V. Kalyanaraman, "Blockchain technology: Beyond bitcoin", Applied Innovation Review, Berkeley, Issue No.2, June 2016.
[4] V. Gupta, " A brief history of blockchain", Harvard Business Review, Fevruary 28, 2017.
[5] K. Delmolino, M. Arnett, A.E. Kosba, A. Miller, E. Shi, "Step by step towards creating a safe Smart contract: lessons and insights from a cryptocurrency lab.", IACR Cryptol ePrint Arch 460, 2015.
[6] https://blockgeeks.com/guides/smart-contracts/
[7] P. Satyavolu, A. Sangamnerkar "Blockchains smart contracts: Driving the next wave of innovation across manufacturing value chains". https://www.cognizant.com/whitepapers/BlockChains-smart-contracts-driving-the-next-wave-of-innovation-across-manufacturing-value-chains-codex2113.pdf
[8] I. Kozlenkova, G.T.M. Hult, D.J. Lund, J.A. Mena, P. Kekec, "The role of marketing channels in supply chain management", Journal of Retailing, 91 (4), 2015, pp 586–609.
[9] https://techcrunch.com/2016/11/24/Blockchain-has-the-potential-to-revolutionize-the-supply-chain/
[10] https://techcrunch.com/2016/10/08/how-Blockchain-can-change-the-music-industry/
[11] https://gendal.me/2015/02/10/a-simple-model-for-smart-contracts/
[12] https://www.draglet.com/Blockchain-applications/smart-contracts/use-cases
[13] P. Boucher, S. Figueiredo Do Nascimento, M. Kritikos, "How Blockchain technology could change our lives", European Parliament. PE 581.948, 2017.
[14] https://github.com/capiman/sha256-sat-bitcoin
[15] https://atozforex.com/news/top-countries-where-bitcoin-is-legal/
[16] https://www.reuters.com/article/us-jpmorgan-cyber-bitcoin-idUSKCN11P2DE
[17] http://www.loc.gov/law/help/bitcoin-survey/
[18] https://cointelegraph.com/news/russian-minister-we-will-never-consider-bitcoin-legal
[19] https://www.buybitcoinworldwide.com/confirmations/
[20] http://www.altcointoday.com/bitcoin-ethereum-vs-visa-paypal-transactions-per-second/
[21] M. Rosenfeld, "Analysis of hashrate-based double-spending", arXiv:1402.2009.
[22] https://en.bitcoin.it/wiki/Confirmation
[23] https://people.xiph.org/ greg/attack-success.html
[24] S. Nakamoto, "Bitcoin: A peer-to-peer electronic cash system", 2008, http://bitcoin.org/bitcoin.pdf.

Using QR codes for easier and more secured business communication

Katerina Bashova
School of Computer Science and Information Technology
University American College Skopje
Skopje, Macedonia
bashova.katerina@gmail.com

Veno Pachovski
School of Computer Science and Information Technology
University American College Skopje
Skopje, Macedonia
pachovski@uacs.edu.mk

Adrijan Bozhinovski
School of Computer Science and Information Technology
University American College Skopje
Skopje, Macedonia
bozinovski@uacs.edu.mk

***Abstract*—The objective of this paper is to find means for an alternative document transfer, in a way that will be compact, simple, light while being readable (understandable) only by the receiver. As a possible solution that will meet the latter requirements, the Quick Response (QR) code is chosen. The QR code can be easily generated and has the potential to store vast amount of data. Another advantage of the QR code is that it is readable by portable devices, like smart phones or tablet, which already has installed some kind of QR code reader. Therefore, we will explore some strategies that will enable encoding the document using the lowest possible QR code version. Additionally several types of encryption will be explored, to find the most suitable one, so that the document is encrypted before being encoded (as a QR code).**

Keywords—QR code, data compression, data encoding, encryption

I. INTRODUCTION

In this modern digital era, it is common for any type of document (financial reports, invoices and other) to be in some universal digital format that will be able to withhold within vast amount of data, while at the same time being easy for transfer. The most common used file formats are the following: .doc and .docx - Microsoft Word file, .odt – Open Office Writer document file, .pdf - PDF file, .rtf - Rich Text Format, .tex - A LaTeX document file, .txt - Plain text file, .wks and .wps- Microsoft Works file, .wpd - WordPerfect document, .csv - Comma separated value file, .dat - Data file, .xml - XML file.

But the question is whether the data that was stored in some of those file formats, could be stored in some alternative way that will be compact, simple, light while being readable (understandable) only by the receiver. The possible solution is the QR code, which has gained popularity and usage in everyday life since 1994. The QR code is vastly used in:

- Manufacturing, for product traceability, process control, order and time tracking, inventory and equipment management;
- Warehousing and logistics, for item tracking;
- Retailing, for point-of-purchase product identification, sales management, inventory control;
- Health care, for medical records management, patient identification, medication tracking, equipment and device tracking;
- Life sciences, for specimen tracking;
- Transportation, for fleet management, ticketing and boarding passes;
- Office automation, for document management;
- Marketing and advertising, for mobile marketing, electronic tickets, coupons, payments and loyalty programs; [2]
- Graveyard use; [3][4]
- Visa Application Documents; [5]
- and other.

A. *What is QR code?*

The QR (Quick Response) Code is a two-dimensional (2-D) matrix code, that unlike the standard one-dimensional (1-D) barcode that only contains data horizontally, the QR code contains data both vertically and horizontally (see Fig. 1). By being able to store the data in both directions, it can be printed out in smaller size. Additionally it is dirt and damage resistant, meaning it has error correction capability. Data can be restored even if the symbol is partially dirty or damaged. Another feature of the QR code is that it can be read from any angle (omni-directional) thus providing high speed reading. This is possible due to the position detection patterns that are located at the three corners of the symbol (see Fig. 2) [2][6].

Fig. 1. Comparison between QR code and bar code [2].

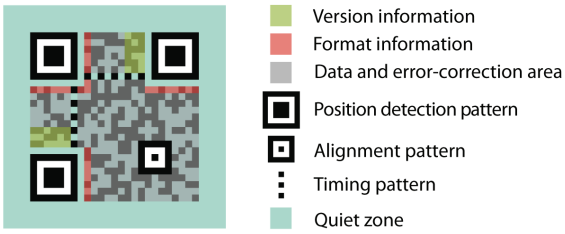


Fig. 2. Structure of QR code [2].

B. How it works and versions

There are various types of QR codes such as QR code Model 1 (versions 1-14), QR code Model 2 (the most commonly used, versions 1-40), Micro QR code (versions M1-M4), iQR code (version 1-61), SQRC and FrameQR [7]. For the purpose of this paper, we chose the most commonly used type the QR code Model 2.

As we previously mentioned, there are 40 versions of QR code and every version has a different module configuration. A module is the number of black and white dots that make up the QR code. "Module configuration" refers to the number of modules contained in a symbol. The smallest version 1 contains 21 by 21 modules, while the highest version 40 177 by 177 modules. On every next version number comprises four additional modules per side [8]. In table I, details of each versions modules and data capacity combined with the four error correction levels are illustrated.

TABLE I. QR CODE VERSIONS AND INFORMATION CAPACITY [8]

Ver.	Modules	ECC Level	Data bits (mixed)	Numeric	Alphanumeric	Binary	Kanji
1	21x21	L	152	41	25	17	10
		M	128	34	20	14	8
		Q	104	27	16	11	7
		H	72	17	10	7	4
...	...	...	...	..	...	...	...
40	177x177	L	23,648	7,089	4,296	2,953	1,817
		M	18,672	5,596	3,391	2,331	1,435
		Q	13,328	3,993	2,420	1,663	1,024
		H	10,208	3,057	1,852	1,273	784

As mentioned before, the QR code has error correction capability to restore data if the code is dirty or damaged. From table II, we can see that there are four error correction levels available. When using higher error correction level, the correction capability is increased as well as the amount of data QR code size. The most recommended error correction level for general usage is M, while the Q or H levels are usually used in factory environments where the chances are high for the QR code to be damaged or get dirty. The lowest level L is used for clean environment with the large amount of data [9].

TABLE II. QR CODE ERROR CORRECTION CAPABILITY [9][10]

ECC Level	Percentage of correction
L (Low)	Aproxx 7%
M (Medium)	Aproxx 15%
Q (Quartile)	Aproxx 25%
H (High)	Aproxx 30%

II. DESCRIPTION OF THE SCENARIOS

For the purpose of this research, three scenarios were defined. In the first scenario, we read the document as byte array and encode it into QR code. The reversed operation would be decoding into byte array and save it as the original document (with the same file extension) (see Fig. 3). The sample from this scenario is the referent sample upon which we will compare the rest of the generated samples from the other scenarios.

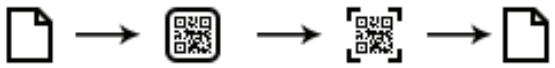


Fig. 3. Scenario 01 Encoding and decoding.

In the second scenario, we read the document as byte array, compress it and encode it into QR code. The reversed operation would be decoding into byte array, decompress it and save it as the original document (with the same file extension) (see Fig. 4).

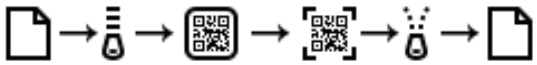


Fig. 4. Scenario 02 Compression, encoding, decompression and decoding.

In the third scenario, we read the document as byte array, encrypt it and encode it into QR code. The reversed operation would be decoding into byte array, decrypt it and save it as the original document (with the same file extension) (see Fig. 5).

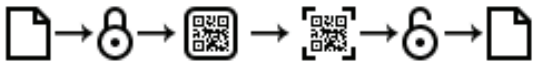


Fig. 5. Scenario 03 Encryption, encoding, decryption and decoding.

After each QR code creation as well as reading, the following data will be recorded and used for further analysis:

- Title, which indicates which scenario was used for data recording,
- Name of the file,
- Extension of the file,
- Number of white spaces (as char) (if the extension is .txt),
- Number of controls (if the extension is .txt),
- Number of characters (without white space chars) (if the extension is .txt),
- Number of characters (with white space chars) (if the extension is .txt),
- Total number of characters (if the extension is .txt),
- Size (in bytes),
- MD5 checksum.

For all three encoding scenarios, we will use QR code version 40 with error correction level M (approx 15%). Several samples were tested and for all those that were bigger than 1305 bytes, ZXing throws exception "Data too big for requested version". In order to avoid this exception, a sample with size 1293 bytes and with encoding UFT-8 was chosen. For this research, we developed prototype desktop application using Microsoft Visual Studio Community 2015. Additional library that was used is the open source ZXing.Net [11][12], that was installed using the NuGet Package Manager. For the encryption and decryption System.Security.Cryptography namespace and TripleDESCryptoServiceProvider class were used. The length of the used password was four characters. For the compression and decompression, System.IO.Compression namespace and GZipStream class were used.

Each scenario is isolated in separated forms and three different classes with methods that are used in one or more scenarios. Each form has a tab control with two tab pages, one for encoding and one for decoding. The encoding page has two combo boxes, the first is for choosing ECL (error correction level, which is by default set to M) and the second is for choosing QR version (which is default set to version 40). Afterwards there are two buttons, three text boxes, two labels and one picture box. The encoding tab pages in the first and in the second scenarios are identical, whereas the encoding page tab in the third scenario has and additional text box for password. The sample that was chosen is a real invoice that contains personal information and therefore in all figures from Fig. 6 to Fig. 11 the data from the invoice (before encoding and after decoding) is blurred. The encoding tab pages for each scenario can be seen from Fig. 6 to Fig. 8.

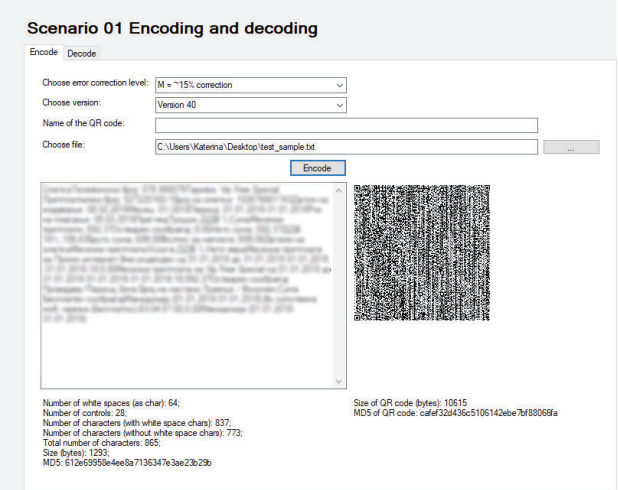


Fig. 6. Tab page of encoding in scenario 01.

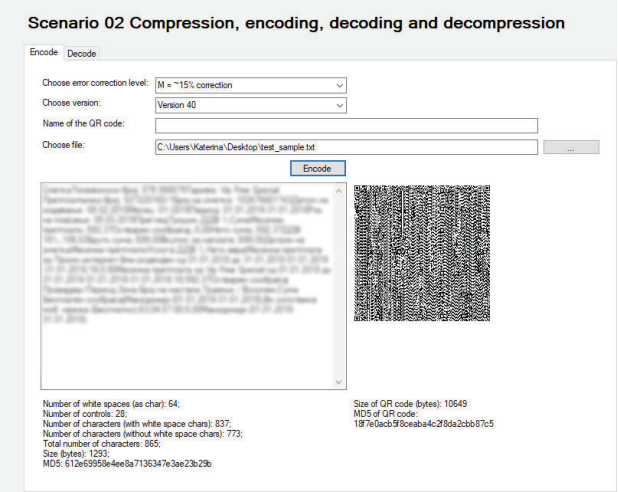


Fig. 7. Tab page of compressing and encoding in scenario 02.

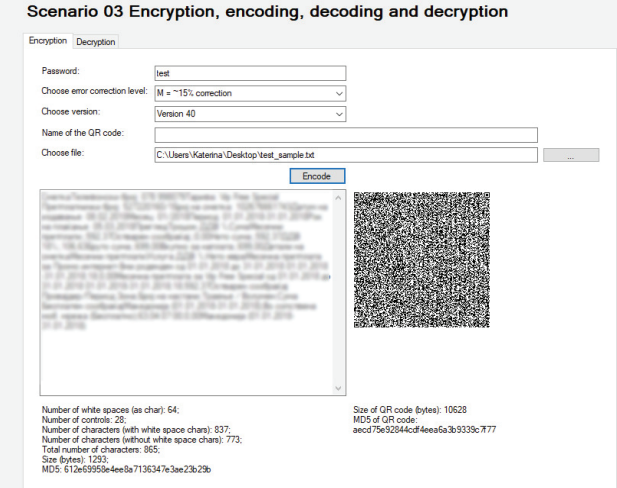


Fig. 8. Tab page of encryption and encoding in scenario 03.

The decoding page has two text boxes, two buttons, two labels and one picture box. Again the decoding tab pages in the first and in the second scenarios are identical, whereas the decoding page tab in the third scenario has and additional text box for password. The decoding tab pages for each scenarios can be seen from Fig. 9 to Fig. 11.

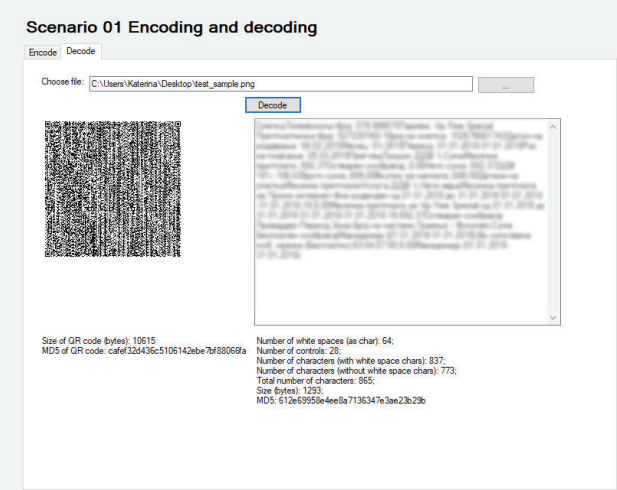


Fig. 9. Tab page of decoding in scenario 01.

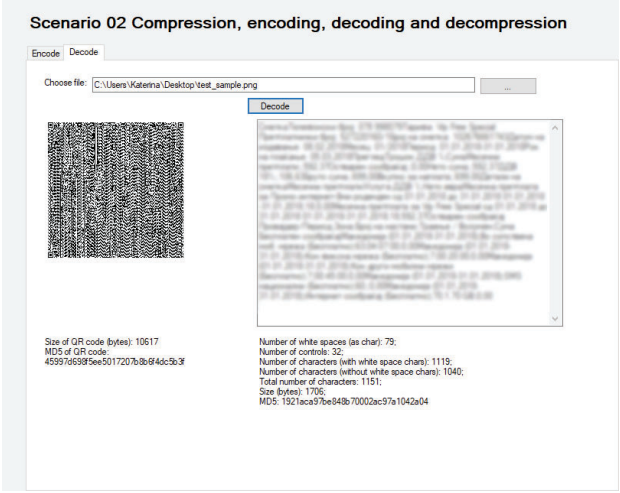


Fig. 10. Tab page of decompression and decoding in scenario 02.

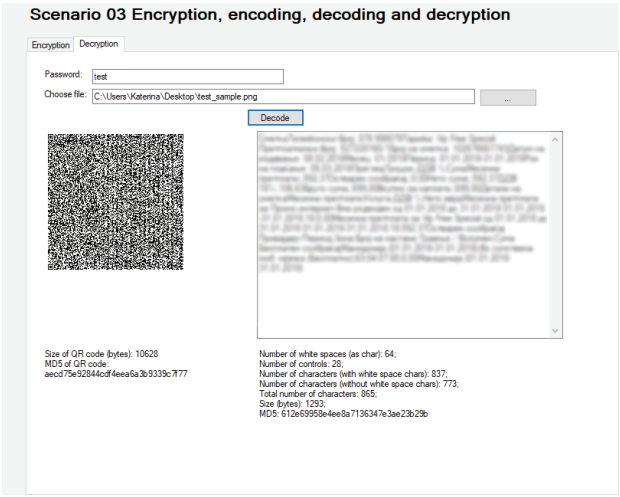


Fig. 11. Tab page of decryption and decoding in scenario 03.

III. Analysis

The results from all three scenarios are shown in the tables III, IV, V and VI. With (1) the percentage difference was calculated between the reference sample and each of the QR codes, as well as between each QR code. In the former case, variable a is the size of the sample and variable b is the size of the selected QR code (encoded, compressed-encoded or encrypted-encoded). In the latter case, variable a and variable b are the sizes of two selected QR codes.

$$\%\text{diff} = \left| \frac{a-b}{\frac{a+b}{2}} \right| \times 100\% \qquad (1)$$

The size difference between the referent sample and the QR code from the first scenario is 9322 bytes (~156,57%). The size difference between the referent sample and the QR code from the second scenario is 9356 bytes (~156,69%). The size

difference between the referent sample and the QR code from the third scenario is 9335 bytes (~156,61%).

The size difference between the QR code from the first scenario and the QR code from the second scenario is 34 bytes (~0,32%). The size difference between the QR code from the first scenario and the QR code from the third scenario is 13 bytes (~0,12%). The size difference between the QR code from the second scenario and the QR code from the third scenario is 21 bytes (~0,20%).

From the tables III to IV as well from the graphic in Fig. 12, we can see that the QR code from the first scenario that was only encoded has the smallest size in bytes, while the highest size in bytes has the QR code from the second scenario that was only compressed and encoded.

TABLE III. RESULTS FFROM SCENARIO 01

	Scenario 01 Encoding and decoding		
	Referent	*Encoded*	*Decoded*
Extension	.txt	.png	.txt
Size	1293 bytes	10615 bytes	1293 bytes
MD5	612e69958e4ee8a7136347e3ae23b29b	cafef32d436c5106142ebe7bf88066fa	612e69958e4ee8a7136347e3ae23b29b

TABLE IV. RESULTS FROM SCENARIO 02

	Scenario 02 Compression, encoding and decoding, decompression		
	Referent	*Encoded*	*Decoded*
Extension	.txt	.png	.txt
Size	1293 bytes	10649 bytes	1293 bytes
MD5	612e69958e4ee8a7136347e3ae23b29b	18f7e0acb5f8ceaba4c2f8da2cbb87c5	612e69958e4ee8a7136347e3ae23b29b

TABLE V. RESULTS FROM SCENARIO 03

	Scenario 03 Encryption, encoding and decoding, decryption		
	Referent	*Encoded*	*Decoded*
Extension	.txt	.png	.txt
Size	1293 bytes	10628 bytes	1293 bytes
MD5	612e69958e4ee8a7136347e3ae23b29b	aecd75e92844cdf4eea6a3b9339c7f77	612e69958e4ee8a7136347e3ae23b29b

TABLE VI. SIZE DIFFERENCES

	Ref.	*1*	*2*	*3*
Ref.		9322 (~156,57%)	9356 (~156,69%)	9335 (~156,61%)
1	9322 (~156,57%)		34 (~0,32%)	13 (~0,12%)
2	9356 (~156,69%)	34 (~0,32%)		21 (~0,20%)
3	9335 (~156,61%)	13 (~0,12%)	21 (~0,20%)	

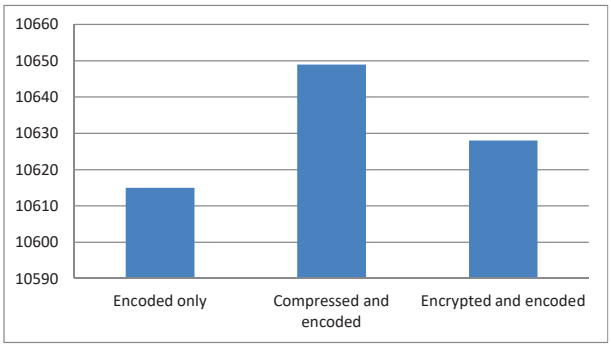


Fig. 12. Size in bytes.

IV. CONCLUSION

The primary hypothesis was that the document that is compressed and encoded as a QR code would have the smallest size in bytes. However, surprisingly from the previously presented results, it is quite the opposite; the QR code from scenario 01 (the encoded only one) has the smallest size of all three QR codes, while the the QR code from scenario 02 (the encrypted one) is in the middle. The possible causes for this kind of results could be: image file format in which the QR code is saved and the GZip class methods may generate and add additional code to the compressed byte array. Therefore, in the further research we will save the QR code in different image file formats like for example .bmp, .tiff, .jpg and .gif. Also we will be looking for other classes that provide compression and then chose which one has better compressing capacity.

Regarding scenario 03, which implemented the DES algorithm, in further research we will implement other algorithms in order to test if any type of encryption would generate QR code with smaller size.

This research shows that it is possible to compress or encrypt a document and then to encode it into a QR code. The limitations of these methods are that maximum version 40 must be selected and the document must be in the size boundaries recommended by the version combined with the error-correction level.

REFERENCES

[1] Computer Hope (2018). What are the most common file types and file extensions? Available: https://www.computerhope.com/issues/ch001789.html [accessed: 11.03.2018])

[2] DENSO ADC (2011) . QR Code® Essentials. Archived from the original on 12.05.2013. Avaiable: http://www.nacs.org/LinkClick.aspx?fileticket=D1FpVAvvJuo%3D&tabid=1426&mid=4802 [accessed: 11.03.2018]

[3] Noval, Asami (2008). Japanese gravestones memorialize the dead with QR codes. Wired. Avaiable: https://www.wired.com/2008/03/japanese-graves/ [accessed: 15.03.2018]

[4] BBC News (2012). QR codes on headstones offered by Poole undertakers. Avaiable: http://www.bbc.com/news/uk-england-dorset-19506286 [accessed: 11.03.2018]

[5] Ministry of Foreign Affairs of Japan (2018). VISA / Residing in Japan. Avaiable: http://www.mofa.go.jp/j_info/visit/visa/ [accessed: 11.03. 2018]

[6] DENSO WAVE INCORPORATED (2013).What is a QR Code? Avaiable: http://www.qrcode.com/en/about/ [accessed: 11.03. 2018]

[7] DENSO WAVE INCORPORATED (2015). Types of QR Code Avaiable: http://www.qrcode.com/en/codes/ [accessed: 11.03. 2018]

[8] DENSO WAVE INCORPORATED (2013). Information capacity and versions of QR Code. Avaiable: http://www.qrcode.com/en/about/version.html [accessed: 11.03. 2018]

[9] DENSO WAVE INCORPORATED (2013). Error correction feature. Avaiable: http://www.qrcode.com/en/about/error_correction.html [accessed: 11.03. 2018]

[10] TEC-IT Datenverarbeitung GmbH (2018). QR Code Barcode Generator Software: 2D Strichcode Generator. Archived from the original on 15.09.2012. Avaiable: https://archive.is/20120915040144/http://www.tec-it.com/de/support/knowbase/symbologies/qrcode/Default.aspx [accessed: 15.03.2018]

[11] Michael Jahn (2017). ZXing.Net. GitHub. Avaiable: https://github.com/micjahn/ZXing.Net/ [accessed: 16.03.2018]

[12] Michael Jahn (2017). ZXing.Net 0.16.2. NuGet. Avaiable: https://www.nuget.org/packages/ZXing.Net/ [accessed: 16.03.2018]

Enumeration of the Closed Knight's Paths of Length 10

Stoyan Kapralov
Dept of Math & Computer Science
Technical University of Gabrovo
Gabrovo, Bulgaria
s.kapralov@gmail.com

Valentin Bakoev
Dept af Algebra & Geometry
Veliko Tarnovo University
Veliko Tarnovo, Bulgaria
v_bakoev@yahoo.com

Kaloyan Kapralov
Skyscanner Bulgaria
Sofia, Bulgaria
kaloyan.kapralov@gmail.com

***Abstract*—The classification of closed knight's paths of length 6 and 8 has been obtained in a previous article by the same authors. In this paper we present the results for enumeration of all geometrically distinct closed knight's paths of length 10.**

Keywords—computational combinatorics, knight graph, closed knight's path, enumeration up to equivalence

I. INTRODUCTION

In graph theory, *a knight graph* is a graph that represents all legal moves of the knight chess piece on a chessboard. Each vertex of this graph represents a square of the chessboard, and each edge connects two squares that are a knight's move apart from each other. More specifically, an *m* x *n* knight graph is a knight graph of an *m* x *n* chessboard [1]. An example of a knight graph is shown in Fig. 1.

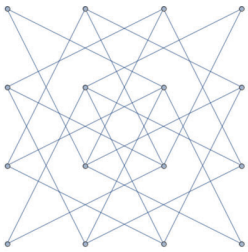


Fig. 1. The 4 x 4 knight graph.

A Hamiltonian cycle on the knight graph is called a knight's tour. The first scientist who studied the problem of finding a knight's tour in the *m* x *n* knight graph was the great Leonard Euler [2].

An excellent survey of information about paths in knight graphs is collected and maintained by G. Jelliss [3].

In this paper we consider cycles of smaller length than the knight's tour, and we call them *closed knight's paths*.

We will number the vertices of the *m* x *n* knight graph with integers from 1 to *mn*, row-wise, as in Fig. 2.

1	2	3	4
5	6	7	8
9	10	11	12
13	14	15	16

Fig. 2. The numeration of the cells.

Consider the following two sequences of vertices:

1 7 14 5 11 4 6 15 8 10,

2 8 10 3 5 14 12 6 15 9.

These are examples of closed knight's paths of length 10 in the 4 x 4 knight graph, illustrated in Fig. 3a and Fig. 3b, respectively.

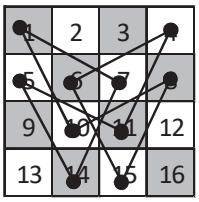


Fig. 3a. The 1 7 14 5 11 4 6 15 8 10 path.

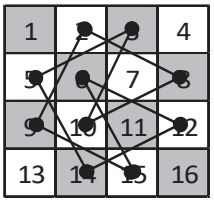


Fig. 3b. The 2 8 10 3 5 14 12 6 15 9 path.

In an *m* x *n* knight graph, we consider only the closed paths that could not be embedded in a smaller knight graph. For example, we consider the sequence: 1 7 9 2 11 4 6 12 3 10 as a closed knight's path of length 10 in the 3 x 4 knight graph but not as a closed knight's path of length 10 in the 4 x 4 knight graph, see Fig. 4.

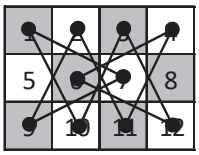


Fig. 4. The 1 7 9 2 11 4 6 12 3 10 path.

Assume that the board cells are squares. Each closed knight's path of length k can be seen as a k-sided polygon with vertices in the centers of the corresponding cells. In the general case these polygons are self-intersecting. An enumeration of some non-intersecting knight's tours with longest possible length is presented in [4].

Definition. Two closed knight's paths of the same length are equivalent if the corresponding polygons are congruent.

We say that two non-equivalent closed knight's paths (in the sense of the above Definition) are *geometrically distinct*.

The authors of the current paper have enumerated the closed knight's paths of length 6 and length 8 in [5]. The results stated in [5] are that there are exactly 25 such paths of length 6 and 478 paths of length 8. However, in [5], the equivalence relation is defined slightly differently, counting non-congruent k-sided polygons, as equivalent, because they are comprised of the same *set of k traversed* vertices. For instance, the following three paths in the 5 x 6 knight graph are considered equivalent:

3 7 15 11 24 28 20 16, see Fig. 5a,

3 7 15 28 20 16 24 11, see Fig. 5b,

3 7 20 28 15 11 24 16, see Fig. 5c,

because the *set of traversed* vertices is identical for all three as illustrated in Fig. 6.

According to the present definition of equivalence these three paths are considered as non-equivalent, because the corresponding 8-polygons are non-congruent.

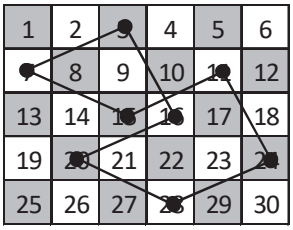


Fig. 5a. The 3 7 15 11 24 28 20 16 path.

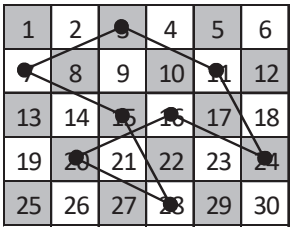


Fig. 5b. The 3 7 15 28 20 16 24 11 path.

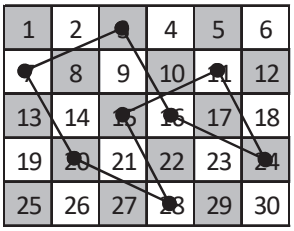


Fig. 5c. The 3 7 20 28 15 11 24 16 path.

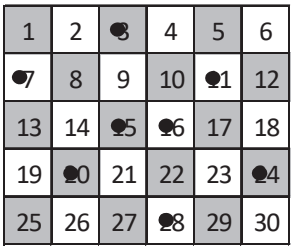


Fig. 6. A set of vertices that could be traversed in multiple non-equivalent ways.

In essence, the results of [5] could be restated as follows: the number of *geometrically distinct* closed knight's paths of length 6 is 25, and the number of *geometrically distinct* closed knight's paths of length 8 is 480 (as opposed to 478).

II. Closed Knight's Paths of Length 10

In this section we study the closed knight's paths of length 10.

It is easy to see that a necessary condition for the existence of closed knight's paths of length $k > 2$ in an m x n knight graph is $3 \le m, n \le k+1$.

We solve separately the subtasks of generating all non-equivalent closed knight's paths of length 10 in an m x n knight graph for $3 \le m \le n \le 11$.

In each class of equivalent paths there is one path which is lexicographically smaller than any other path in the set. We call this path *minimal*.

We implement a backtracking algorithm which generates in lexicographical order all paths one by one, checks if the current path is minimal and discards it in case it is not minimal.

To check if a given path P is minimal we consider all paths that can be obtained from P by a sequence of one or more of the following transformations:

- reflection
- rotation (in case $m = n$).

If we obtain a path that is smaller (in lexicographic order) than P, this means that the path P is not minimal.

The results are summarized in Table I and the computations show that the following theorem holds.

Theorem. There are exactly 12000 geometrically distinct closed knight's paths of length 10.

The computer programs used for obtaining the results about enumeration of closed knight's path are written in Julia programming language [6]. The programs and files with all non-equivalent closed knight's paths of length k for $k = 4, 6, 8, 10$ are published in [7].

TABLE I. THE RESULTS

m\n	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**
3	0	1	3	9	15	15	14	8	0
4		9	110	218	331	255	186	43	1
5			241	1045	1122	945	455	81	2
6				864	1839	1169	506	57	2
7					783	954	284	27	0
8						200	106	0	0
9							0	0	0
10								0	0
11									0

ACKNOWLEDGEMENT

The authors are grateful to anonymous reviewers for the helpful comments and suggestions.

The work of the first author was supported in part by Grant 1808C/2018 of the Technical University of Gabrovo, Bulgaria.

REFERENCES

[1] http://mathworld.wolfram.com/KnightGraph.html, Accessed 03/12/2018.

[2] L. Euler, Solution d'une question curieuse qui ne paroit soumise à aucune analyse (Solution of a curious question which does not seem to have been subject to any analysis), Mémoires de l'Academie Royale des Sciences et Belles Lettres, Année 1759, vol.15, pp.310–337, Berlin 1766.

[3] G. Jelliss, Knight's Tours Notes, Available at: http://www.mayhematics.com/t/t.htm, Accessed 03/12/2018.

[4] http://euler.free.fr/knight/index.htm, http://euler.free.fr/knight/index2.html, Accessed 03/12/2018.

[5] S. Kapralov, V. Bakoev and K. Kapralov, "Enumeration of some closed knight paths," Proc. of the International Scientific Conference UNITECH-2017, Nov. 17-18, 2017, Technical University of Gabrovo, Gabrovo, Bulgaria; arXiv:1711.06792 [math.CO]

[6] https://julialang.org/, Accessed 04/19/2018.

[7] S. Kapralov, V. Bakoev and K. Kapralov, https://drive.google.com/drive/folders/16jHC5bnG6TGDsttc_01j1skDPKHXepdD?usp=sharing, Accessed 04/19/2018.

Irreducible Polynomials in the Construction of Uniformly Distributed Sequences

Vesna Dimitrievska Ristovska
Faculty of Computer Sciences and Engineering
Univ. Cyril and Methodius, Skopje, Macedonia
vesna.dimitrievska.ristovska@finki.ukim.mk

Vasil Grozdanov
Faculty of Natural Sciences and Mathematics
Department of Mathematics
Univ. Neophit Rilski, Blagoevgrad, Bulgaria
vassgrozdanov@yahoo.com

***Abstract*—In this paper we present some numerical and graphical results based on irreducible polynomials with the goal -to generate some classes of uniformly distributed sequences. In the computations, we use our algorithm, proposed in a previous paper.**
Irreducible polynomials are used as a tool in the algorithm to get linear homogeneous recurrence relations with constant coefficients in the finite field in base b, where b is an arbitrary prime.
In this paper we present some results in the cases when the dimension $s-$ is set to be 1, 2 and 3. Obtained experimental results are graphically presented.

Keywords: irreducible polynomial, primitive polynomial, linear homogeneous recurrence relations, (t,m,s)-sequence.

I. Motivation

This paper is based on two ideas:

1) to generalize the constructive approach of Sobol' [7], which generate classes of (t,s)-sequences over the field $\mathbf{F_2}$ by using monocyclic difference operators over $\mathbf{F}_2$, in the way to use **arbitrary** base b, not only **base b=2** and
2) to use our previous paper [1] where we obtained an effective algorithm to generate monocyclic difference operators are over $\mathbf{F}_b$, where b is a prime number.

The main our goal in this paper is to make computations and to check how "good" are the irreducible polynomials as a tool in the construction of a class of $(t,s)-$sequences, by using linear homogeneous recurrence relations with constant coefficients in the finite field in base b, where b is an arbitrary prime.

II. Introduction

First, we will give the concept of a class of sequences with well uniformly distributed points in $[0,1)^s$, following Niederreiter [6]. Let $b \geq 2$ be a fixed integer and b will denote the base in which are constructed the considered sequences. Next we will give the definition of a $(t,m,s)-$net and a $(t,s)-$sequence.

Definition 1 *Let $0 \leq t \leq m$ be integers. A point set P consisting of b^m points in $[0,1)^s$ forms a $(t,m,s)-$net in base b, if every subinterval $J = \prod_{j=1}^{s} \left[\frac{a_j}{b^{d_j}}, \frac{a_j+1}{b^{d_j}}\right)$ of $[0,1)^s$, with integers $d_j \geq 0$ and integers $0 \leq a_j < b^{d_j}$ for $1 \leq j \leq s$ and of volume b^{t-m}, contains exactly b^t points of P.*

Definition 2 *Let $t \geq 0$ be a given integer. The sequence $(\mathbf{x}_n)_{n\geq 0}$, $\mathbf{x}_n \in [0,1)^s$, is a $(t,s)-$sequence in base b if for all $l \geq 0$ and $m \geq t$ the point set $\{\mathbf{x}_{lb^m}, \ldots, \mathbf{x}_{(l+1)b^m-1}\}$ is a $(t,m,s)-$net.*

It is obvious that a $(t,m,s)-$net is extremely well distributed if the parameter t is small.

III. Mathematical Preliminaries: Linear Homogeneous Recurrence Relations and Irreducible Polynomials Over $\mathbf{F}_b$

In the following we will give some notations and statements about the linear homogeneous recurrence relations and irreducible and primitive polynomials over the field $\mathbf{F}_b$.

A relation of the form

$$Lu_i = u_{i+m} + a_{m-1}u_{i+m-1} + \ldots + a_1u_{i+1} + a_0u_i, \quad (1)$$

where for $i \in \mathbf{Z}$, $u_i \in \mathbf{F}_b$, for $0 \leq i \leq m-1$ $a_i \in \mathbf{F}_b$, $a_0 \neq 0$ is called linear homogeneous recurrence relation (difference operators) over $\mathbf{F}_b$ of order m with constant coefficients (LHRR).

A solution of the equation

$$Lu_i = 0 \quad (2)$$

we will call the sequence $\ldots u_{-2}, u_{-1}, u_0, u_1, u_2, \ldots$, of elements from $\mathbf{F}_b$ which satisfied this equation for all integer i. The solution will be denoted with $\{u_i\}$. The solutions of the equation $Lu_i = 0$ have cyclic character. It is easy to show that the solution $\{u_i\}$ is periodical with a period ω_1 such that $\omega_1 \leq b^m - 1$.

A cycle of the solutions of the equation (2) we will call a set of solutions which are different by the shifting of the numeration, i.e. if $\{u'_i\}$ is a solution of the equation (2), then an arbitrary another solution $\{u_i\}$ of the cycle can be presented of the form $u_i = u'_{i+\alpha}$ for some integer α.

The relation Lu_i is called monocyclic if the equation $Lu_i = 0$ has only one solution with period $\omega = b^m - 1$.

To every LHRR (1) corresponds a polynomial of degree m over $\mathbf{F}_b$ of the form

$$Lu_i \leftrightarrow P(x) = x^m + a_{m-1}x^{m-1} + \ldots + a_1x + a_0, \quad (3)$$

where for $i = 0, 1, \ldots m-1$ $a_i \in \mathbf{F}_b$.

Following Lidl and Niederreiter [5] we will give the next definition:

Definition 3 *An irreducible polynomial is a non-constant polynomial that may not be factored into the product of two non-constant polynomials.*

Following Sobol' [7] we will recall the theoretical bases of the construction of sequences of $b-$adic rational type, the so-called $BR-$ sequences. So, the details are as follows:

Definition 4 *Let $V_1, V_2, \ldots, V_j, \ldots,$ be an arbitrary sequence of $b-$adic rational numbers, where for $j \geq 1$ we have that $0 < V_s < 1$. The numbers of this sequence we will call direction numbers. A $BR-$ sequence $\{r(i)\}_{i\geq 0}$ which corresponds to the direction numbers $\{V_j\}_{j\geq 1}$ is defined as: if an arbitrary integer number i has the $b-$adic presentation*

$$i = e_m e_{m-1} \ldots e_2 e_1,$$

then we replace

$$r(i) = e_1 V_1 * e_2 V_2 * \ldots * e_m V_m,$$

*where * is the operation digit-by-digit summation modulo b and $e_j V_j = \underbrace{V_j * \ldots * V_j}_{e_j - times}$.*

We can represent the direction numbers V_j in the form of $b-$ adic fractions:

$$V_j = 0, v_{j1} v_{j2} \ldots v_{ji} \ldots,$$

where all $v_{ji} \in \mathbf{F}_b$. In this sense the setting of the sequence $\{V_s\}$ is equivalent to setting an infinite matrix, which elements are from the field $\mathbf{F}_b$.

$$(v_{ji}) = \begin{bmatrix} v_{11} & v_{12} & \ldots & v_{1i} & \ldots \\ v_{21} & v_{22} & \ldots & v_{2i} & \ldots \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ v_{j1} & v_{j2} & \ldots & v_{ji} & \ldots \\ \ldots & \ldots & \ldots & \ldots & \ldots \end{bmatrix}. \quad (4)$$

This matrix is called direction matrix. Using results in our paper [1], the following two theorem hold:

Theorem 1 *Let in the direction matrix (4) we have that for $j = 1, 2, \ldots$ $v_{jj} \neq 0$, i. e. $v_{jj} \in \{1, 2, \ldots, b-1\}$, and for $i > j$ $v_{ji} = 0$. Then the corresponding $BR-$ sequence $\{r(i)\}_{i\geq 0}$ is a $(0, 1)-$ sequence in base b.*

Let Lu_i be an arbitrary monocyclic LHRR over the field $\mathbf{F}_b$ of order m. We will generate the direction numbers $V_1, V_2, \ldots, V_i, \ldots$ as a solution of the equality

$$V_{i+m} * a_{m-1} V_{i+m-1} * \ldots * a_1 V_{i+1} * a_0 V_i = b^{-m} V_i, \quad (5)$$

i.e. $LV_i = b^{-m} V_i$, as in the relation Lu_i should the symbol + to be replaced by the operation *. The initial conditions $V_1, V_2, \ldots, V_m$ of the equation (5) can be chosen in different manners, but for our purposes it is necessary to satisfy the next conditions: if the $b-$adic presentation of the number V_j is

$$V_j = 0, v_{j1} v_{j2} \ldots v_{ji} \ldots,$$

we assume that for all $j \geq 1$ $v_{jj} \neq 0$ and for $i > j$ $v_{ji} = 0$. In this way the matrix

$$B = \begin{bmatrix} v_{11} & 0 & \ldots & 0 \\ v_{21} & v_{22} & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots \\ v_{m1} & v_{m2} & \ldots & v_{mm} \end{bmatrix}$$

is triangular and nonsingular. On the main diagonal there are nonzero numbers and over it stand zeros. We will say that the $BR-$ sequence $\{r(i)\}_{i\geq 0}$ with these direction numbers $\{V_i\}_{i\geq 1}$ corresponds to the relation Lu_i.

In the next theorem [1] it is shown the possibility to use monocyclic LHRR as a tool to construct $(t, s)-$sequences.

Theorem 2 *Let b be a prime number and $L_1, L_2, \ldots, L_s$ are different monocyclic LHRR of orders $m_1, m_2, \ldots, m_s$ over $\mathbf{F}_b$. For $k = 1, 2, \ldots, s$ let $(P^{(k)}(i))_{i\geq 0}$ be the $BR-$ sequence which corresponds to the relation L_k. Then the sequence $(P(i))_{i\geq 0}$ of points of the form*

$$P(i) = (P^{(1)}(i), P^{(2)}(i), \ldots, P^{(s)}(i))$$

is a $(t, s)-$ sequence in base b with a parameter

$$t = \sum_{k=1}^{s} (m_k - 1). \quad (6)$$

It is well known that of a monocyclic relation of the form (1) corresponds an irreducible polynomial. The question was: how good uniformly distributed sequences we will obtained if we use irreducible polynomials?

Zierler [9] proved that necessary and sufficient condition that the relation (1) to be monocyclic is the polynomial (3) to be primitive. A polynomial is primitive if it is irreducible, a divisor of the binomial $x^\omega - 1$ (ω is the period of the corresponding monocyclic relation), and it is not a divisor of a binomial $x^q - 1$ of degree $q < \omega$.

When we use primitive polynomials, the obtained results in our previous paper [10] are very good, but the next question was: How good uniformly distributed sequences we will obtain if we use irreducible instead of primitive polynomials.

IV. Algorithm and Numerical Simulations

A. Algorithm

According to Theorem 2 and exposed algorithm in [1], in this paper we propose this algorithm for construction of sequences, when in $s-$ dimensional case we set base b to be arbitrary prime.

1) Input the dimension s, the base b, the number of the points of the net and the coefficients of the monocyclic relations L, where b is arbitrary prime.
2) For initial direction numbers V_j we use arbitrary random numbers which satisfies the condition of Theorem 1.
3) Construct the direction matrix which corresponds to the LHRR L.
4) For $k = 1, \ldots, s$ a generate the one-dimensional sequence $(P^{(k)}(i))_{i\geq 0}$.

5) Construction of the sequence $P(i) = ((P^{(1)}(i), P^{(2)}(i), \ldots, P^{(s)}(i)))_{i \geq 0}$, which is $s-$ dimensional sequence.

For the purposes of this work to visualize the $(t, s)-$ sequences constructed by the proposed algorithm, a computer programs is written. It constructs one-dimensional $(t, s)-$ sequences, when in $s-$ dimensional case we set different basis b_k for each coordinate $k = 1, \ldots, s$ where b_k is arbitrary prime.

B. Visualizations Obtained by Software Simulations

On the following Fig.1, Fig. 2, Fig. 3 and Fig. 4 we present some results of this program for different dimensions s and different bases b_k, in the case when we use an irreducible polynomial, which is a primitive polynomial.

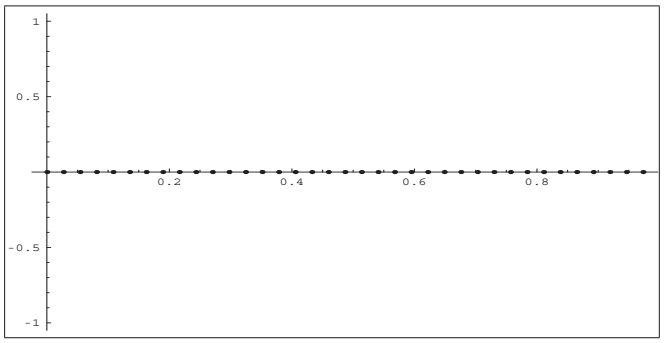


Fig. 1. s=1, N=37, b1=37, a1={35}

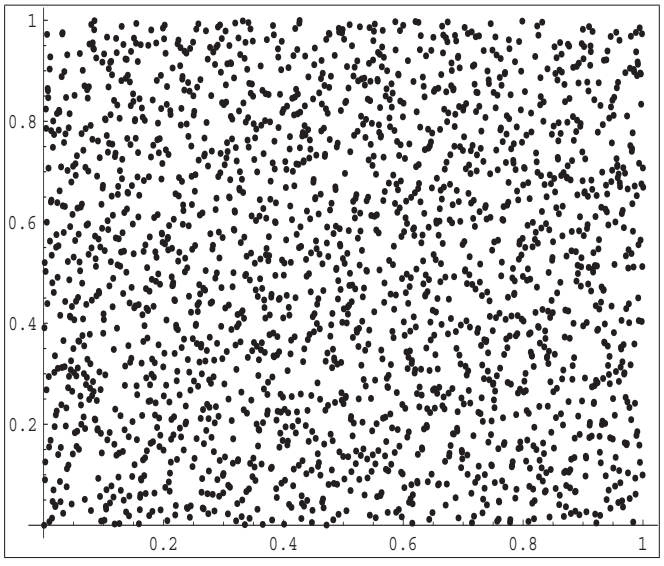


Fig. 2. s=2, N=2048, b1=2,a1={1}, b2=7, a2={5,5}

On the Fig. 5, Fig. 6 and Fig. 7 we present results for different dimensions s and different bases b_k, in the case when we use an irreducible polynomial, which is not a primitive polynomial.

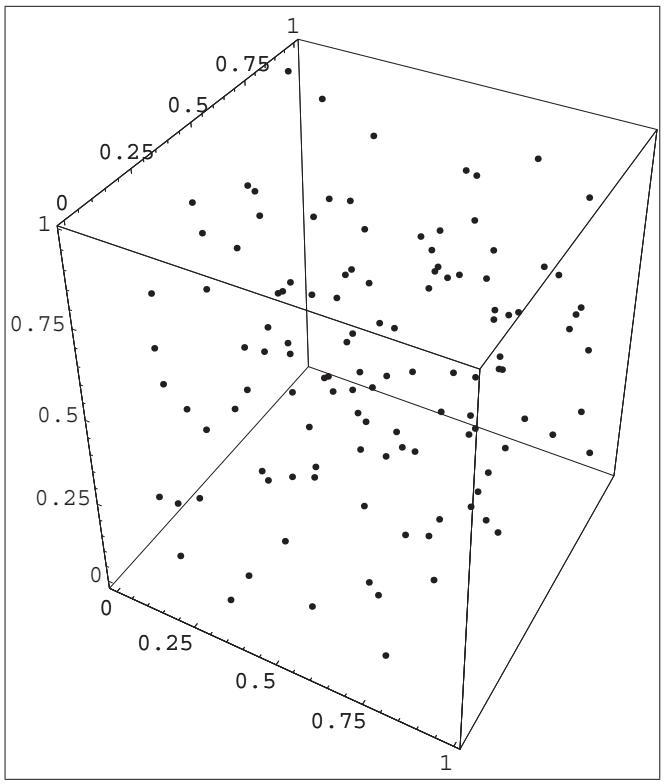


Fig. 3. s=3, N=121, b1=11,a1={3,2,2}, b2=23, a2={13},b3=7, a3={3,0,4,5}

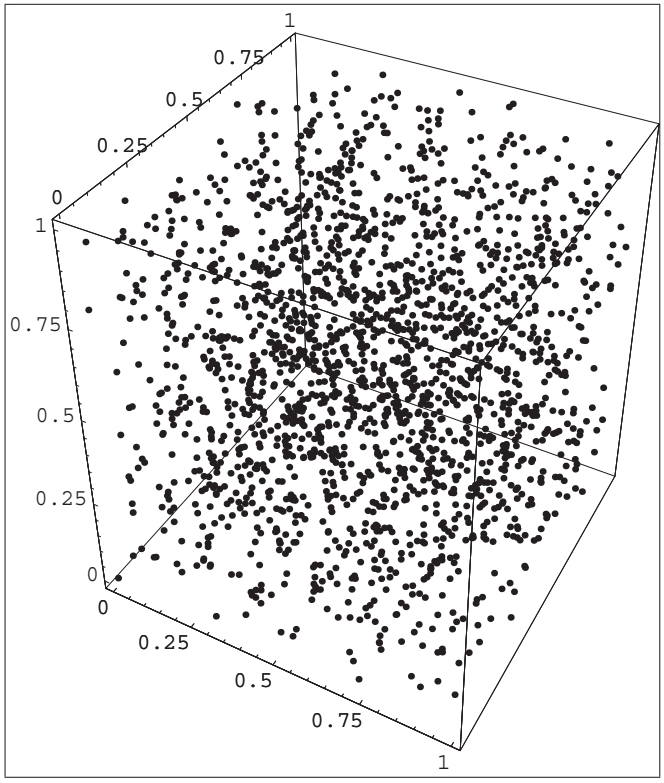


Fig. 4. s=3,N=2000, b1=7,a1={2}, b2=11,a2={6,10,2,3,1}, b3=19,a3={5,17}

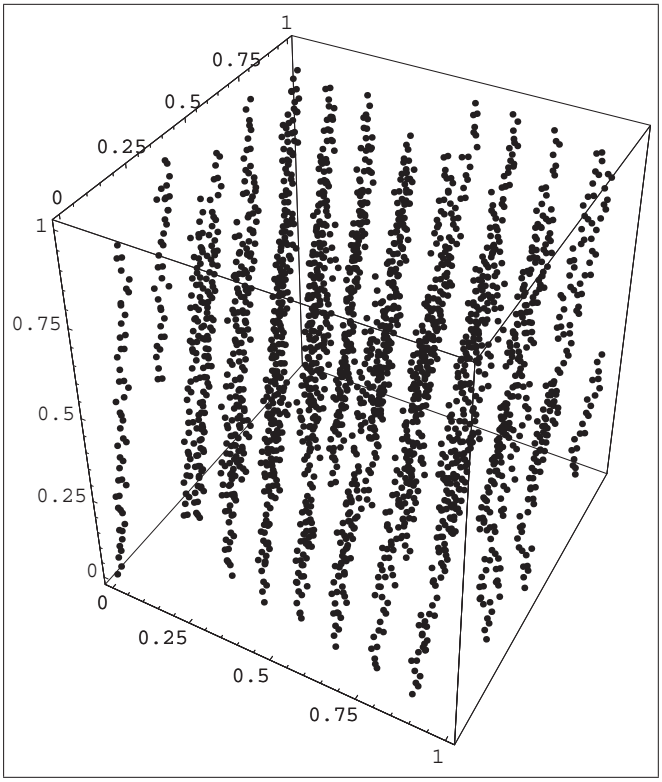


Fig. 5. s=3, N=343*7, b=7,a1={3},a2={2,3},a3={5,5} (Nonprimitive pol.)

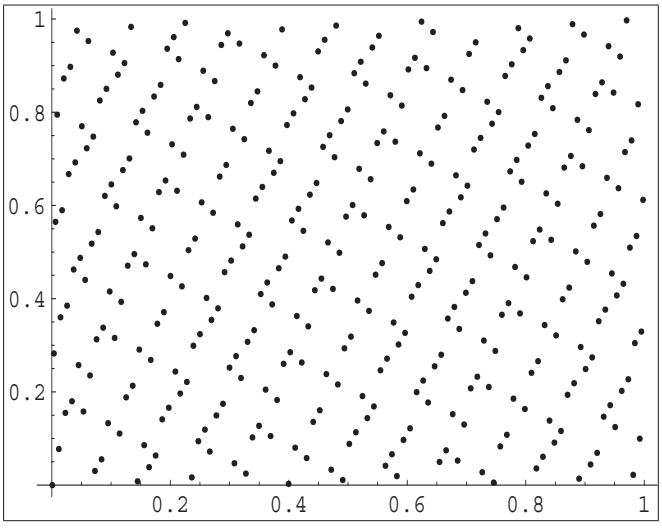


Fig. 6. s=2, N=361, b=19,a1={8},a2={3,0,3}(Nonprimitive pol.)

V. Conclusion

From the presented graphical results, we can see that in the case when we use irreducible polynomial, which is not a primitive polynomial on the figures are shown sequences, which are not uniformly distributed. So, obtained experimental results and their visualizations have verified our expectations, that necessary and sufficient condition for the obtained sequence to be very well uniformly distributed (i.e. relation (1) to be monocyclic) is the polynomial (3) to be primitive, i.e. not only irreducible polynomial.

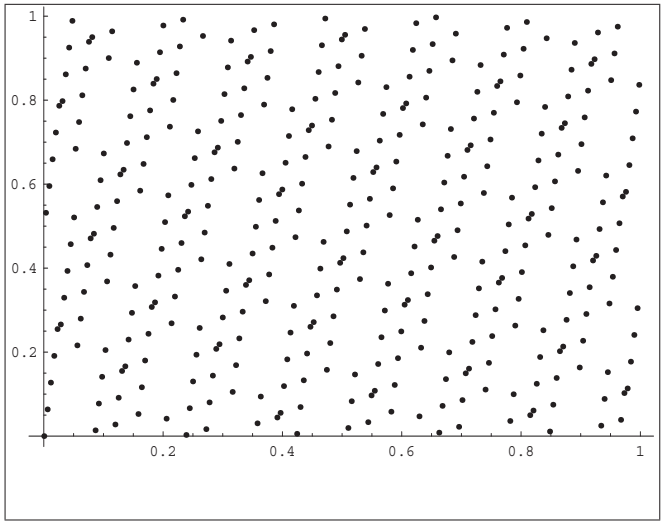


Fig. 7. s=2, N=361, b=19,a1={3},a2={3,1}(Nonprimitive pol.)

Acknowledgment

This research was partially supported by Faculty of Computer Sciences and Engineering, Univ. Ss. Cyril and Methodius, Skopje, as a part of the project "Testing of PRNG based on irrational numbers".

References

[1] V. Dimitrievska Ristovska, V. Grozdanov and A. Atanasov, An effective algorithm for constructing of (t,s)-sequences over **F**b, Proceedings of the V Congress of UMM, Ohrid, 2014

[2] H. Faure, Discrépance de suites associées à un systeme de numeration (en dimension s), Acta Arithmetica, XLI, (1982), 337-351.

[3] H. Faure, Variation on (0,s)-Sequences, Journal of Complexity, 17, 2001, 741-753.

[4] L. Kuipers and H. Niederreiter, Uniform distribution of sequences, John Wiley & Sons, New York., 1974.

[5] R. Lidl and H. Niederreiter, Introduction to finite fields and their applications, Cambridge University Press, New York, USA, 1986.

[6] H. Niederreiter, Random Number Generator and Quasi-Monte Carlo Methods, CBMS - NSF Series in Applied Mathematics, 63, SIAM, Philadelphia, 1992.

[7] I. M. Sobol', Mnogomernye kvadraturnye formuly i funktsii Haara, Izdat. Nauka, Moscow, 1969.

[8] J. G. Van der Corput, Verteilungsfunktionen, Proc. Kon. Akad. Wetensch. Amsterdam, 1935, 38, No 8, 813-821.

[9] V. Zierler, Linear recurring sequences, J. Soc. Industr. Appl. Math., 7 (1), (1959), 31-48.

[10] V. Dimitrievska Ristovska and V. Grozdanov Primitive polynomials as a tool in generation of $(t, s)-$ sequences, Proceedings of CIIT Conference 2015, Molika, Bitola (2015)

Solution of the coupled Boussinesq–Burger's equations by reduced differential transform method

Mohammed O. Al-Amr
Department of Mathematics
College of Computer Sciences and Mathematics
University of Mosul
Mosul, Iraq
E-mail: alamr@uomosul.edu.iq

***Abstract*— In this paper, the reduced differential transform method (RDTM) is utilized to obtain the approximate solution of the coupled Boussinesq–Burger's equations. The series solution is obtained with easily computed components. A comparison is made between our results and the exact solutions. The simplicity and consistency of the proposed method are revealed. Therefore, it can be further extended to solve different types of nonlinear partial differential equations.**

Keywords*— *reduced differential transform method, Boussinesq–Burger's equations, partial differential equations.

I. Introduction

In the last decades, Nonlinear partial differential equations (NPDEs) have become essential tools to model complex phenomena that arising in different aspects of science and engineering such as optical fibers, fluid dynamics, plasma physics, solid state physics, hydrodynamics and acoustics. Therefore, constructing exact and approximate solutions of NLPDEs is of great importance in mathematical sciences [1-10]. Among the NPDEs, we consider the coupled Boussinesq–Burger's equations [11-13]

$$\begin{aligned} u_t - \frac{1}{2}v_x + 2uu_x &= 0, \\ v_t - \frac{1}{2}u_{xxx} + 2(uv)_x &= 0. \end{aligned} \tag{1}$$

The functions $u(x,t)$ and $v(x,t)$ correspond to the horizontal velocity field and the height of the water surface over a bottom horizontal level. As a nonlinear long-wave equation, (1) can describe the evolution of shallow water waves and appears in the investigation of the flow of fluids in dynamical systems. Recently, many researchers have paid their attention to secure the solution of the coupled Boussinesq–Burger's equations [11-20].

The aim of this paper is to derive the approximate analytical solution of the coupled Boussinesq–Burger's equations via the reduced differential transform method. The layout of this article is as follows: Section 2 deals with the overview of the utilized method. In Section 3, we exploit the RDTM for solving the governing equations. Conclusions are given in Section 4.

II. Overview of the Method

As one of the most effective solution methods, the RDTM was developed by the Turkish mathematician Yildiray Keskin [21] in 2009. Since that, it has been widely used for solving various types of differential equations by many scientists [22-26].

Consider a function of two variables $u(x,t)$ and suppose that it can be represented as a product of two single-variable functions, i.e., $u(x,t)=f(x)g(t)$. In the sense of one-dimensional differential transform, one can represent the function $u(x,t)$ as follows:

$$u(x,t)=\left(\sum_{i=0}^{\infty}F(i)x^i\right)\left(\sum_{j=0}^{\infty}G(j)t^j\right)=\sum_{k=0}^{\infty}U_k(x)t^k, \tag{2}$$

where, $U_k(x)$ is called t-dimensional spectrum function of $u(x,t)$.

Let us introduce the following basic definitions of the RDTM [21,22]:

Definition 2.1 If a function $u(x,t)$ is analytic and differentiated continuously with respect to time t and space x in the domain of interest, then let

$$U_k(x)=\frac{1}{k!}\left[\frac{\partial^k}{\partial t^k}u(x,t)\right]_{t=0}, \tag{3}$$

where, the t-dimensional spectrum function $U_k(x)$ is the transformed function. Note that, the lowercase $u(x,t)$ represents the original function, while the uppercase $U_k(x)$ stands for the transformed function.

Definition 2.2 The differential inverse transform of $U_k(x)$ is defined as follows:

$$u(x,t)=\sum_{k=0}^{\infty}U_k(x)t^k. \tag{4}$$

Then, combining equation (3) and (4) we write

$$u(x,t)=\sum_{k=0}^{\infty}\frac{1}{k!}\left[\frac{\partial^k}{\partial t^k}u(x,t)\right]_{t=0}t^k. \quad (5)$$

From the above definitions, it can be established that the concept of the RDTM is developed from the power series expansion.

To demonstrate the key idea of the RDM, consider the following nonlinear partial differential equation, which is written in an operator form

$$Lu(x,t)+Ru(x,t)+Nu(x,t)=g(x,t), \quad (6)$$

with initial condition

$$u(x,0)=f(x), \quad (7)$$

where, $L=\frac{\partial}{\partial t}$, R is a linear operator which has partial derivatives, $Nu(x,t)$ is a nonlinear operator and $g(x,t)$ is an inhomogeneous term.

According to the RDTM, we can construct the following recursive formula:

$$(k+1)U_{k+1}(x,t)=\mathrm{G}_k(x)-RU_k(x)-NU_K(x), \quad (8)$$

where $U_k(x), RU_k(x), NU_k(x)$ and $G_k(x)$ are the transformations of the functions $Lu(x,t), Ru(x,t), Nu(x,t)$ and $g(x,t)$ respectively.

From initial condition (7), we write

$$U_0(x)=f(x). \quad (9)$$

Substituting (9) into (8) and by straightforward iterative calculation, we get the following $U_k(x)$ values. Then, the inverse transformation of the set of values $\{U_k(x)\}_{k=0}^{n}$ gives the n-terms approximation solution as

$$\tilde{u}_n(x,t)=\sum_{k=0}^{n}U_k(x)t^k. \quad (10)$$

Consequently, the exact solution of (6) reads

$$u(x,t)=\lim_{n\to\infty}\tilde{u}_n(x,t). \quad (11)$$

Some fundamental operations of the RDTM are presented in Table 1.

III. Application of the RDTM

In this section, the RDTM is implemented for solving the coupled Boussinesq–Burger's equations given by (1) with initial conditions [13]

$$\begin{aligned} u(x,0)&=\frac{c}{2\alpha}-\frac{\alpha}{2}\tanh(\alpha x),\\ v(x,0)&=-\frac{\alpha^2}{2}\operatorname{sech}^2(\alpha x), \end{aligned} \quad (12)$$

where α and c are constants.

Table I. The Fundamental Operations of RDTM

Functional Form	**Transformed Form**
$u(x,t)$	$U_k(x)=\frac{1}{k!}\left[\frac{\partial^k}{\partial t^k}u(x,t)\right]_{t=0}$
$w(x,t)=u(x,t)\pm v(x,t)$	$W_k(x)=U_k(x)\pm V_k(x)$
$w(x,t)=\alpha u(x,t)$	$W_k(x)=\alpha U_k(x)$ (α is a constant)
$w(x,t)=x^m t^n$	$W_k(x)=x^m\delta(k-n)$, $\delta(k)=\begin{cases}1, & k=0\\ 0, & k\neq 0\end{cases}$
$w(x,t)=x^m t^n u(x,t)$	$W_k(x)=x^m U_{k-n}(x)$
$w(x,t)=u(x,t)v(x,t)$	$W_k(x)=\sum_{r=0}^{k}V_r(x)U_{k-r}(x)=\sum_{r=0}^{k}U_r(x)V_{k-r}(x)$
$w(x,t)=\frac{\partial^r}{\partial t^r}u(x,t)$	$W_k(x)=\frac{(k+r)!}{k!}U_{k+r}(x)$
$w(x,t)=\frac{\partial}{\partial x}u(x,t)$	$W_k(x)=\frac{\partial}{\partial x}U_k(x)$

According to the RDTM and Table 1, the differential transform of (1) reads

$$\begin{aligned} (k+1)U_{k+1}(x)&=\frac{1}{2}\frac{\partial}{\partial x}V_k(x)-2A_k(x),\\ (k+1)V_{k+1}(x)&=\frac{1}{2}\frac{\partial^3}{\partial x^3}U_k(x)-2B_k(x)-2C_k(x), \end{aligned} \quad (13)$$

where the t-dimensional spectrum functions $U_k(x)$ and $V_k(x)$ are the transformed functions. $A_k(x), B_k(x)$ and $C_k(x)$ are the transformed nonlinear terms given by:

$$A_k(x)=\sum_{r=0}^{k}U_r(x)\frac{\partial}{\partial x}U_{k-r}(x), \quad (14)$$

$$B_k(x)=\sum_{r=0}^{k}U_r(x)\frac{\partial}{\partial x}V_{k-r}(x), \quad (15)$$

$$C_k(x)=\sum_{r=0}^{k}V_r(x)\frac{\partial}{\partial x}U_{k-r}(x). \quad (16)$$

Starting with initial conditions (12) and by simple iterative steps, we get

$$
\begin{aligned}
U_0(\mathrm{x}) &= \frac{c}{2\alpha} - \frac{\alpha}{2}\tanh(\alpha x),\\
U_1(\mathrm{x}) &= \frac{c\alpha}{2\cosh(\alpha x)^2},\\
U_2(\mathrm{x}) &= \frac{c^2\alpha\sinh(\alpha x)}{2\cosh(\alpha x)^3},\\
&\vdots
\end{aligned}
\tag{17}
$$

and

$$
\begin{aligned}
V_0(x) &= -\frac{\alpha^2}{2}\operatorname{sech}^2(\alpha x),\\
V_1(x) &= -\frac{\alpha^2 c\sinh(\alpha x)}{\cosh(\alpha x)^3},\\
V_2(x) &= -\frac{1}{2}\frac{\alpha^2\left(2\cosh(\alpha x)^2-3\right)\alpha^2}{\cosh(\alpha x)^4},\\
&\vdots
\end{aligned}
\tag{18}
$$

and so on. Similarly, we can acquire the rest of components with the aid of MAPLE software.

Taking the inverse transformation of the sets $\{U_k(x)\}_{k=0}^{n}$ and $\{V_k(x)\}_{k=0}^{n}$ yields the following n-terms approximate solutions:

$$
\begin{aligned}
\tilde{u_n}(x,t) &= \sum_{k=0}^{n} U_k(x)t^k\\
&= \frac{c}{2\alpha} - \frac{\alpha}{2}\tanh(\alpha x) + \frac{c\alpha}{2\cosh(\alpha x)^2}t\\
&\quad + \cdots + \frac{1}{n!}\left[\frac{\partial^n}{\partial t^n}\left(\frac{c}{2\alpha} - \frac{\alpha}{2}\tanh(\alpha x - ct)\right)\right]_{t=0} t^n,
\end{aligned}
\tag{19}
$$

$$
\begin{aligned}
\tilde{v_n}(x,t) &= \sum_{k=0}^{n} V_k(x)t^k\\
&= -\frac{\alpha^2}{2}\operatorname{sech}^2(kx) - \frac{\alpha^2 c\sinh(\alpha x)}{\cosh(\alpha x)^3}t + \cdots\\
&\quad + \frac{1}{n!}\left[\frac{\partial^n}{\partial t^n}\left(-\frac{\alpha^2}{2}\operatorname{sech}^2(\alpha x - ct)\right)\right]_{t=0} t^n.
\end{aligned}
\tag{20}
$$

Accordingly, the exact solutions of (1) read

$$
\begin{aligned}
u(x,t) &= \lim_{n\to\infty}\tilde{u_n}(x,t) = \frac{c}{2\alpha} - \frac{\alpha}{2}\tanh(\alpha x - ct),\\
v(x,t) &= \lim_{n\to\infty}\tilde{v_n}(x,t) = -\frac{\alpha^2}{2}\operatorname{sech}^2(\alpha x - ct).
\end{aligned}
\tag{21}
$$

In order to measure the validity of the resultant solutions, we plot the 4th order approximate solutions in Fig. 1 and Fig. 2. Also, the absolute errors are listed in Table 2 and depicted in Fig. 3 and Fig. 4. One can see that the approximate solutions are accurate in comparison with the exact solutions.

IV. Conclusions

In this work, the reduced differential transform method is well availed to acquire the approximate analytical solution of a nonlinear long-wave equation, namely the coupled Boussinesq–Burger's equations. The numerical results are presented accompanied by 3D graphics. A comparison is made between the approximate solution and the exact solution to demonstrate the validity of the first one. It is observed that the computational work of this approach is smaller than other existing ones. Various NLEEs might be investigated by the RDTM in future works.

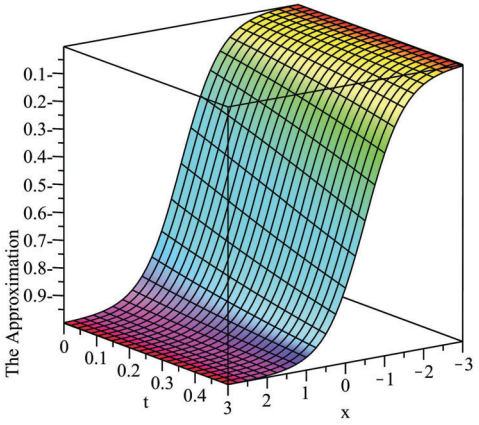


Fig. 1. 3D profile of $\tilde{u_4}(x,t)$ when $c=1$ and $\alpha=-1$.

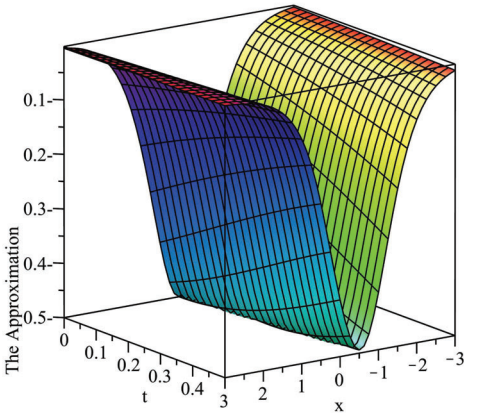


Fig. 2. 3D profile of $\tilde{v_4}(x,t)$ when $c=1$ and $\alpha=-1$.

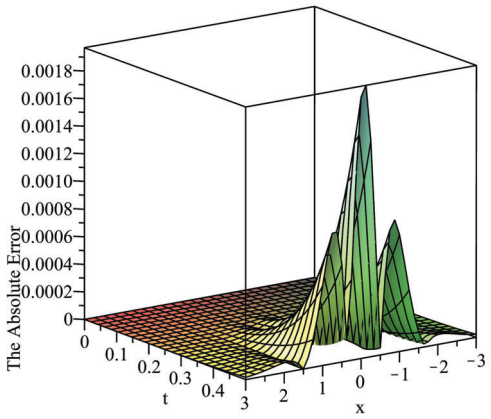


Fig. 3. 3D profile of the absolute error of $\tilde{u}_4(x,t)$ when $c=1$ and $\alpha=-1$.

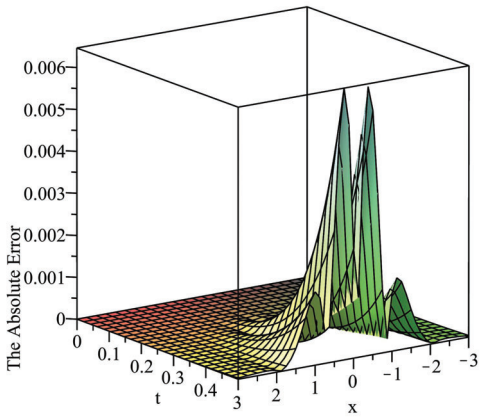


Fig. 4. 3D profile of the absolute error of $\tilde{v}_4(x,t)$ when $c=1$ and $\alpha=-1$.

Table II. The Absolute Errors of $\tilde{u}_4(x,t)$ and $\tilde{v}_4(x,t)$ for Various Values of t and x When $c=1$ and $\alpha=-1$

t	x	$\lvert u(x,t)-\tilde{u}_4(x,t)\rvert$	$\lvert v(x,t)-\tilde{v}_4(x,t)\rvert$
0.1	-2	2.42000×10^{-8}	2.30000×10^{-9}
	-1	2.40600×10^{-7}	5.56100×10^{-7}
	0	6.63910×10^{-7}	1.88000×10^{-7}
	1	2.22400×10^{-7}	5.73700×10^{-7}
	2	2.36000×10^{-8}	6.91000×10^{-9}
0.3	-2	5.78000×10^{-6}	9.37800×10^{-7}
	-1	6.26378×10^{-5}	1.24797×10^{-4}
	0	1.56306×10^{-4}	1.31519×10^{-4}
	1	4.92939×10^{-5}	1.39100×10^{-4}
	2	5.69220×10^{-6}	2.49745×10^{-6}
0.5	-2	7.23781×10^{-5}	3.74940×10^{-5}
	-1	8.45610×10^{-4}	1.37640×10^{-3}
	0	1.89191×10^{-3}	2.60947×10^{-3}
	1	5.76568×10^{-4}	1.74699×10^{-3}
	2	7.16051×10^{-5}	3.97898×10^{-5}

References

[1] A.M. Wazwaz, "A sine–cosine method for handling nonlinear wave equations," Math. Comput. Modelling, vol. 40, no. 5-6, pp. 499–508, 2004.

[2] R. Hirota, "Exact envelope-soliton solutions of a nonlinear wave equation," J. Math. Phys., vol. 14, no. 805, pp. 805–809, 1973.

[3] J.H. He and X.H. Wu, "Exp-function method for nonlinear wave equations," Chaos Solitons Fractals, vol. 30, no. 3, pp. 700–708, 2006.

[4] A.J. Al-Sawoor and M.O. Al-Amr, "Numerical solution of a reaction-diffusion system with fast reversible reaction by using Adomian's decomposition method and He's variational iteration method," Al-Rafidain J. Comput. Sci. Math., vol. 9, no. 2, pp. 243–257, 2012.

[5] A.J. Al-Sawoor and M.O. Al-Amr, "A new modification of variational iteration method for solving reaction-diffusion system with fast reversible reaction," J. Egyptian Math. Soc., vol. 22, no. 3, pp. 396–401, 2014.

[6] E. Fan, "Uniformly constructing a series of explicit exact solutions to nonlinear equations in mathematical physics," Chaos Solitons Fractals, vol. 16, no. 5, pp. 819–839, 2003.

[7] S. Liu, Z. Fu, S.D. Liu and Q. Zhao, "Jacobi elliptic function expansion method and periodic wave solutions of nonlinear equations," Phys. Lett. A., vol. 289, no. 1-2, pp. 69-74 , 2001.

[8] M.O. Al-Amr, "Exact solutions of the generalized (2+1)-dimensional nonlinear evolution equations via the modified simple equation method," Comput. Math. Appl., vol. 69, no. 5, pp. 390–397, 2015.

[9] M.O. Al-Amr and S. El-Ganaini, "New exact traveling wave solutions of the (4+1)-dimensional Fokas equation," Comput. Math. Appl., vol. 74, no. 6, pp. 1274-1287, 2017.

[10] A. Zerarka, S. Ouamane and A. Attaf, "On the functional variable method for finding exact solutions to a class of wave equations," App. Math. Comput., vol. 217, no. 7, pp. 2897–2904, 2010.

[11] A. Chen and X. Li, "Darboux transformation and soliton solutions for Boussinesq-Burgers equation," Chaos, Solitons and Fractals, vol. 27, no. 1, pp. 43-49, 2006.

[12] Z. Wang and A. Chen, "Explicit solutions of Boussinesq-Burgers equation," Chinese Physics, vol. 16, no. 5, pp. 1233–1238, 2007.

[13] A.M. Wazwaz, "A variety of soliton solutions for the Boussinesq-Burgers equation and the higher-order Boussinesq-Burgers equation," Filomat, vol. 31, no. 3, pp. 831–840, 2017.

[14] Z. Liang, Z. Li-Feng and L. Chong-Yin, "Some new exact solutions of Jacobian elliptic function about the generalized Boussinesq equation and Boussinesq–Burgers equation," Chinese Physics B, vol. 17, no. 2, pp. 403-410, February 2008.

[15] L. Gao, W. Xu, Y. Tang and G. Meng, "New families of travelling wave solutions for Boussinesq–Burgers equation and (3 + 1)-dimensional Kadomtsev–Petviashvili equation," Physics Letters A, vol. 366, no. 4-5, pp. 411–421, 2007.

[16] A. Esen, O. Taşbozan and S. Kutluay, "Approximate Analytical Solutions of the Damped Burgers and Boussinesq-Burgers Equations," Çankaya University Journal of Science and Engineering, vol. 11, no. 1, pp. 65–76, 2014.

[17] M. Khalfallah, "Exact traveling wave solutions of the Boussinesq–Burgers equation," Math. Comput. Model., vol. 49, no. 3-4, pp. 666–671, 2009.

[18] A.S. Abdel Rady, E.S. Osman and M. Khalfallah, "Multi-soliton solution, rational solution of the Boussinesq–Burgers equations," Commun. Nonlinear Sci. Numer. Simulat., vol. 15, no. 5, pp. 1172–1176, 2010.

[19] S. Kumar, A. Kumar and D. Baleanu, "Two analytical methods for time-fractional nonlinear coupled Boussinesq–Burger's equations arise in propagation of shallow water waves," Nonlinear Dyn., vol. 85, no. 2, pp. 699–715, 2016.

[20] L.K. Ravi, S.S. Ray and S. Sahoo, "New exact solutions of coupled Boussinesq–Burgers equations by Exp-function method," J. Ocean Eng. Sci., vol. 2, no. 1, pp. 34–46, 2017.

[21] Y. Keskin and G. Oturanc, “Reduced differential transform method for partial differential equations,” Int. J. Nonlinear Sci. Numer. Simul., vol. 10, no. 6, pp. 741–749, 2009.

[22] M.O. Al-Amr, “New applications of reduced differential transform method,” Alexandria Eng. J., vol. 53, no. 1, pp. 243–247, 2014.

[23] A.J. Al-Sawoor and M.O. Al-Amr, “Reduced differential transform method for the generalized Ito system,” Int. J. Enhanc. Res. Sci. Technol. Eng., vol. 2, no. 11, pp. 135-145, 2013.

[24] R. Abazari and M. Abazari, “Numerical study of Burgers–Huxley equations via reduced differential transform method,” Comp. Appl. Math., vol. 32, pp. 1-17 , 2013.

[25] M.A. Abdou and A.A. Soliman, “Numerical simulations of nonlinear evolution equations in mathematical physics,” Int. J. Nonlinear Sci., vol. 12, pp. 131-139, 2011.

[26] M.S. Mohamed and K.A. Gepreel, “Reduced differential transform method for nonlinear integral member of Kadomtsev–Petviashvili hierarchy differential equations,” J. Egyptian Math. Soc., vol. 25, pp. 1–7, 2017.

Building And Selection An Optimal Mathematical Model Describing A Scientific Or Engineering Process

Radoslav Mavrevski
Department of Electrical Engineering, Electronics and Automatics,
Faculty of Engineering,
University Center for Advanced Bioinformatics Research
South-West University "Neofit Rilski", 66 Ivan Mihaylov Str.
2700 Blagoevgrad, Bulgaria
radoslav_sm@abv.bg

Metodi Traykov
Department of Electrical Engineering, Electronics and Automatics,
Faculty of Engineering,
University Center for Advanced Bioinformatics Research
South-West University "Neofit Rilski", 66 Ivan Mihaylov Str.
2700 Blagoevgrad, Bulgaria
metodi_043@abv.bg

***Abstract*—The presented work demonstrates different stages in the modeling of relation-ships between mass phenomena and processes by regression analysis using the least squares method. For this purpose, is used the software package Matlab and the built-in fitting functions. Mathematical models can be used to predict process output, calibrate, or optimize processes. Since the modeling of different devices and phenomena is important to both engineering and science, engineers and scientists have very different reasons for doing mathematical modeling. In this study the experimental data are taken from existing literature and represent example of a of the length of the service depended on the number of repaired components.**

Keywords—Modeling; regression analysis; least squares method; optimal model

I. INTRODUCTION

The mathematical and computer modeling is an integral part of research in many areas. Models and modeling are quite common in engineering, physical, medico biological and other sciences. The Mathematical model is a description of the system using mathematical means. The essential moment in the mathematical model is the degree of accuracy, respectively the adequacy with the real object, which is closely related to paradigms in the particular scientific field [1]. The problem of choosing an optimal model is one of the main problems in data analysis. The study of choosing an optimal mathematical model is current scientific problem, as evidenced by the large number of world-known scientists working on this problem.

In order to determine the individual optimal models in the candidate model classes in modeling, are used various regression (fitting) methods, as a the most widely used least squares method (LS), robust regression (RR) and minimax (MM). In essence, these methods represent criteria for the optimism of the model with respect to the values of the parameters in a given class of models. The LS method for linear models as they are used in the present work are relatively easy to compute, and so they have early history in their use. [2].

The mathematical procedure for this method is to find the best approximation of the curve to a given set of points by minimizing the sum of the squares of the deviations (residuals) of the points from curve. [3].

In modeling of the relationships between variables, choosing of the optimal model describing experimental data is an particular scientific problem, starting with a careful analysis of the experimental data available, and goes through the following stages: (a) creating a scatter plot of the data and checking whether there is any trend in these data (Plotting points by *X* and *Y*) and whether there are among obvious wrong data; if there are obvious erroneous data, then we ignore them and do again the measurements or observations; (b) when there is an obvious trend in the set of data, we try to find the class model, expressing this trend of interrelationships of the studied factors; while observing the cloud with the experimental data, we turn heuristically to classes of models (linear quadratic, cubic, exponential, hyperbolic and etc.), which describe the set of experimental data in optimal way concerning to minimize the approximation error; (c) after the selection of classes of models, we use different fitting methods for finding the best models in the given classes; finding the individual model in the given classes is usually made by various fitting methods, such as the most widely used method of least squares fitting or other such as robust fitting, minimax fitting etc.; и (d) the selected candidate "optimal" models in each class mentioned above, R^2, are compared by adjusted R^2, the residual sum of squares (RSS), or some other known criterion.

II. MHETODS AND TEHNOLOGIES

A. Regression analysis

In mathematical and computer modeling in the present work, regression analysis was used. Regression analysis is a method for examining of relationships, between one dependent and multiple independent variables. It includes techniques for modeling and analyzing of the relationships between several variables, as a focusing is on the relationships between dependent variable and one (single factor regression analysis) or more (multifactor regression analysis) independent variables. In particular, regression analysis helps to understand how the value of the dependent variable changes when the independent variables change.

The regression analysis is widely used for predicting and forecasting. It is also used to understand which of the independent variables are related to the dependent variable and to explore the forms of that relationship. Regression analysis can

be used to make a inference to established relationships between dependent and independent variables. However, it can lead to "false" relationships, so caution is recommended [4].

The regression model is used to model the relationship between dependent variable y and independent variables $x = (x_1, x_2, ..., x_p)^T$.

Linear regression of the least squares method is the most widely used modeling method. It is not only the most widely used modeling method but is adapted to a wide range of situations that are outside its direct scope. It has a major role in many other modeling methods: nonlinear regression on the least squares, weighted regression of lest squares and etc.

B. *Least squares methods*

In linear regression as well as nonlinear regression, it is found that the scattering of the data around the optimal curve follows the normal distribution. This assumption leads to the goal of the lest square method: minimize the sum of squares of vertical deviations or (distances between *Y*-values) between points and the curve [1]. The criterion of optimality with respect to the parameters of this method is to minimize the sum:

$$\sum_{\alpha=1}^{n} (y_\alpha - \hat{y}_\alpha)^2 \qquad (1)$$

where y_α is the measured value of the dependent variable, a $\hat{y}_\alpha$ is the calculated value?

C. *Matlab*

Matlab provides a plurality of graphical user interfaces and curve fitting functions using the least squares method and its modifications. For linear regression, the function lsqcurvefit in Matlab is used in the current work and has the following syntax:

$$x = lsqcurvefit(fun, x0, xdata, ydata), \qquad (2)$$

where *fun* is fitting function, *x0* is a vector with the initial values (starting) of the searched parameters in the model, *xdata* и *ydata,* are experimental data, respectively on *X* и *Y*.

III. Results and Discusion

The experimental data in this work is taken from the book „Regression Analysis by Example“ by Walter Hewhart и Samuel Wilks [5]. The data is an illustrative example of a company that trades and repairs computers. The purpose is to investigate the relationship between service duration and the number of electronic components in the computer that need to be repaired or replaced. The data consists of the length of the service in minutes (the variable for response) and the number of repaired components (the forecast variable). The sample of service records is presented in Table I.

TABLE I. Duration Of Service In Minutes And Number Of Repaired Components.

№	Minutes	Number of components
1	23	1
2	29	2
3	49	3
4	64	4
5	74	4
6	87	5
7	96	6
8	97	6
9	109	7
10	119	8
11	149	9
12	145	9
13	154	10
14	166	10

The scatter plot of experimental data built with Matlab is shown in Fig. 1.

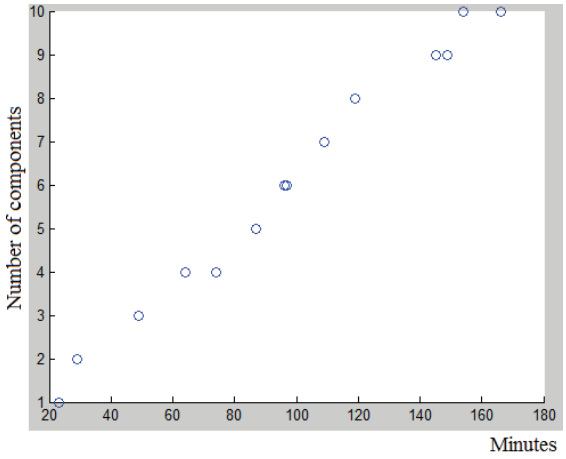


Fig. 1. Scatter plot

After the fitting with mathematical model $y = a + bx$ using the lest squares method in Matlab, we obtain optimal parameter values in the model: a=-0.1896 и b=0.0637 (see Table II). The result of the fitting is shown in Fig.2.

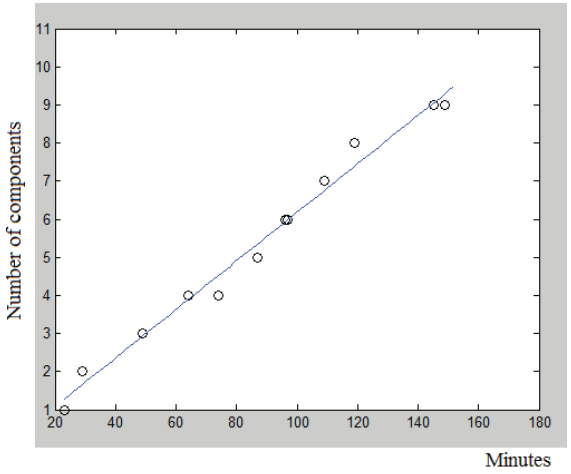


Fig. 2. Experimental data and optimal model

TABLE II. REGRESSION MODELS AND PARAMETER VALUES.

Regression model	Parameters	Criteria		
$y = a + bx$	a=-0.1896	R^2	adjusted R^2	RSS
	b=0.0637	*0.987*	*0.986*	*1.432*

As can be seen form Table II the values of the relevant criteria for the found individual optimal model in linear class model describing number of components - minutes relationship, are very good (R^2 and adjusted R^2 are very close to 1 and RSS is very small).

We use only linear class model becaose, only this class have a accordance with trend in the set of data, observinging the experimental data cloud.

The linear regression of the lest squares has earned its place as the main modeling tool for processes due to its efficiency and completeness. Although there are types of data that are better described by functions that are nonlinear in parameters, many processes in science and engineering are well described by linear models. This is because either the processes are inherently linear, or because in short bounds, each process can be well approximated to a linear model.

Estimates of unknown parameters obtained from linear least squares regression are the optimal estimates from a wide range of possible parameter estimates in the usual assumptions used to model the processes. Practically, the linear regression of the lest squares makes very efficient use of the data

The major disadvantages of linear least squares are possibly poor extrapolating properties and sensitivity to deviations. This means that linear models cannot be effective to extrapolate process results for which we cannot collect data in the area that interests us. Sure, extrapolation is potentially dangerous, regardless of the model type.

Nonlinear regression of the lest squares expands the linear regression of the lest squares to use with much larger and more general class functions. [6] The greatest advantage of nonlinear least squares regression over many other techniques is the wide range of functions that can be adapted.

ACKNOWLEDGMENT

This work is partially supported by the project of the Bulgarian National Science Fund, entitled: Bioinformatics research: protein folding, docking and prediction of biological activity, NSF I02 /16 12.12.14.

REFERENCES

[1] K. Sadanori, G. Kitagawa, Information Criteria and Statistical Modeling, New York: Springer Science+Business Media, LLC, 2008.

[2] R. L. Mason, R. F. Gunst, J. L. Hess, Statistical Design and Analysis of Experiments, New Jersey: Published by John Wiley & Sons, Inc., Hoboken, 2003.

[3] S. Weisberg, Applied linear regression. 2 ed., New York: Wiley., 1985

[4] J. S. Armstrong, F. Collopy, Error Measures For Generalizing About Forecasting Methods: Empirical Comparisons, International Journal of Forecasting, vol. 8, no. 1, pp. 69-80, 1992.

[5] W. Hewhart, S. Wilks, Regression Analysis by Example. 4 ed., New York: Mount Sinai School of Medicine, 2006.

[6] R. Mavrevski, P. Milanov, M. Traykov, N. Pencheva, Assessment of different model selection criteria by generated experimental data, WSEAS Transactions on Computers, vol. 16, pp. 260-268, 2017, ISSN:2224-2872.

Cloud based Data Acquisition and Annotation Architecture for Weed Control

Petre Lameski, Eftim Zdravevski, Vladimir Trajkovik, Andrea Kulakov
Faculty of Computer Science and Engineering
Sts. Cyril and Methodius University, Skopje, Macedonia
E-mails: {petre.lameski,eftim.zdravevski,vladimir.trajkovik,andrea.kulakov}@finki.ukim.mk

***Abstract*—In this paper we present a short evaluation of a cloud based architecture for data acquisition and annotation. We evaluate the implemented system for annotation and give initial results on the ability of the system to produce accurate labels on the data. The used data is consisted of plant field images. The users partially annotate the data and we use segmentation algorithms for enriching the annotation of the images. We compare three different segmentation algorithms used for the annotation. The results show that Grabcut algorithm is better than Watershed and nearest-neighbor approaches, but there is still room for improvement.**

***Index Terms*—weed control, image processing, data acquisition, data annotation, data segmentation**

I. Introduction

The novel algorithms for classification and segmentation in Machine Vision need large amount of data to generate models with substantial classification performance. While deep learning algorithms are very good at extracting information from unannotated data using unsupervised learning methods [1], there is still need for annotated data, especially in niche domains where such data does not exist. The image annotation is a well established research domain and there are quite a few examples in the literature [2].

One of the ways to boost the process of labeling data is to use crowd sourcing [3], however for niche domains, such plant species recognition or weed detection, such approach could not be quite applicable since it is difficult by untrained people to sometimes distinguish different plants. This can also be overcome by using artificial models of the plants which gives good results when building models using deep learning [4]. In agricultural environment there are differences between some types of plants based on their geographic location and the used seeds, resulting in non-standardized types, especially for individual and small scale farmers. For this reason, a platform for image data annotation from plants is needed to resolve the lack of annotated data for different plant phenotypes, especially for resolving the problem of detecting weeds in plantations.

In this paper we present an architecture for data annotation that uses cloud based processing. The expert is able to manually select parts of the plants that need to be annotated and segmentation algorithms are used for inferring the rest of the plants. We present the results of several conventional algorithms for manual or semi-automated segmentation and give conclusion and future directions for the system development .

The paper is organized as follows. In Section II, we describe the used algorithms for segmentation and the used architecture and in Section III, we present the obtained results and give a short conclude the paper.

This work was partially financed by the Faculty of Computer Science and Engineering at the Sts. Cyril and Methodius University, Skopje, Macedonia. We also acknowledge the support of Microsoft Azure for Research through a grant providing resources for this work.

II. Methods

The general architecture of a weed control system is presented in [5]. In this paper we describe in more detail the module of the architecture that is used for data annotation. The general flow is already described in the previous work, so the focus of this paper is on the annotation problem, proposing several methods for resolving the data annotation. An example input image and the user selection are depicted in Figure 1. The images are taken from the dataset described in [6].

We use three different segmentation algorithms to enrich the dataset. The first algorithm is one of the most used algorithms for semi-automated segmentation of images, Grabcut [7]. The algorithm uses the user input to initialize the foreground and background regions and can be applied for both plant and weed segmentation. We use foreground-background segmentation of the plant pixels by defining as certain background all markings from the weed and the ground and define as certain foreground only the plant markings. For weed detection we repeat the process, while considering the weed markings as foreground, and plant and ground as background. We use similar approach for the other two segmentation algorithms, Watershed [8] and a nearest-neighbor based segmentation that uses only the users input to generate the pixel distribution for foreground and background segmentation. In this case we use the K-D tree [9] approach to search for the nearest neighbor pixel in each of the classes.

III. Results and conclusion

Based on the performed experiments using the dataset the best results were obtained with Grabcut. The Watershed algorithm was too restrictive and failed to identify other plant pixels than the ones that were marked by the user, and the K-D tree based search was too inclusive. The plant and weed

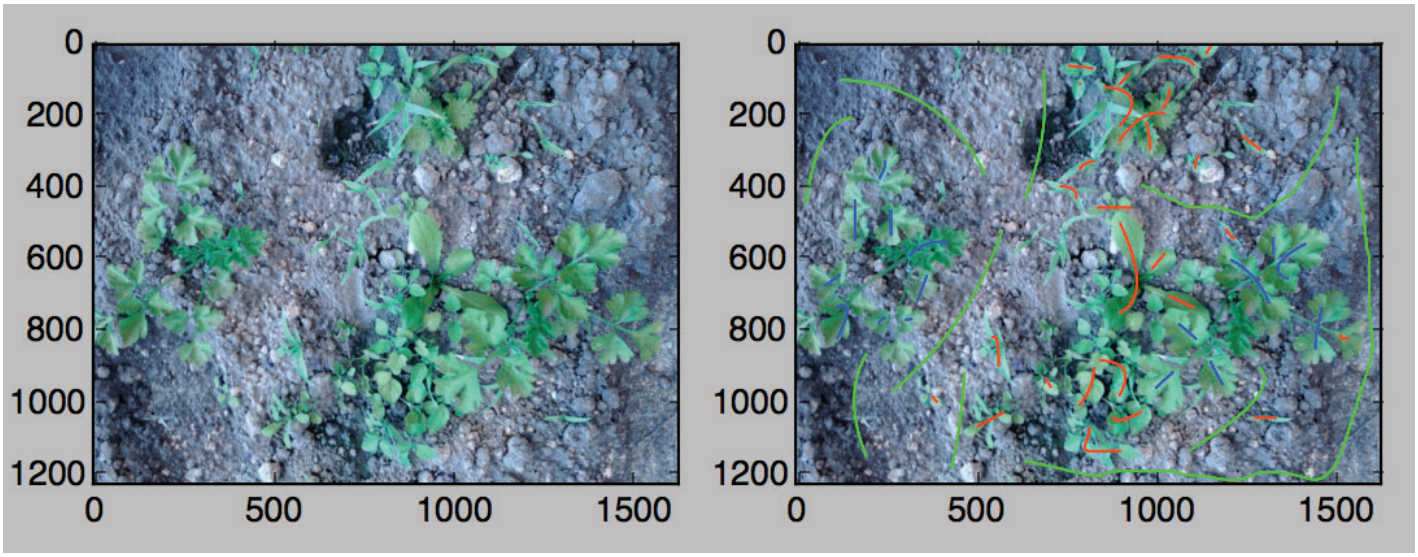


Fig. 1. User selection from plant image, different color lines annotate different parts: weed, useful plant and ground

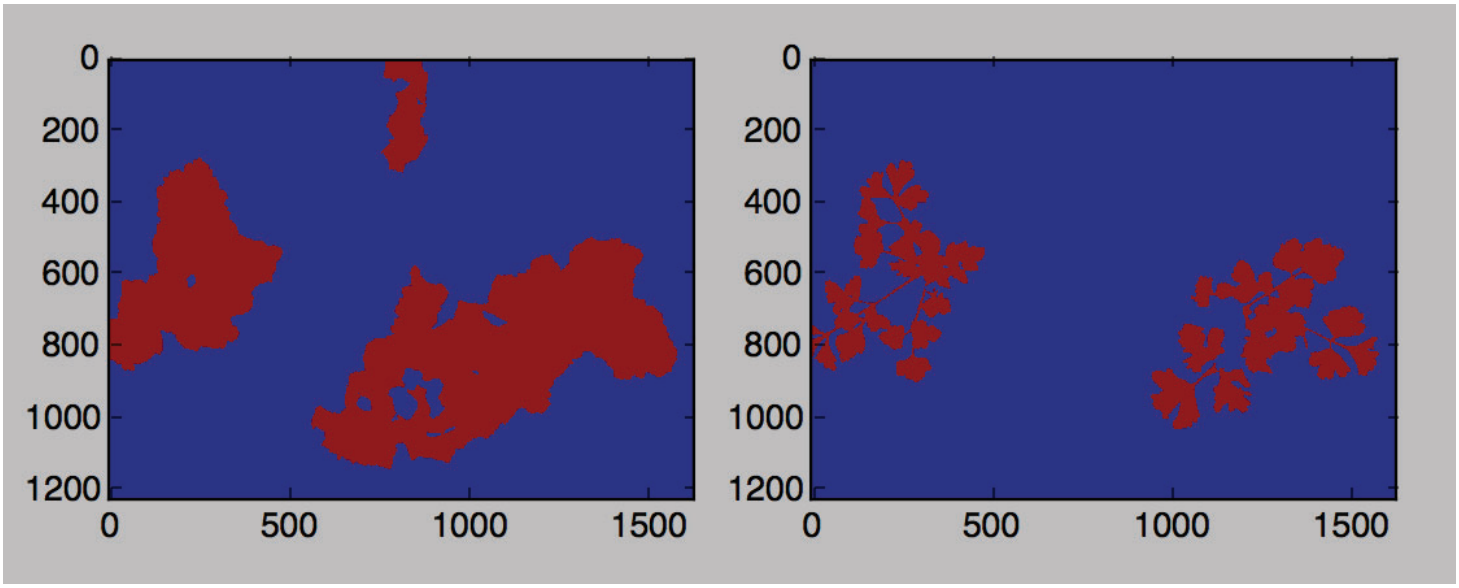


Fig. 2. Segmentation result using Grabcut

pixels are too similar in color so the K-D tree search included most of the weed pixels when searching for plant pixels and vice versa. The results obtained from the Grabcut algorithm are depicted in Figure 2. As it can be observed, the Grabcut algorithm successfully identifies most of the target pixels. It also includes some of the ground pixels in the vicinity of the plant images, but for some of the classification algorithms this would not be a problem, especially if we are targeting weed patches. However, Grabcut is also not the best choice if there are overlapping leaves of the plants, such as in some of the examples in the analyzed dataset, because it tends to also include pixels from the weed plant which could be a problem for the annotation process.

In the future, additional color spaces should be considered for the segmentation task, including vegetation indexes that use RGB images. Unsupervised segmentation algorithms could also be considered prior to including the user input to obtain patches or groups of pixels that belong to a certain class or object.

REFERENCES

[1] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature*, vol. 521, no. 7553, p. 436, 2015.

[2] A. Hanbury, "A survey of methods for image annotation," *Journal of Visual Languages & Computing*, vol. 19, no. 5, pp. 617–627, 2008.

[3] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Computer Vision and Pattern*

Recognition, 2009. CVPR 2009. IEEE Conference on. IEEE, 2009, pp. 248–255.

[4] J. Ubbens, M. Cieslak, P. Prusinkiewicz, and I. Stavness, "The use of plant models in deep learning: an application to leaf counting in rosette plants," *Plant Methods*, vol. 14, no. 1, p. 6, Jan 2018. [Online]. Available: https://doi.org/10.1186/s13007-018-0273-z

[5] P. Lameski, E. Zdravevski, V. Trajkovik, and A. Kulakov, "Cloud-based architecture for automated weed control," in *Smart Technologies, IEEE EUROCON 2017-17th International Conference on*. IEEE, 2017, pp. 757–762.

[6] ——, "Weed detection dataset with rgb images taken under variable light conditions," in *International Conference on ICT Innovations*. Springer, 2017, pp. 112–119.

[7] C. Rother, V. Kolmogorov, and A. Blake, "Grabcut: Interactive foreground extraction using iterated graph cuts," in *ACM transactions on graphics (TOG)*, vol. 23, no. 3. ACM, 2004, pp. 309–314.

[8] S. Beucher, "The watershed transformation applied to image segmentation," *SCANNING MICROSCOPY-SUPPLEMENT-*, pp. 299–299, 1992.

[9] M. Greenspan and M. Yurick, "Approximate kd tree search for efficient icp," in *3-D Digital Imaging and Modeling, 2003. 3DIM 2003. Proceedings. Fourth International Conference on*. IEEE, 2003, pp. 442–448.

Development of Rapid-Testing Technology for the Screening of Food Safety

Ge Long[1,2]
1. School of Electrical & Automatic Engineering
Changshu Institute of Technology
Changshu, PR China
2. Changshu Science and Technology Innovation Park,
Changshu, PR China

Shuo Pan, Lv Gang, Zhu Peiyi*
School of Electrical & Automatic Engineering
Changshu Institute of Technology
Changshu, PR China
* zhupy@cslg.edu.cn

***Abstract*— Considering the complex light path system, low detection efficiency and low detection precision in the traditional screening of food safety, some improvement methods are proposed for the traditional screening equipment of food safety. The equipment is adopted rear spectroscopic measurement technology and linear array Charge-coupled Device (CCD) which can realize real-time detection within visible range. Automatic liquid surface sensing and tracking technology in the process of sample injection are accepted for improving the test accuracy. Multicomponent pigment measurement experimental results prove its stability which verifies the reasonableness of software and hardware designing in the new screener.**

Keywords—food safety; rapid-testing technology; screening test; holographic concave grating; full automation equipment

I. INTRODUCTION

Food safety is a major issue concerning public health, which becomes the focus issue of all countries around the world [1]. Typically it may take four days or more to complete the entire procedure by laboratory routine detection method. Even if screening test, we should speed two days to complete testing [2, 3]. Due to the acceleration of food circulation speed, the use of laboratory routine detection methods are seldom effective means of monitoring food safety because of the time required to obtain results [4]. In most instances, preliminary screening of food safety can best be accomplished through the use of rapid detection equipment [5-7]. Rapid methods are more sensitive than conventional assays, but the increased sensitivity has also created interesting challenges and problems for the regulatory agencies and the food industry [8].

According to the characteristics of modern micro-spectrum technology [9], some improvement methods are proposed for the traditional screening equipment of food safety. The equipment is adopted rear spectroscopic measurement technology which can realize real-time detection within visible range. We present flat field holographic concave grating to achieve spectrometer and reduce the wavelength error. Meanwhile, The equipment adopts a new Charge-coupled Device (CCD) sensors which is replaced the traditional optical path generation and detection system. CCD sensors can obtain the full band spectral signal in the visible range and the whole band was detected simultaneously. In order to enhance the detection speed, refine sample and reagent automatic extraction technology are recommend. Automatic liquid surface sensing and tracking technology in the process of sample injection are accepted for improving the test accuracy. The new technology can simplify the complexity of the rapid detection system, reduce the volume and improve the test speed, which can meet the needs of the rapid development of food safety testing industry in China. Paper organization is as follows: Section II describe the system design technical scheme. Section III shows every subsystem architecture and design methodology. Then finishes the debugging of the sample machine in Section IV. Section V contains the results and conclusion besides.

II. THE SYSTEM DESIGN TECHNICAL SCHEME

A. Food Safety Detection Based on Lambert-Beer Law

Absorption spectroscopy is employed as an analytical chemistry tool to determine the presence of a particular substance in a sample and, in many cases, to quantify the amount of the substance present. The new food safety detection equipment is based on Absorption spectroscopy which obeys Lambert-Beer Law. The Law relates the absorption of light to the properties of the material through which the light is traveling. The bigger the concentration of the colored solution, or the thicker the transmitted solution, or the greater the intensity of the incident light is, the higher the intensity of the absorbed light is. Assume that monochromatic light source irradiates a colored solution, the intensity of the transmitted light will weaken because the solution absorbs part of the intensity of the light, thus

$$A = K \cdot C \cdot L \tag{1}$$

where A is the absorbance, K is the absorptivity, C is the concentration of the colored solution, L is the thickness of the solution. Equation 1 is the mathematical expression for the law of the light absorption, which is called Lambert-Beer Law. The absorptivity $K = A/(C \cdot L)$ is the absorbance of the colored solution when the concentration and the thickness are per unit. The bigger K is, the higher the sensitivity of the colorimetric analysis is when the wave length of the incident light, the type of the solution and temperature are certain. Assume the absorbance of some kind of reference substance is A_1, the concentration of the colored solution is C_1, and the thickness of the solution is L_1. In the same situation, if we know the absorbance of the reference substance is A_2, and the

thickness of the solution is L_1. According to lambert beer's law, it will be

$$A_1 = K \cdot C_1 \cdot L \quad (2)$$

$$A_2 = K \cdot C_2 \cdot L \quad (3)$$

Thus the unknown concentration of the colored solution is

$$C_2 = \frac{A_2 C_1}{A_1} \quad (4)$$

When the source of the light is certain, according to the Lambert-Beer Law, the concentration of the colored solution can be calculated if we measure the intensity of the monochromatic absorption light.

B. *The Traditional Food Safety Detection*

Different substances have their specific absorption spectra, and the corresponding wavelength of the maximum value for the absorption of light is called the maximum absorption wavelength. In the experiments of food additive testing, generally the monochromatic light, which has the maximum absorption wavelength from the target detection material, is used as the measuring beam. The traditional photoelectric color comparator needs a monochromator which can turn the compound light into a specific wavelength monochromatic light. The traditional food safety detection has a monochromator [8], and the structure is shown in Fig. 1.

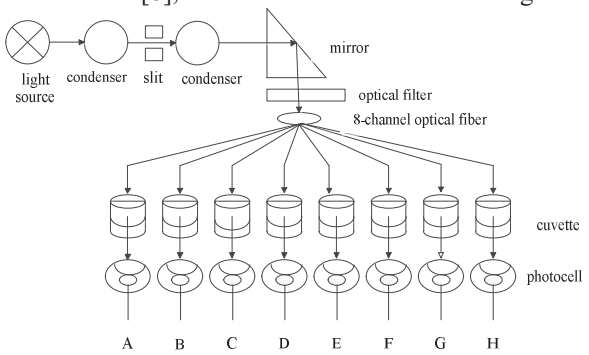


Fig. 1 The structure of the traditional food safety detection.

In the traditional food safety detection, the compound ligh from source is divided into a monochromatic light by the monochromator. The monochromator is usually realized with a grating or a narrow-band filter, which is the most complicated part in the photoelectric colorimetric unit. Therefore, it needs a mechanical structure to switch the grating or narrow band filter to generate different wavelengths monochromatic light. The traditional food safety detection has many short-coming such as the complex light path system, low detection efficiency and low detection precision.

C. *The Technical Scheme of Food Safety Detection*

In order to decrease the complexity of light path system, a new rapid screening of food safety system was presented. The new method is based on the rear spectroscopic technology which could simplify the traditional optical path system. A technical scheme of the food safety detection is presented in Fig. 2.

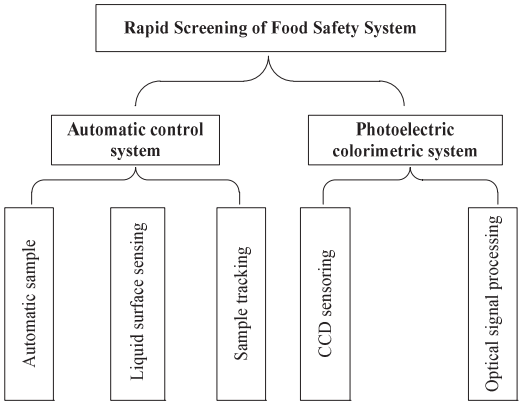


Fig. 2 The technical scheme of the food safety detection.

Fig. 2 illustrates the rapid screening of food safety system has two segments: automatic control system and photoelectric colorimetric system. Automatic control system accomplishes the automatic processing of sample, colorimetric, collection and data processing. Automatic addition of samples and reagents are accomplished by the automatic mechanical arm. In the processing of sample, the detector will automatically control mechanical arm to adjust the sample and reagent into the reaction cup according to the timed sequence and sample size in the system. By means of automatic cleaning the sample probe, liquid surface sensing and tracking technology, the contact area between the surface of the sample probe and the liquid is minimized, which can reduce cross contamination. Photoelectric colorimetric system adopts the rear spectroscopic technology which can simplify the traditional optical path system and reduce the volume and weight of the detector.

III. Design And Realization Of The Subsystem

A. *The Photoelectric Colorimetric System*

In order to decrease the complexity of light path system, a new CCD sensors is present which is based on the rear spectroscopic technology. The scheme of the photoelectric colorimetric system is shown in Fig. 3.

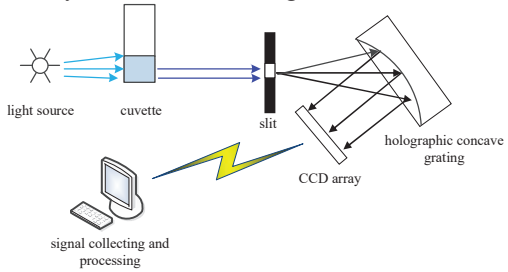


Fig. 3 The photoelectric colorimetric system

We present flat field holographic concave grating to achieve spectrometer and reduce the wavelength error. Meanwhile, the equipment adopts a new Charge-coupled Device (CCD) sensors which is replaced the traditional

optical path generation and detection system, the photoelectric colorimetric system. CCD sensors can obtain the full band spectral signal in the visible range and the whole band was detected simultaneously. Fig. 3 illustrates the light from source is sent off in different directions, then through an unscattered solution in the cuvette. Some of the light is absorbed by the solution, and some of light will pass the solution and reach the flat field holographic concave grating. After splitting by the grating, many monochromatic lights shoot CCD detector. The strong or weak lights will be converted into a current. The higher the current is, the greater the intensity of the light is.

B. The Automatic Control System

The control part refers motion control of the whole instrument. According to the function in Fig. 2, it is divided into four modules: sample module, reagent module, stirring module and cleaning module. Every module has different function, but all the structures are similar. All four modules need its' motors to drive the component in the module, so the hardware design of these modules is similar. In this paper, the design of hardware is illustrated by taking the sample module as an example. The sample module is divided into control layer and motion execution layer, and its hardware structure diagram is shown in Fig.4.

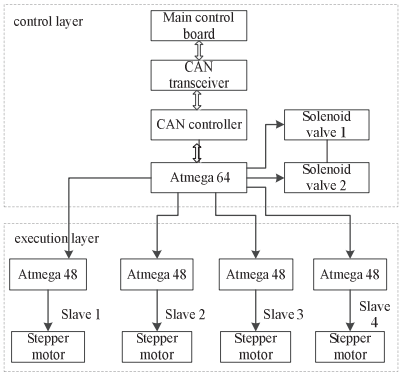


Figure 4 The hardware structure diagram of the sample module

The control layer is composed of ATmega 64 single chip microcomputer and its peripheral circuit. The motion execution layer is composed of the minimum system of the ATmega 48 and a driving stepper motor. The slave 1 controls the stepper motor which is used to drive the specimen tray rotating. The slave 2 controls the stepper motor which is used to drive the sample arm rising and falling. The slave 3 controls the stepper motor which is used to drive the sample arm swing. The slave 4 controls the stepper motor which is used to drive the pump sucking.

C. Liquid Surface Sensing and Tracking Technology

The sample module is the source of the screening equipment of food safety, which is the key of measurement and analysis. There are many ways to improve the detection speed and accuracy, but one of the most important methods is to reduce the amount of liquid on the surface of the probe. At present the most common method is to adopt liquid surface sensing and tracking technology, which can not only control the depth of the probe into the liquid but judge the liquid volume. The liquid surface sensing can avoid the detector virtual sampling to false detection. In this paper, the method is presented by the capacitance testing based on single – needle contact. The method can reduce the chance of sample contamination.

IV. System Debugging And Experimental Analysis.

To verify the function and performance of the new screening equipment of food safety, several commonly used synthetic pigment solutions were tested. We prepared a series of standard solutions of lemon yellow and carmine pigment. Firstly, a 0.25g lemon yellow pigment sample, and amount of water were placed in a 1000mL volumetric flask. A 10mL solution was extracted in another 500mL volumetric flask after shook well. Then added ammonium acetate solution in the 500mL volumetric flask and shook well, which was the different standard concentrations of lemon yellow pigment solution. Similarly, the different standard concentrations of carmine pigment solution were got. Absorption spectrum from 300 nm to 650 nm of the solution was measured by linear array CCD, then determined the maximum absorption wavelength, compared the results with the value from the calibration of the spectrophotometer. Based on the principle of additive property [10], we tested orange flavored drink (it contains lemon yellow and carmine pigments) by the new detector. We heated the drink to remove the CO_2 after it was aerated. Then took an appropriate amount solution and transferred into a 10mL cuvette. Absorption spectrum from 380 nm to 650 nm of the solution was measured by linear array the new food safety detector. The absorption spectrum from orange flavored drink was shown in Fig. 5.

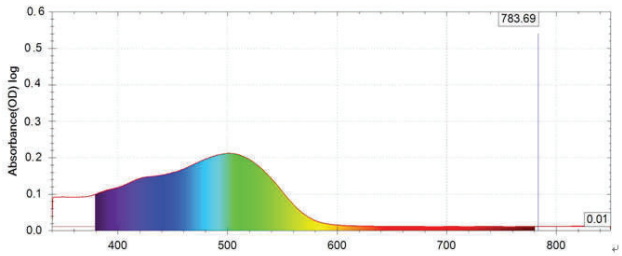


Fig.5 Absorptions of the full spectrum from orange flavored drink

From Fig. 5, the absorbance at 426 nm and 507nm were measured. Thus, we can get the contents of lemon yellow pigment and carmine pigment in the drink by solving the simultaneous equation. The results were shown in Table 1.

Table 1 The concentration of two pigment in the drink

No.	Absorbance (426 nm)	Absorbance (507 nm)	Concentration (lemon yellow)	Concentration (carmine)
1	0.186	0.145	4.38	3.79
2	0.185	0.147	4.32	3.90
3	0.186	0.148	4.34	3.87
4	0.187	0.152	4.33	4.00
5	0.184	0.149	4.25	3.96
Aver.			4.32	3.90
RSD (%)			1.10	2.09

As shown in Table 1, the concentration of lemon yellow pigment in drink was 4.32×10^{-3} mg/mL, and carmine pigment in drink was 3.90×10^{-3} mg/mL. The contents of lemon yellow pigment and carmine pigment didn't exceeds *food additive usage sanitation standard* (GB2760). Therefore, it is thought that the food additive used was safe. The method of the new detector was effective and accurate.

Stability is one of the key indicators of instrument. In general, if the instrument has been put into operation for half a year, its stability index still conforms to the specified standard, indicating that the instrument has a good stability. We prepared a lemon yellow solution, and assuming the nominal value of the absorbance at 426 nm is 1.000. Measuring the absorption of the lemon yellow solution by the new screening equipment every 5 minutes, the results were shown in Table 2. We prepared a carmine pigment solution by the same steps. and assuming the nominal value of the absorbance at 507 nm is 1.000. Measuring the absorption of the carmine pigment solution by the new screening equipment every 5 minutes, the results were shown in Table 2.

Table 2 The stability of the screening equipment

Solution	Wave-length	Initial value	5min	10min	20min	Stability
lemon yellow	426	0.994	0.993	0.996	0.996	0.002
carmine	507	0.992	0.995	0.991	0.996	0.004

The stability of the two tests is calculated, and the value is the maximum deviation value between the current and the initial value. As shown in Table 2, The stability of the instrument has met the industry standard, which are within the range of ±0.010A. Assuming the nominal value of the two different absorbance is 0.500. Repeat 4 times measurement and record the value of the instrument, and the results of the relative standard deviation as shown in table 3. The RSD value is usually used to indicate the absorbance repeatability of the instrument, and the range should be less than 0.5%.

Table 3 Repeated measurement of the instrument absorbance

Solution	Wave-length	1	2	3	4	RSD (%)
lemon yellow	426	0.499	0.495	0.497	0.4996	0.30
carmine	507	0.503	0.501	0.504	0.506	0.25

V. CONCLUSION

In this paper, a new rapid detection equipment of food safety based on rear spectroscopic technology and CCD was provided. We redesigned the photoelectric colorimetric system. The optical path system was simplified by the rear spectroscopic technology, thus the new equipment reduced the volume and weight of the instrument than traditional food safety detection. The flat field holographic concave grating and linear array CCD can collect the plane full spectrum signal at one time, so the new equipment can apply to multicomponent pigment measurement. After repeated practice and verification, the new equipment can reach up to 350nm-1100nm and meet the requirements of food additives measurement.

Acknowledgment

The authors gratefully acknowledge the financial support from the Prospective Joint Research Project of Jiangsu Province (BY2016050-02), the project of talent peak of six industries of Jiangsu province (2017-XYDXX-105), Jiangsu Laboratory of Lake Environment Remote Sensing Technologies Open Project Fund (JSLERS-2017-004), and the Project of Patent Navigation in Suzhou, the Science and Technology Development Plan Project of Changshu (CR201711).

References

[1] Lee K M, Runyon M, Herrman T J, et al. Review of Salmonella, detection and identification methods: Aspects of rapid emergency response and food safety[J]. Food Control, 2015, Vol. 47, pp.264-276.

[2] Tao C, Li G. A rapid one-step immunochromatographic test strip for rabies detection using canine serum samples[J]. Letters in Applied Microbiology, 2014, Vol. 59, No. 2, pp. 247-251.

[3] Z Yan, L Zhou, Y Zhao,et al. Rapid quantitativdetection of Yersinia pestis by lateral-flow immunoassay and up-converting phosphor technology-based biosensor [J]. Sens. Actuators B Chem. 2006, No. 119, pp. 656–663.

[4] Friedman M. Anticarcinogenic, cardioprotective, and other health benefits of tomato compounds lycopene, α-tomatine, and tomatidine in pure form and in fresh and processed tomatoes.[J]. Journal of Agricultural & Food Chemistry, 2013, Vol. 61, No. 40, pp. 9534-50.

[5] Li Z, Zhu Y, Zhang W, et al. A Low-Cost and High Sensitive Paper-Based Microfluidic Device for Rapid Detection of Glucose in Fruit [J]. Food Analytical Methods, 2017, Vol. 10, No. 3, pp.666-674.

[6] Kohl S K. Demonstration of Absorbance Using Digital Color Image Analysis and Colored Solutions [J]. Journal of Chemical Education, 2006, Vol. 83, No. 4, pp. 644-646.

[7] Visconti P, Lay-Ekuakille A, Primiceri P, et al. Hardware Design and Software Development for a White LED-Based Experimental Spectrophotometer Managed by a PIC-Based Control System [J]. IEEE Sensors Journal, 2017, Vol. 17, No. 8, pp. 2507-2515.

[8] Esperanza M, Cid Á, Herrero C, et al. Acute effects of a prooxidant herbicide on the microalga Chlamydomonas reinhardtii: Screening cytotoxicity and genotoxicity endpoints[J]. Aquatic Toxicology, 2015, No.165, pp.210-221.

[9] Zhou Q, Li X, Ni K, et al. Holographic fabrication of large-constant concave gratings for wide-range flat-field spectrometers with the addition of a concave lens [J]. Optics Express, 2016, Vol. 24, No. 2, pp. 732-738.

[10] Zhang Y J, Ma X, Zhang Y H, et al. The application of near-infrared diffuse reflection spectra based on the principle of linear additive in tobacco redrying formula [J]. Spectroscopy & Spectral Analysis, 2011, Vol. 31, No. 2, pp. 390.

Application of machine learning and time-series analysis for air pollution prediction

Vladimir Stojov *, Nikola Koteli *, Petre Lameski * Eftim Zdravevski*

* Faculty of Computer Science and Engineering
Sts. Cyril and Methodius University, Skopje, Macedonia
E-mails: vladimir.stojov@gmail.com,
{nikola.koteli,petre.lameski,eftim.zdravevski}@finki.ukim.mk

***Abstract*—Medical research studies show that low air quality can have a direct effect on the increased number of diseases, especially respiratory defects, but also on the increased mortality rate in people. Luckily, harmful particles and substances in the air can easily be detected and measured by using affordable sensors. The number of this type of sensors deployed in the city of Skopje, Macedonia continuously grows. The increased coverage of monitored regions, and the elevated public interest in solving this problem for obvious reasons, make the prediction of high levels of air pollution extremely beneficial. According to the available historical data, the problem of low air quality is proving to be more serious during the winter, that is during the heating season. If weather forecast is available, there is an opportunity to predict the air quality. This work reviews recent advances in air quality predictions using time-series analysis techniques, machine learning and deep learning. We proposes and evaluate two approaches for air quality prediction: combination of LSTM and convolutional neural networks and one-dimensional convolutional neural networks. The results show a promising accuracy of about 78% in predicting the level of air pollution.**

***Index Terms*—air pollution, prediction systems, deep learning, time-series analysis**

I. Introduction

After several key scientific man-made discoveries in the first half of the eighteenth century, a period of industrialization followed. Apart from the technological advances associated with this time, the period of the industrial revolution is also known as the first wave of global aero-pollution caused by man himself [1]. Except for the increased mortality rate to serve as an indicator, the instruments that were at mans disposal were not progressive enough to be able to determine the presence of pollutants, or measure their quantity in the air. Since then, the problem with pollution has only gotten worse.

After several key scientific man-made discoveries in the first half of the eighteenth century, a period of industrialization followed. Apart from the technological advances associated with this time, the period of the industrial revolution is also known as the first wave of global aero-pollution caused by man himself [1]. Except for the increased mortality rate to serve as an indicator, the instruments that were at mans disposal were not progressive enough to be able to determine the presence of pollutants, or measure their quantity in the air. Since then, the problem with pollution has only gotten worse.

This work was partially financed by the Faculty of Computer Science and Engineering at the Sts. Cyril and Methodius University, Skopje, Macedonia. We also acknowledge the support of Microsoft Azure for Research through a grant providing resources for this work.

As the human population rises, the need for food, clothing, medicine, as well as many other materials and goods consequently grows. This dependence has a direct impact over the reduced storage capacity of warehouses and garbage dumps. In order to cope with the increased demand for goods, the industrial facilities are always up and running. Factories, furnaces, mines, smelters, waste disposal units and similar plants that do not have appropriate filters installed, release a significant amount of particles and gases which may pose a threat to humans health. In addition, densely populated areas and cities are always accompanied by a vast number of vehicles, as well as a high percentage of housings that use non-ecological energetic resources as a fuel for heating. These also play a major role in the air pollution of the inhabited areas. Medical research studies show that low air quality can have a direct effect on the increased number of diseases, especially respiratory defects, but also over the increased mortality rate in humans [2], [3]. At the same time, this global issue indirectly impacts the economy in a negative manner [4].

The number of installed sensors able to detect harmful particles and substances continuously grows in the city of Skopje, Macedonia. Thanks to this increased coverage of monitored regions, as well as the few mobile and web applications which citizens use to keep up with the status of the air quality, the topic of air pollution has been drastically actualized in the last few years [5]. The geological characteristics of the city of Skopje, Macedonia, which is situated in a valley, influence the meteorological conditions in a way that they prevent the movement of air in periods with zephyr or no wind at all. According to the available historical data, the problem of low air quality is proving to be more serious during the winter, that is during the heating season [6]. Most of the sensors that are spread throughout Skopje, are capable of detecting the presence of PM10 and PM2.5 particles, as well as the presence of NO2, CO, O3, and SO2 gases. These types of particles and substances are most commonly present in the air, and are identified as damaging to human health.

It is a fact that the level of pollution depends on the present weather conditions that can be predicted using firmly established scientific methods that are constantly being improved by meteorologists and scientists of related scientific fields. Therefore, there is an opportunity to design and build an air quality prediction model, with the help of air quality historical data and the results from measurements of meteorological parameters and weather phenomena for the same time period. The meteorological parameters that are monitored and recorded by the National Hydrometeorological Service of Macedonia are: precipitation, air temperature, relative humidity, air pressure, direction and speed of wind. At the same time, changes in weather phenomena are also recorded, which may be: sunshine, cloudiness, fog, rain, hail or frost.

In addition to the dependency of air quality on weather conditions, another key characteristic is that all these data records are marked with a timestamp, or time interval. This greatly helps in data fusion of the two types of records. Furthermore, it can facilitate air quality prediction, provided that the weather forecast for the future period is known.

If the authorities in charge have relevant information of this type a few days in advance, they would have enough time to plan and enforce appropriate actions. This plan may include increased level of control over the industrial capacities or even a temporary work halt, more frequent public transport timetables, or timely alert so that the citizens can be prepared accordingly for the upcoming period [7]. Such measures would have a significant effect on those groups of citizens who are most affected by the decline in air quality, such as infants, senior citizens and those people with chronic respiratory diseases. Furthermore, the social network impact on the involvement of the authorities in charge is quite significant [8] and for the case of Skopje is one of the main channels for increasing awareness of the problem.

One of the goals of this research is to analyze several existing prediction models and to compare the level of success of their application in prediction of data that belongs to the domain being studied, such as seismic events prediction [9], [10]. Another aim is to build an architecture model that incorporates several already established prediction models, which should be able to make accurate air quality predictions based on the corresponding meteorological input data.

This rest of this paper is organized as follows: section II provides an introduction to prediction systems for data series and reviews other relevant approaches. Next, section III describes the methods including the analyzed architecture and section IV describes obtained results. Finally, section V concludes the paper and provides directions for future research.

II. Prediction Systems for Data Series

In some of the research studies done so far, analyzes on similar fields have been conducted already [6], [11]–[14]. Namely, due to the complex nature of the relationship between the meteorological and air quality data, the nonlinearity of the modeling approach is inevitable. Therefore, neural networks

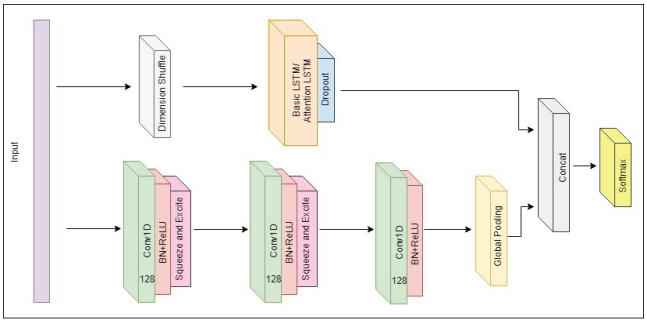


Fig. 1. MLSTM-FCN Architecture. Source: [14].

come in handy for solving this type of problem [6], [11]. The expectancy of the time-series data used in this survey is that the time interval between any two measurements is nearly a constant value [15], [16]. The task of the model will be to learn to detect similar patterns in the series of data, and to classify them, such as predicting dangerous methane concentration in mines [17]. The work done so far on similar topics involves prediction models based on LSTM neural networks, convolutional neural networks, deep learning algorithms, as well as other methods based on feature extraction and classical machine learning algorithms. The work presented in [18] proposes using histogram-based features calculated from the time-series. Such features are easily interpretable, computationally efficient, but also very robust and possibly useful for the prediction of air pollution. The methods presented in [19], [20] facilitate automation of the process of feature extraction and selection from arbitrary time-series data, and could be useful to come up with lightweight and powerful models with the least possible sensors.

One of the proposed models in [14] consists of a fully convolutional section constituted of temporal convolutional layers used as feature extractors, in pair with a LSTM section that process the multivariate time-series input, which initially is dimensionally adjusted to enhance performance. The graphic representation of the model is shown in Fig. 1.

III. Methods

A. Air Pollution Prediction Architectures

This research contains an analysis of the performance and effectiveness of the Long Short-Term Memory (LSTM) neural networks, as a recurrent neural network type suitable for solving this type of prediction problems [21], [22]. The difference between recurrent neural networks and regular feed-forward networks is the concept of time. This concept is introduced by feeding the output of a hidden layer back into itself. The problem with basic recurrent neural networks is that back-propagation gradients for maintaining long-distance connections tend to either vanish or accumulate and explode. Owing to the properties of the architecture, LSTM networks tackle this problem, which causes dissipation of the sensitivity of older input data [23], [24]. The main building block of a LSTM network is the LSTM cell, see Fig. 2. The cell

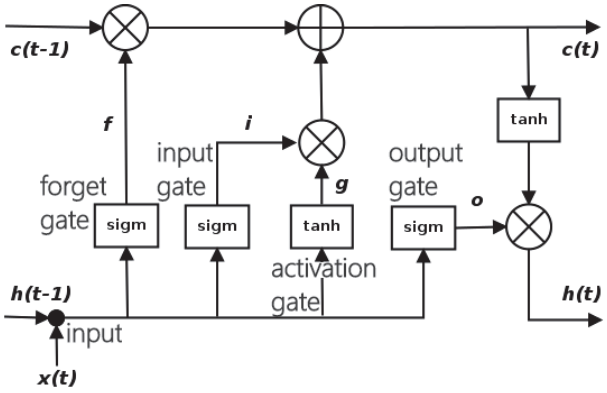


Fig. 2. A single LSTM cell.

maintains an internal memory state over time, backed up by non-linear gates that control the data flow in and out of the cell [25].

This structure facilitates the diminishing effect of the multiplication of tiny gradient values. This is done by first, squashing the input value with a tanh activation function, see 1. The U^g is the weight for the input, V^g is the weight of the previous cell output, and b^g is the input bias. The element x_t is the actual input, and h_{t-1} is the output of the hidden layer.

$$g = tanh(b^g + x_t U^g + h_{t1} V^g) \tag{1}$$

On the new value, element-wise multiplication by the output of the input gate is performed. This acts as an input filter, see 2.

$$i = \sigma(b^i + x_t U^i + h_{t1} V^i) \tag{2}$$

Another specific mechanism that's part of this cell, i.e. the forget gate is responsible for regulating which state is to be forgotten, or memorized, see 3.

$$f = \sigma(b^f + x_t U^f + h_{t1} V^f) \tag{3}$$

The internal state of the cell is named as c_t, see Fig. 4 and it is used to provide a recurrence loop for learning dependencies between time-separated inputs. The output of the forget gate actually determines which previous states should be remembered based on c_{t-1}.

$$c_t = c_{t-1} \circ f + g \circ i \tag{4}$$

The final step of this cell is the output gate, which is expressed in the first part of 5, 6, where the final output value is the second part.

$$o = \sigma(b^o + x_t U^o + h_{t-1} V^o) \tag{5}$$

$$ht = tanh(c_t) \circ o \tag{6}$$

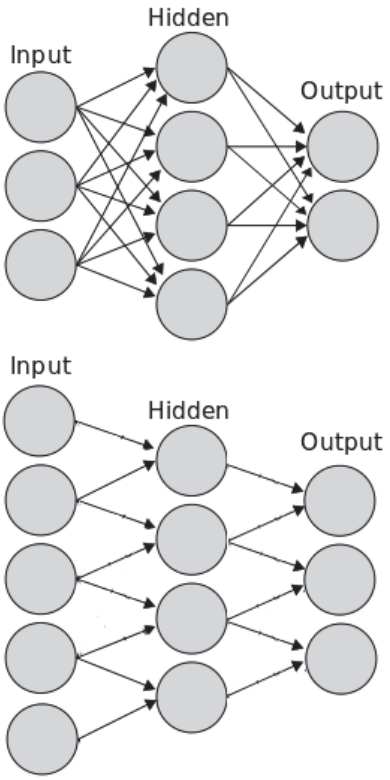


Fig. 3. Fully connected neural network vs. convolutional neural network with filter size [1,2].

In addition to this network type, we also examine the applicability of convolutional neural networks in the aforementioned domain. This type of neural networks have so far proven to be suitable for dealing with classification tasks which most often involve image recognition. However, this neural network model has also made a breakthrough in solving time-series classification or forecasting problems gross2017predicting, wang2017time. A convolutional neural network is comprised of sequential convolutional layers. Each of these layers is associated only to a single, sub-region of the input, i.e. it represents a convolution between the input and a sliding filter at a certain point, see Fig. 3. The filter is a weight matrix. The core idea behind employing this neural network architecture relies on the capability of the base model to learn filters that are adept in detecting specific patterns present in the input. Consequently, these filters can be used in forecasting future values. Standard convolutional networks contain an activation layer, which is useful for transforming the input into a non-linear value, which allows for learning more complex models. In this paper, one of the activation functions we will use is the sigmoid function, see Fig. 4. An interesting feature of these neural networks is that they are capable to process raw sensor information while generating structured data that may contain the key domain specific properties, and can be used in the process of training the model [26], [27]. This is a result of a structural characteristic, i.e separate channels, one for each

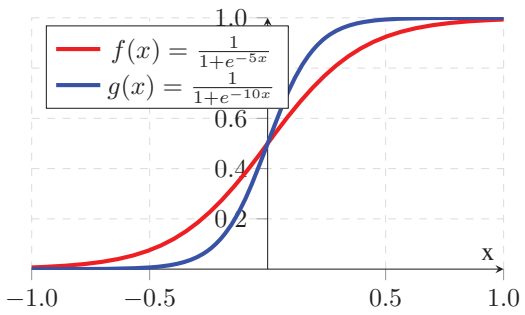


Fig. 4. Sigmoid Activation Function.

key feature. Another positive trait is the possibility to process complex multivariate data [28].

The aim of the research is to achieve a stable model that will reliably generate predictions for air quality based on weather forecast information. In this paper, several approaches for modeling an air quality prediction system will be presented.

B. Evaluated models

The first neural network model that was evaluated is a general model for classification. The results from this evaluation can used for benchmarking and comparison of future models. The key notion supporting this approach is tightly coupled with the nature of convolutional neural networks. Namely, it would be interesting to analyze the applicability of 2-dimensional convolutional layering in the domain of air quality data classification.

The measurements records are scarce at some points, meaning that the distance between two subsequent measurements may be larger than the usual, but the data shows that every measurement contains measurement entries for at least eight continuous hours. This circumstance can be exploited in a constructive manner, i.e. since every eight hour data sequence can be perceived as an image, thus can be used for classification during training. Each image represents a table, or a matrix with 8 rows, and 16 columns of data. Four of these columns contain pollution measurement values, which can be used to produce label sequences required for the supervised fragment of the training task. Every image can be associated with 4 different labels. Each label belongs to one of the four pollution-data columns and is calculated by using a simple categorical activation function, the output of which signals the level of pollution. The model contains two convolutional layers, each characterized by feature maps, and down-sampling sub-layer, see Fig. 5.

C. Data Description

The dataset consists of pollution and meteorological data from the area around and in Skopje, Republic of Macedonia. We took into consideration all of the meteorological and pollution measurement stations. In order to generate data without missing values, we added only the sensors that had

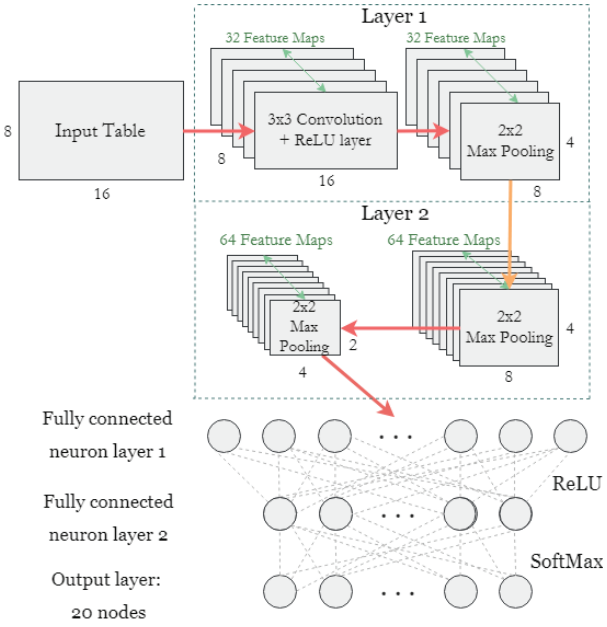


Fig. 5. Neural Network model composed of 2 Convolutional Layers, each with feature maps and sub-sampling.

enough samples measured in continuous intervals of 9 hours. By doing this we obtained data from 21 sensors. Each sample of the data contains measurements for 8 continuous hours and the ninth hour measurements are taken as labels.

D. Prediction Approach

One of the planned approaches involves combination of LSTM and Convolutional neural networks. The main idea behind applying convolutional neural networks is that multivariate time-series can can be viewed as a sequence of space-time images, which is an area where these networks excel. The STaR architecture proposed in [29] builds on this idea as shown in Fig. 10. In the STaR network, a hybrid network model is formed by combining RNN and CNN, streaming a copy of the input to each and concatenating the outputs in the end. Different filters are used on the same input in order to extract and learn different feature representations [29]. In contrast, the effectiveness of the sole LSTM based model shall also be analyzed, in terms of prediction of future values based on the previous N sequential measurement records.

Another method that will be examined is the appliance of one-dimensional convolutional neural networks. The expectation is that this network type is capable of learning how to predict future values based on meteorological and air quality time-series data. The reason for this expectancy is a result of the nature of the strong one-dimensional structure of time-series, which implies the high degree of correlation between spatially nearby variables, and this can be used to extract local features [30].

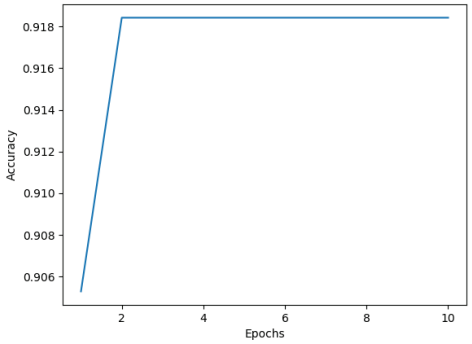


Fig. 6. Convolutional Neural Network model - 10 class labels - Results.

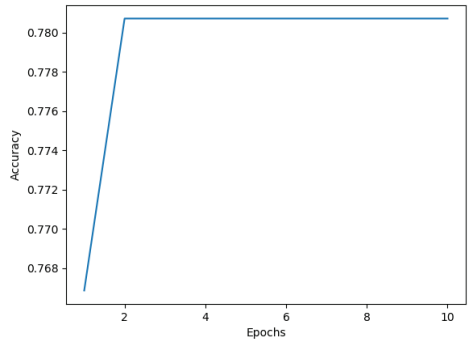


Fig. 7. Convolutional Neural Network model - 20 class labels - Results.

IV. Results

The analysis performed after the initial experiments, where the neural network model described in Section **??** is employed, yields somewhat promising results. The accuracy of the current model is highly dependent on the size of the class set that the pollution values belong to. The results from the first experiment with 20 classes are displayed on Fig. 7. The accuracy is around 78%, while on Fig. 6, a model with 10 classes shows increased accuracy of around 92%. In both cases, each class is a discrete interval of the air pollution. However, the dataset is highly imbalanced, i.e. a few classes are represented by a great number of examples, while the rest are represented by a few. To be more specific, in the scenario where 10 classes are used, the majority of examples (65% and 19%) belong to only 2 classes, and on the other hand the rest examples (16% of the total number) belong to 8 classes, see Fig. 8. In the other scenario, with 20 classes, the distribution is still imbalanced, where the most represented classes contribute with around 80% of the total number of examples, and the rest 17 classes are only represented by a 20% of the entries, see Fig. 9. This leads to the conclusion that the general classification accuracy is not the perfect fit for measuring the effectiveness of the model.

V. Conclusion and Future Work

In this work, we have presented recent advances in air quality predictions using time-series analysis techniques, machine learning and deep learning. We proposed a new architecture model for prediction based on the existing discussed frameworks.

The number of air-quality sensors deployed in the urban areas continuously grows. Combining the data from these sensors, with the weather forecast data, provides an opportunity to predict the air quality.

The paper proposes two approaches for air quality prediction: combination of LSTM and convolutional neural networks and one-dimensional convolutional neural networks.

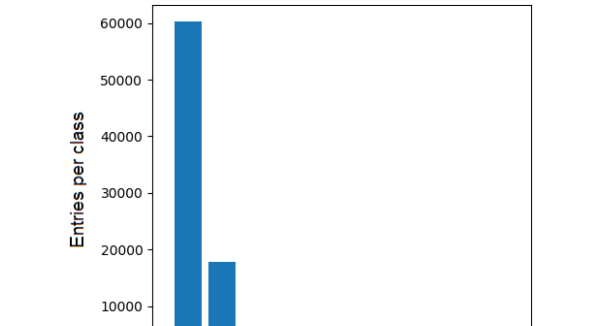


Fig. 8. Measurement entry value distribution by class, with a set of 10 classes.

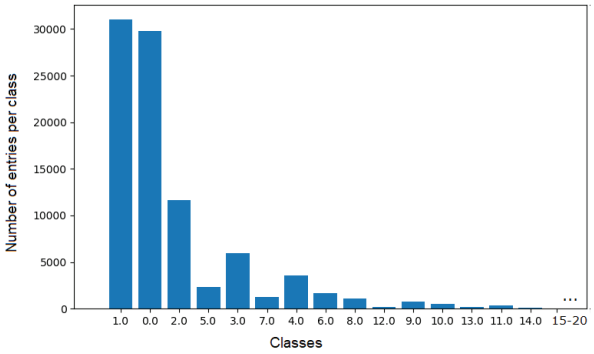


Fig. 9. Measurement entry value distribution by class, with a set of 20 classes.

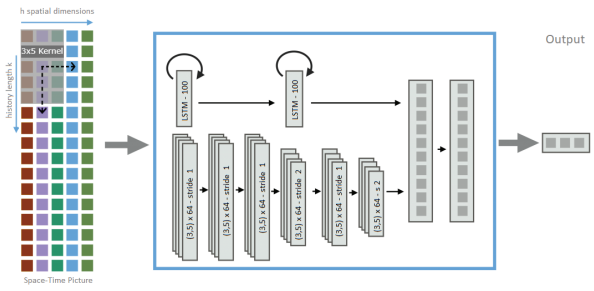


Fig. 10. The Space-Time Convolutional and Recurrent Neural Network (STaR) architecture. Related time-series are arranged in a space-time picture and fed to the input layer of STaR. Source: [29]].

The main idea behind applying convolutional neural networks is that multivariate time-series can can be viewed as a sequence of space-time images, which is an area where these networks excel. In LSTM based approach, LSTM model can be used for prediction of future values based on the previous N sequential measurement records. In one-dimensional convolutional neural networks approach, we rely on the strong one-dimensional structure of time-series, which implies the high correlation of spatially nearby variables, and this can be used to extract local features needed for prediction.

Our initial experiments show that both approaches produce promising results.

REFERENCES

[1] J.-P. Candelone, S. Hong, C. Pellone, and C. F. Boutron, "Post-industrial revolution changes in large-scale atmospheric pollution of the northern hemisphere by heavy metals as documented in central greenland snow and ice," *Journal of Geophysical Research: Atmospheres*, vol. 100, no. D8, pp. 16 605–16 616, 1995. [Online]. Available: http://dx.doi.org/10.1029/95JD00989

[2] R. Wilson and J. Spengler, *Particles in Our Air: Concentrations and Health Effects*, ser. Department of Physics Series. Harvard School of Public Health, 1996. [Online]. Available: https://books.google.mk/books?id=XksfAQAAIAAJ

[3] D. V. Bates, "Health indices of the adverse effects of air pollution: The question of coherence," *Environmental Research*, vol. 59, no. 2, pp. 336 – 349, 1992. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0013935105800404

[4] N. Knzli, R. Kaiser, S. Medina, M. Studnicka, O. Chanel, P. Filliger, M. Herry, F. Horak, V. Puybonnieux-Texier, P. Qunel, J. Schneider, R. Seethaler, J.-C. Vergnaud, and H. Sommer, "Public-health impact of outdoor and traffic-related air pollution: a european assessment," *The Lancet*, vol. 356, no. 9232, pp. 795 – 801, 2000. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0140673600026532

[5] K. Mitreski, M. Toceva, N. Koteli, and L. Karajanovski, "Air quality pollution from traffic and point sources in skopje assessed with different air pollution models," *Journal of Environmental Protection and Ecology*, vol. 17, no. 3, pp. 840–850, 2016.

[6] T. Stafilov, R. Bojkovska, and M. Hirao, "Air pollution monitoring system in the republic of macedonia," *Journal of Environment and Protection Ecology*, vol. 4, pp. 518–524, 2003.

[7] G. Corani, "Air quality prediction in milan: feed-forward neural networks, pruned neural networks and lazy learning," *Ecological Modelling*, vol. 185, no. 2, pp. 513 – 529, 2005. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0304380005000165

[8] K. Budinoski and V. Trajkovik, "Incorporating social network services in egovernment solutions: A case study," *European Journal of ePractice*, vol. 16, pp. 58–70, 2012.

[9] A. Janusz, M. Grzegorowski, M. Michalak, ukasz Wrbel, M. Sikora, and D. lzak, "Predicting seismic events in coal mines based on underground sensor measurements," *Engineering Applications of Artificial Intelligence*, vol. 64, pp. 83 – 94, 2017. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0952197617301215

[10] E. Zdravevski, P. Lameski, and A. Kulakov, "Automatic feature engineering for prediction of dangerous seismic activities in coal mines," in *2016 Federated Conference on Computer Science and Information Systems (FedCSIS)*, Sept 2016, pp. 245–248.

[11] M. Gardner and S. Dorling, "Artificial neural networks (the multilayer perceptron)a review of applications in the atmospheric sciences," *Atmospheric Environment*, vol. 32, no. 14, pp. 2627 – 2636, 1998. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1352231097004470

[12] M. Kolehmainen, H. Martikainen, and J. Ruuskanen, "Neural networks and periodic components used in air quality forecasting," *Atmospheric Environment*, vol. 35, no. 5, pp. 815 – 825, 2001. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S135223100000385X

[13] K. P. Singh, S. Gupta, A. Kumar, and S. P. Shukla, "Linear and nonlinear modeling approaches for urban air quality prediction," *Science of The Total Environment*, vol. 426, pp. 244 – 255, 2012. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0048969712004809

[14] F. Karim, S. Majumdar, H. Darabi, and S. Harford, "Multivariate lstm-fcns for time series classification," *arXiv preprint arXiv:1801.04503*, 2018.

[15] P. Esling and C. Agon, "Time-series data mining," *ACM Computing Surveys (CSUR)*, vol. 45, no. 1, p. 12, 2012.

[16] G. Bontempi, S. B. Taieb, and Y.-A. Le Borgne, "Machine learning strategies for time series forecasting," in *European Business Intelligence Summer School*. Springer, 2012, pp. 62–77.

[17] D. lzak, M. Grzegorowski, A. Janusz, M. Kozielski, S. H. Nguyen, M. Sikora, S. Stawicki, and ukasz Wrbel, "A framework for learning and embedding multi-sensor forecasting models into a decision support system: A case study of methane concentration in coal mines," *Information Sciences*, vol. 451-452, pp. 112 – 133, 2018. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0020025518302822

[18] E. Zdravevski, P. Lameski, R. Mingov, A. Kulakov, and D. Gjorgjevikj, "Robust histogram-based feature engineering of time series data," in *Computer Science and Information Systems (FedCSIS), 2015 Federated Conference on*, ser. Annals of Computer Science and Information Systems, M. P. M. Ganzha, L. Maciaszek, Ed., vol. 5. IEEE, Sept 2015, pp. 381–388. [Online]. Available: http://dx.doi.org/10.15439/2015F420

[19] E. Zdravevski, P. Lameski, V. Trajkovik, A. Kulakov, I. Chorbev, R. Goleva, N. Pombo, and N. Garcia, "Improving activity recognition accuracy in ambient-assisted living systems by automated feature engineering," *IEEE Access*, vol. 5, pp. 5262–5280, 2017.

[20] E. Zdravevski, B. Risteska Stojkoska, M. Standl, and H. Schulz, "Automatic machine-learning based identification of jogging periods from accelerometer measurements of adolescents under field conditions," *PLOS ONE*, vol. 12, no. 9, pp. 1–28, 09 2017. [Online]. Available: https://doi.org/10.1371/journal.pone.0184216

[21] Y. Bengio, P. Simard, and P. Frasconi, "Learning long-term dependencies with gradient descent is difficult," *IEEE transactions on neural networks*, vol. 5, no. 2, pp. 157–166, 1994.

[22] F. Karim, S. Majumdar, H. Darabi, and S. Chen, "Lstm fully convolutional networks for time series classification," *arXiv preprint arXiv:1709.05206*, 2017.

[23] J. C. B. Gamboa, "Deep learning for time-series analysis," *arXiv preprint arXiv:1701.01887*, 2017.

[24] K. Kawakami, "Supervised sequence labelling with recurrent neural networks," Ph.D. dissertation, Ph. D. thesis, Technical University of Munich, 2008.

[25] K. Greff, R. K. Srivastava, J. Koutník, B. R. Steunebrink, and J. Schmidhuber, "Lstm: A search space odyssey," *IEEE transactions on neural networks and learning systems*, vol. 28, no. 10, pp. 2222–2232, 2017.

[26] F. J. Ordóñez and D. Roggen, "Deep convolutional and lstm recurrent neural networks for multimodal wearable activity recognition," *Sensors*, vol. 16, no. 1, p. 115, 2016.

[27] P. Lameski, E. Zdravevski, V. Trajkovik, and A. Kulakov, "Weed detection dataset with rgb images taken under variable light conditions," in *ICT Innovations 2017*, D. Trajanov and V. Bakeva, Eds. Cham: Springer International Publishing, 2017, pp. 112–119.

[28] A. Borovykh, S. Bohte, and C. W. Oosterlee, "Conditional time series forecasting with convolutional neural networks," *arXiv preprint arXiv:1703.04691*, 2017.

[29] W. Groß, S. Lange, J. Bödecker, and M. Blum, "Predicting time series with space-time convolutional and recurrent neural networks," *Proc. of the 25th ESANN*, pp. 71–76, 2017.

[30] Y. LeCun, Y. Bengio *et al.*, "Convolutional networks for images, speech, and time series," *The handbook of brain theory and neural networks*, vol. 3361, no. 10, p. 1995, 1995.

Geo-reference Digitalization of Green Urban Spaces Using GIS: A Case Study of Skopje

Andreja Naumoski, Georgina Mirceva, Kosta Mitreski
Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje
Skopje, Republic of Macedonia
e-mail: {andreja.naumoski, georgina.mirceva, kosta.mitreski}@finki.ukim.mk

***Abstract* — Green urban places are important features of every city these days. Not only for recreation and for sport activities of the inhabitants of the city, but they serve as filters for today's growing problem of aero pollution. Therefore, maintaining and improving them, should be a priority of city planners. In many cases, city planners are using Geographic Information Systems (GIS) as powerful digitalization and planning tool. With GIS, the users can save, analyse and present geo-referenced data on any platform (desktop, mobile, cloud etc.). In that direction, the paper aims to use GIS in the process of green urban spaces digitalization in order to improve them and further become available to city planners for their use. As a case study, we have taken into account two Skopje city parks, by digitalizing the walking pathways as well as green spaces and park see-sites. The obtained geo-referenced map is the result that will provide a framework for further analysis and improvements of the green urban spaces. For future work, we plan to expand the analysis by combining outdoor activities with the digitalized green public map in real-time.**

Keywords — GIS, green urban spaces, data management, data analytics, digitalization.

I. INTRODUCTION

Geo-referenced data contains not only the futures needed for the users, but also these features contain a geographical component, namely; the location of that particular event happened. In today digitalized world, GIS has provided an excellent framework for storing, analysing, visualizing vast amount of time dependent data, as well as geo-referenced data. GIS package contains variety of practical tools for managing the data and the geographical maps. One of the most used GIS software is ArcGIS ESRI [1]. This software contains several major packages like: Arc Map – software application used for geo-referenced data handling and operations, map analytics and visualization; Arc Catalog – package for geo-reference data accommodation and administration; Arc Toolbox – various tools, scripts, geoprocessing, interoperability, geographical coordinate system manipulation and model building in a form of Arc Map extensions; Arc 3D Globe – tool for managing and visualizing 3D geo-reference data. Using these of the many software packages inside of the ArcGIS software, the user can collect, stock, manage and geographically integrate vast amount of various types of geo-referenced data from various sources in one place. Sounds like big data, which in fact it is. Bigdata analytics in GIS is one way to integrate various types of information on a geographical map. Combined with the multi-platform option that is available in GIS, the user can deep dive into the data, dig more integrated and general conclusions and share on various platforms. This make GIS ideal for presenting the results to a wider audience and for the city planners. Several research studies have revealed the importance of GIS in planning and analysing data regarding green urban spaces. These studies focus on the relationship between the green urban spaces and their accessibility to the city inhabitants [2], as well as entire distribution of the green urbans spaces within the cites [3]. Other researchers focus on the influence of the green urban spaces that have on their residence [4, 5]. Furthermore, researcher investigate not only the influence of the green urban spaces on the physiological state of the city inhabitants, but also from the epidemiological status [6], by decreasing the rate of disease spreading. More important, the influence of the urban green spaces on young children and their development have been research by [7], showing that green urban spaces have statically significant impact on children health that live no more than 500 meters from the nearest green urban space. Others have developed various indices to evaluate the relationship between the area of the urban green spaces and overall city building area [8]. This why having a GIS map of the green urban spaces is vital for future development of the city, mostly of the inhabitant's health.

The structure of the paper is organized as follows: Section 2 presents the procedure of the digitalization of two green urban spaces in city of Skopje the two final maps, while Section 3 concludes the paper and gives direction for future work.

II. METHODS AND MATERIALS

In order to make and use GIS potential, we need to acquire geographical map and geo-referenced data. From the many tools that GIS offers, we use the map downloader to download a map directly from Google Maps. Later this map, it is important to select the part of the map that shows the city of Skopje and its surrounding. Depending of the internet connection the map is downloaded and imported into GIS. The second part of our analysis requires geo-referenced data, which we obtained using smartphone in KML format [9]. KML is an XML notation for expressing annotations and visualisation within 2D and 3D maps, used in Google Earth software. This software was the first that was able to view and edit the KML format, later on other programs used KML format to develop their own maps and software support. This was done not for desktop, but also for mobile device, that we used for our project. We use the smartphone to track the pathway of a particular user or group, and then using the smartphone, variety

applications export the data in KML format and load into GIS. The next step of our analysis is to apply Analysis/Measurement tool to measure the pathways that are generated by the smartphone. After we complete this step we proceed on to the digitalization step. As we mentioned digitalization requires to use geodatabase with geo-referenced data to work. For this purpose, we use Arc Catalog tool, where we first create new geodatabase and we add the geo-reference data by creating feature classes. It is important to note that feature classes coordinate projection should be the same as the base map coordinate projection. For each of these feature classes we then create tables, as we need.

Now, we start the digitalization by marking the target area borders (Create Area/Polygon Features -> Area). Each geographical map in ArcMap consist from several layers that are linked with the geo-referenced data stored in the database. In order to start editing the map, we select "Start Editing" option in the ArcMap. With Create feature tool, we select the elements that are needed to be drawn on the map, and we can track them in the Table of Content. For more precise drawing on the map we select Show/Hide Area/Line Vertical tool in order to show the pathways that we gather with the smartphone app. The software package also has several preinstalled elements, like Lines, Polygons, Points and etc., so the user can select and immediately use them. More complex operations are also supported, when the user have a need to merge two or more objects (lines, polygons and etc.), he can use the Merge tool. The conceptual diagram of this process is represented with Fig. 1.

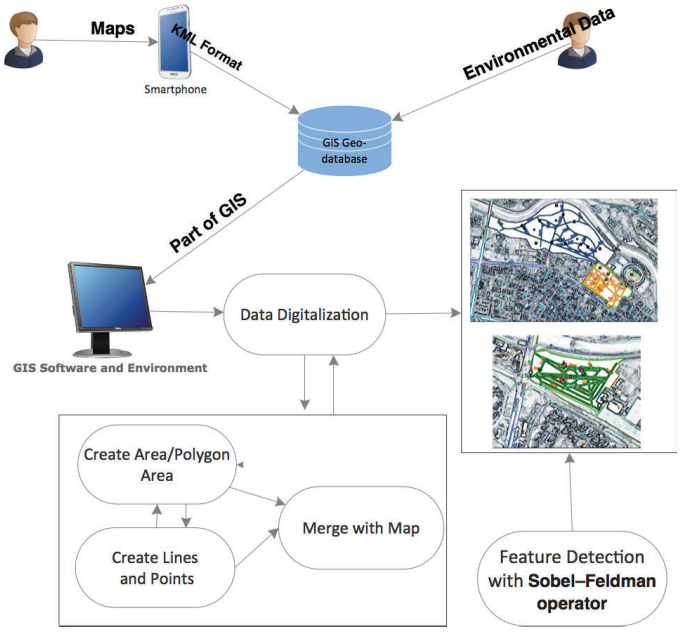


Fig. 1. Diagram flow of the digitalization and future detection process

Polygons are one of the most widely use objects on the GIS map. They can be reinserted from other maps, or the user can draw them on the map by connection several dots that we draw on the area of interest on the map. After the drawn polygon is placed on the map, then we can select the type of the feature on the map: For example: State Park. Later, we can change the colour, or we can some other features. By using various tools, we can add or delete features drawn on the map. For example, by using the tool for overlay control centre we can delete or add the feature that exist on the map. Other interesting features of every green urban space are recreational objects, like children playgrounds, pedestrian walkways, churches, small ponds, and bridges connecting them, sport stadiums, tennis or basketball playground and etc. This can be done by using the tool for creating Point/Text features. The final maps of one of the two city parks is shown in Fig. 2.

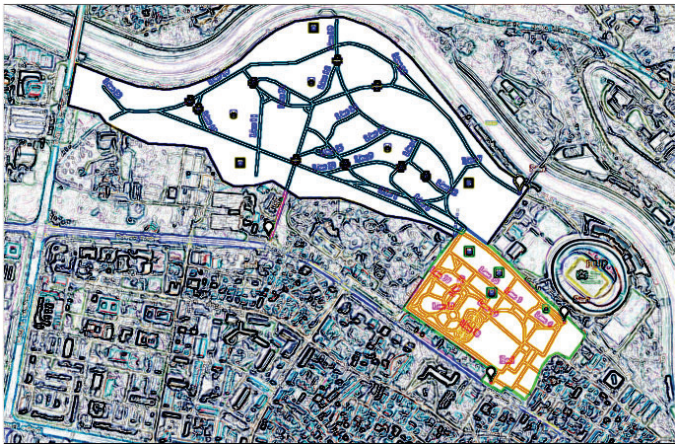


Fig. 2. Digitalized map and feature detection of the first Skopje green space

The yellow and the blue colours are representing the green urban space, while the lines that are drawn within are the pedestrian walkways. Additionally, we apply Sobel-Feldman operator [10] to detect features of the finished maps in order to speed up the process of identifying object on the map.

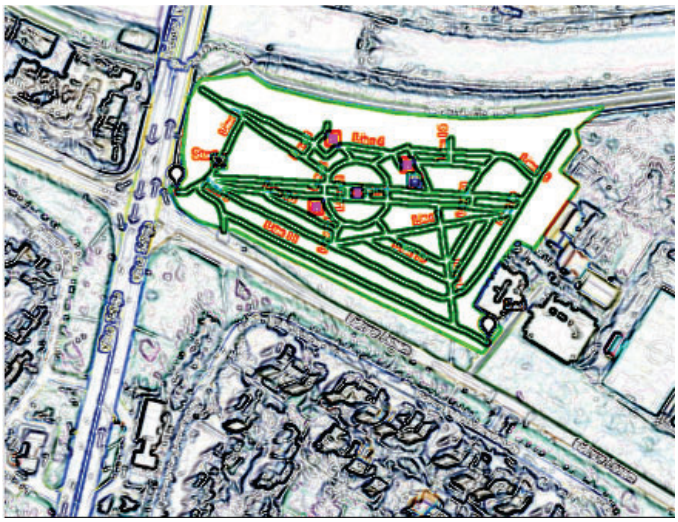


Fig. 3. Digitalized map and feature detection of the second Skopje green space

III. Conclusion

By digitalizing the geo-referenced green urban spaces in this paper, we provided two final maps that depicts the

pedestrian walkways as well as some points of interest. Additionally, in this paper we have presented the process of data and map collection, through the process of map making and digitalization of the geo-referenced data. Furthermore, we added some points of interest inside the green urban space, as well as places that are important for each visitor and later apply feature detection using Sobel-Feldman operator. In future, we plan to connect this geo-referenced map and all-important places with variety sports activates in real-time, as well as annotating the detected objects on the map.

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] ESRI. ArcGIS Desktop: Release 10.4. Redlands, CA: Environmental Systems Research Institute, 2017.

[2] M. Cetin, "Using GIS analysis to assess urban green space in terms of accessibility: case study in Kutahya," International Journal of Sustainable Development & World Ecology, vol. 22, no. 5, pp. 420-424, 2015.

[3] M. Wen, X. Zhang, C.D. Harris, J.B. Holt, and J.B. Croft, "Spatial disparities in the distribution of parks and green spaces in the USA," Annals of Behavioural Medicine, vol. 45, no. 1, pp. 18-27, 2013.

[4] J. Qin, X. Zhou, C. Sun, H. Leng, and Z. Lian, "Influence of green spaces on environmental satisfaction and physiological status of urban residents," Urban forestry & urban greening, vol. 12, no. 4, pp.490-497, 2013.

[5] A.C. Lee, and R. Maheswaran, "The health benefits of urban green spaces: a review of the evidence," Journal of public health, vol. 33, no. 2, pp. 212-222, 2011.

[6] M. van den Berg, W. Wendel-Vos, M. van Poppel, H. Kemper, W. van Mechelen, and J. Maas, "Health benefits of green spaces in the living environment: A systematic review of epidemiological studies," Urban Forestry & Urban Greening, vol. 14, no. 4, pp. 806-816, 2015.

[7] I. Markevych, C.M. Tiesler, E. Fuertes, M. Romanos, P. Dadvand, M.J. Nieuwenhuijsen, D. Berdel, S. Koletzko, and J. Heinrich, "Access to urban green spaces and behavioural problems in children: results from the GINIplus and LISAplus studies," Environment international, vol. 71, pp. 29-35, 2014.

[8] K. Gupta, P. Kumar, S.K. Pathan, and K.P. Sharma, "Urban Neighborhood Green Index–A measure of green spaces in urban areas," Landscape and Urban Planning, vol. 105, no. 3, pp. 325-335, 2012.

[9] "OGC KML 2.3 Standard". OGC. 2015.

[10] I. Sobel, History and Definition of the Sobel Operator. 2014.

EU Directive 2008/50/EC on Ambient Air Quality and Cleaner Air for Europe in Context of the Republic of Macedonia

Ilija Rumenov
University "Ss. Cyril and Methodius"- Faculty of Law "Iustinianus Primus" Skopje,R.Macedonia
i.rumenov@pf.ukim.edu.mk

Martina Toceva
Ministry of environment and physical planning of Republic of Macedonia(MoEPP)- Department of Macedonian Environmental Information Center Skopje,R.Macedonia
m.toceva@moepp.gov.mk

Kosta Mitreski
University "Ss. Cyril and Methodius"- Faculty of Computer Science and Engineering, Laboratory for Eco-informatics Skopje,R.Macedonia
kosta.mitreski@finki.ukim.mk

Abstract*— The EU Directive 2008/50/EC on ambient air quality and cleaner air for Europe provides for rules in order to reduce pollution to levels which minimize harmful effects on human health, paying particular attention to sensitive populations and to improve the monitoring and assessment of air quality including the deposition of pollutants and to provide information to the public. This article provides for analysis of the ambient air quality and measurement systems in the Republic of Macedonia in context of this EU Directive and it offers for examination of the Directive requirements in respect of air quality management area delineation - zones and agglomerations. The evaluation is made on the basis of the regular measurements obtained from the existing measuring networks as well as via measurements of the emissions from stationary and mobile sources in several years period. Results from the analysis of air pollution parameters SO_2, NO_2, NOx, PM_{10} and ozone concentrations are presented both for specific regions and on the entire state territory.

Keywords— *EU directives, air pollution, zones, models*

I. INTRODUCTION

Air is fundamental natural resource for sustenance of life and for other activities in the biosphere. Its paramount importance proposes significant protection as a natural resource and as a right. As a right, protection is given on universal and on regional level although, the importance of environment was neglected in the past because of the lack international concern for the global protection of environment [1]. However such position has been changed in recent years with the rise of the conscience for the environmental problems with protection offered by European Convention of Human Rights (ECHR) and in the EU [2]. In 2008 the EU has adopted a new Directive 2008/50/EC (so called CAFÉ directive) on ambient air quality and cleaner air for Europe [3] which repealed the old Council Directive 96/62/EC of 27 September 1996 on ambient air quality assessment and management [4] and all of other Directives (1999/30/EC, 2000/69/EC, 2002/3/EC) and also the Council Decision 97/101/EC of 27 January 1997[5]. With that the EU has replaced all of these five instruments with one in the interests of clarity, simplification and administrative efficiency. For achieving the goals in the CAFÉ directive, Member States have the following responsibilities: a) assessment of ambient air quality; (b) approval of measurement systems (methods, equipment, networks and laboratories); (c) ensuring the accuracy of measurements; (d) analysis of assessment methods; (e) coordination on their territory if Community-wide quality assurance programs are being organized by the Commission; (f) cooperation with the other Member States and the Commission. Also another obligation form this Directive is that the Member States should establish zones and agglomerations throughout their territory, upon which air quality assessment and air quality management should be carried.

The detailed limits for each substance of interest are set out in the Directive and specify the requirements for air quality assessment in each of the zones.

In this aspect, the monitoring of the situation in the Member States of the EU has shown for example, that total of 19 % of the EU-28 urban population was exposed to PM10 (Particular matter 10) levels above the daily limit value and approximately 53 % was exposed to concentrations exceeding the stricter WHO AQG (World Health Organization – Air quality guidelines) value for PM10 in 2015. This represents an increase compared with 2014, but the magnitude of the change may be considered as being within the expected year-to-year variability [6].

Republic of Macedonia faces tremendous problems in correlation to the health of the population arising out of the ambient air quality. Every year, the concentration of the pollutants exceeds the thresholds for the protection of the human health. The problem is evident in every part of the country, but it is mostly present in the urban environments such as Skopje and Tetovo [7]. Most problematic pollutants are the PM10 and NOx which seriously endanger the health of the population. The most relevant legal sources are the Ambient Air Quality Act [8] and the Decree on the thresholds and types of pollutants of the ambient air and thresholds for alarming, deadlines for achieving the thresholds, margins of tolerance of the thresholds, goals and long term goals [9].

This paper is part of the continuous work performed by the Laboratory for Eco-informatics at the Faculty of Computer Science and engineering at the University "Ss. Cyril and Methodius" Skopje, Republic of Macedonia and represents follow up on the new developments in the EU based on the adoption of the new EU directive[10]. The rest of the paper is organized as follows. In section 2 methods that are used for air quality methods are presented, while in Section 3 the results from the conducted measurements with analysis on the key components that affect this quality issue pointed by the expert group.

II. METHODS

The requirements laid down in the Directive for air quality assessment methods in each of the zones depend on how deep pollution levels in the zones fall below the limit values [14].

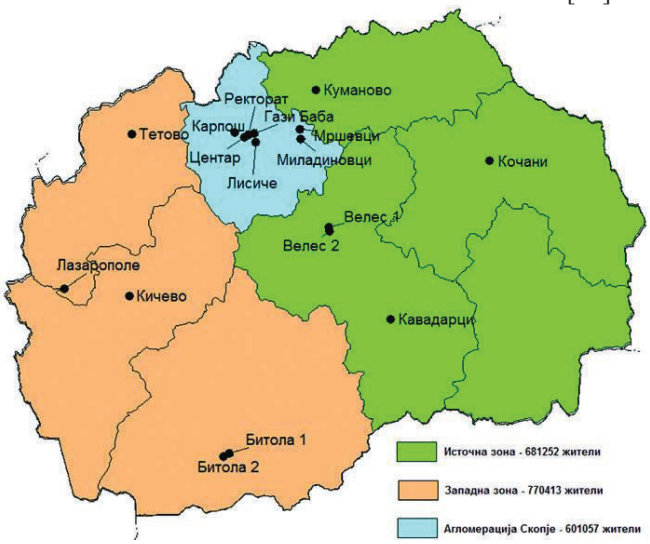


Fig.1. Zones and agglomeration in the Republic of Macedonia

The borders of administrative units may serve for delineating zones or combining administrative areas with similar air quality characteristics (Fig.1).

The ambient air quality directive does not stipulate measurements any longer as the only tool for determining levels in a zone, and envisages – depending on pollution levels – the use of modeling techniques and expert estimates and their combinations.

It is important to notice that in this context the distinction between measurement and other assessment methods (interpretation, spatial interpolation of measurements, modelling) is not as clear-cut as is often thought. The Guidance on Preliminary Assessment [11], suggests that three components be used as part of the assessment process;

- preliminary measurements(monitoring system),
- modeling, and
- air emission inventories.

A. Monitoring systems and methodologies applied in Macedonia

Assessment is done with all the data available from the automatic monitoring systems in the MoEPP measurement programs.

a) Automatic Air Monitoring System (MoEPP)

State automatic network for air quality monitoring consists of 18 stations that consistently measure air quality in different parts of the country. This number of stations meets the minimum number of monitoring stations on national level and is in consent with the requirements stated in the national legislation and European directives. In order to obtain credible results, equipment for monitoring must regularly be maintains and calibrate. Unfortunately, the poor regular maintenance of the instruments and deficiency of spare parts results with lower coverage with data especially in past a few years [7]. Laboratory for Eco-informatics used one mobile monitoring station (Air pointer-MLU-Austria) for this purposes. Our database system used all the data available for acquisition from the MoEPP and from our station for Air pollution assessment and modeling purposes.

b) Air pollution modeling

Air pollution dispersion models can be used for impact assessment of certain sources of emissions and categories the sources on the quality of air. They can be used to support the process decisions by providing information on the impact of measures to improve air quality and emission reductions, also to support traffic and urban planning. In order to achieve quality results, quality data from the meteorological observations and detailed emission information are needed. The weakness availability of quality input data limits usage of the dispersion models in the country. MoEPP used Local (UDM-FMI) and (CAR-FMI) and regional (SILAM) models developed by the Finnish Meteorological Institute for assessment of air quality. On the other hand Laboratory for Eco-informatics used Street canyon Operational Street Pollution Model (OSPM) and HYSPLIT (HYbrid Single-particle Lagrangian integrated trajectory) model. The idea in this article is to compare output results from different models and also to predict and improve air quality in R.Macedonia.

c) Air emission inventories

There are natural and anthropogenic sources of suspended particles in the atmosphere. Anthropogenic sources are combustion of fuels for production energy, incineration, heating in households and combustion of fuels from vehicles. Especially in cities, important local sources represent road traffic (vehicles and dust from the roads), as well and burning wood or coal for heating in households. Production energy and industrial emissions generates different sizes of suspended particles, depending on manufacturing process. The size of the suspended particles is very significant from a health aspect, since the finest particles get deeper in the human body and cause it more serious health impact.

III. RESULTS

The biggest individual sources of pollution are REK Bitola (located at Pelagonija Valley), FENI Industry (located at the most famous vinery area), Jugohrom ALZAR (Polog Valley) and MakSteel, ArcelorMittal and USJE (Skopje Valley).

Only REK BITOLA and TEC-Oslomej Thermal Power Stations would likely affect SO_2 levels in air significantly when judged against the directive. Simple calculation [12] shows that annual average values would probably be about 30 µg/m^3 or less. This seems to correspond with the slight elevation in SO_2 levels noted in measurements in Bitola when compared to other towns (Fig.2).

Similar simple calculation shows that TEC Oslomej, near Kichevo, would give rise to an annual average of about 25 µg/m^3 or less from a 180 m stack.

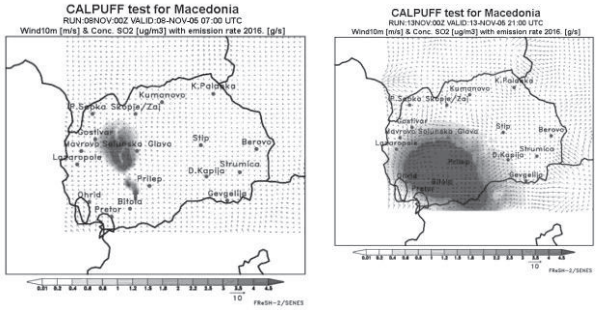


Fig 2. Emission of SO_2 according to CALPUFF model [13] and corresponding puffs before 10 years.

A. Exceedance of air quality limit values in regions of Macedonia (PM10 and PM2.5)

According to Fig.3, annual average PM_{10} values exceeded the EU annual limit values in all stations except Lazaropole, an EMEP station located at 1100 m a.s.l. in the west region.

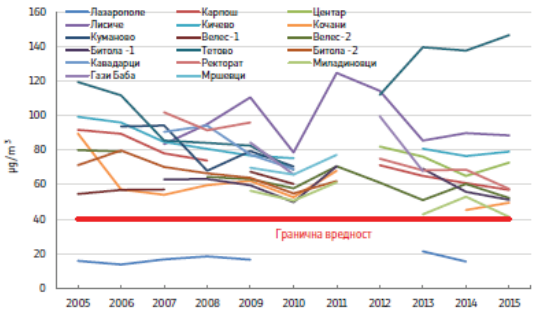


Fig.3 Average annual values for PM10 for the period 2005-2015 [7]

TABLE I. ILUSTRATE EU STANDARDS FOR PM10 AND PM2.5

Parameter	Average period	Value	Comments
PM10	Average daily value	50 µg/m^3	Not to exceed more than 35 days a year
PM10	Average annual value	40 µg/m^3	

The highest annual average concentrations of PM10 were measured in Tetovo and Skopje (Lisice) and exceeded 120 µg / m3. Levels of concentrations remain stable for the entire period between 2005 and 2015 year. It is estimated that the average value of PM10 in urban locations is approximately 80 µg / m3. Concentrations of PM10 in urban areas have pronounced and equal seasonal variations [7]. PM10 concentrations are high in the period December – January (Fig.4).

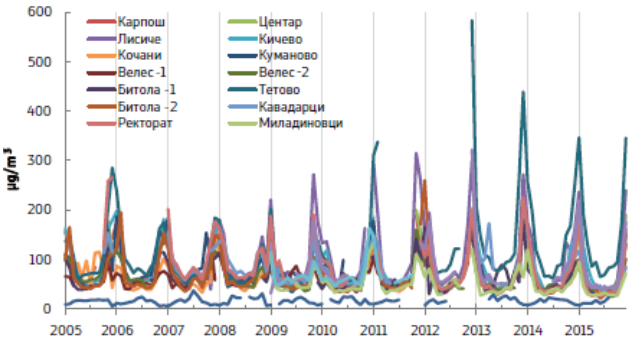


Fig.4 Average annual values for PM10 for the period 2005-2015.

Concentrations of fine suspended particles (PM2.5) are measured in two monitoring stations in Skopje since 2012.The average annual value of the concentrations is about 40-50 µg / m3, which is two times higher than the limit value.

B. Exceedance of air quality limit values in regions of Macedonia (NOx)

The main share of national emissions of NO2 originates from the energy sector (41% in 2014) and traffic (40% in 2014) (Fig.5). The total amount of NOx emissions in 2014 is approximately 32000 tons (MoEPP, 2016).

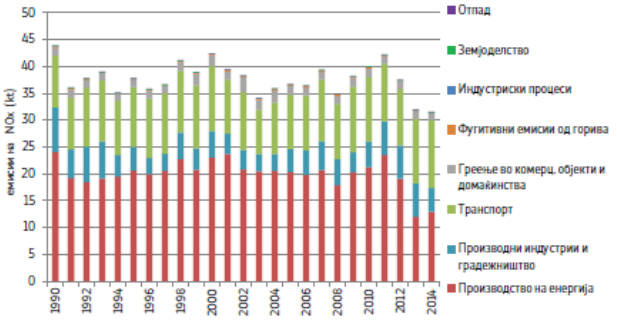


Fig.5 National emissions of NOx in the period 1990-2014, by sectors(MoEPP, 2016).

NOx emissions have seen a downward trend since 2011, which is the result of the reduced operation of the thermal power plant REK Oslomej, that is, the reduced consumption of coal and gasification on a heating plant.

TABLE II. ILUSTRATE EU STANDARDS FOR NO2

Parameter	Average period	Value	Comments
NO_2 (for Human health)	One hour	200 µg/m3	Not to exceed more than 18 hours a year
NO_2 (for Human health)	Average annual value	40 µg/m^3	

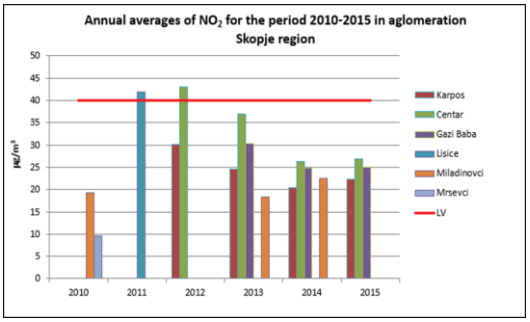


Fig. 6. Annual average concentrations of NO2 for the period 2010-2015 measured in Agglomeration Skopje region

Annual limit of NOx set for the protection of Human health in Macedonia is not exceeded only at measurement stations in Lisice(2011) and Center(2012) (Skopje region) (Fig.6)

C. Exceedance of air quality limit values in regions of Macedonia (SO_2)

Concentrations of SO2 in the air are visibly reduced in the past years, because the consumption of lignite and fuel oil is reduced. However, total national SO2 emissions are still high. Because of that it is necessary to introduce technologies to reduce emissions of SO2 especially in the main thermal power plants. The main share (over 90% in 2014) of national SO2 emissions comes from the energy sector that includes the production of electricity energy and heat

TABLE 3 ILUSTRATE EU STANDARDS FOR SO_2

Parameter	Average period	Value	Comments
SO_2 (for Human health)	Maximum one-hour average value	350μg/m^3	24 hours a year
SO_2 (for Human health)	Maximum daily eight-hour average value	125μg/m^3	3 days a year

In the past ten years, the reduction in SO2 concentrations is a relatively systematic trend in all monitoring stations. Since 2007 there have been no registered exceeding of limit values of SO2. (Fig.7)

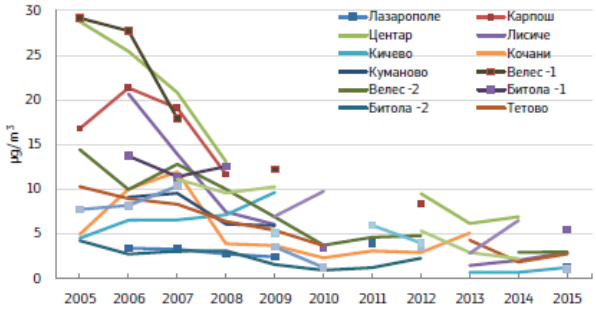


Fig.7 Average annual concentrations of SO2

D. Exceedance of air quality limit values in regions of Macedonia (CO)

The limit value of carbon monoxide in R. Macedonia still goes beyond a few days in a year in some cities. This is probably related to the old vehicles and widespread use of wood for heating in households. In cities throughout Europe, CO concentrations have dropped significantly since catalysts have become mandatory for new ones vehicles with petrol engines in 1992 year.

The main share in the national emissions of CO consists of heating in households (over 60% in 2014) and transport (27% in 2014) (MoEPP, 2016).(Fig.8).

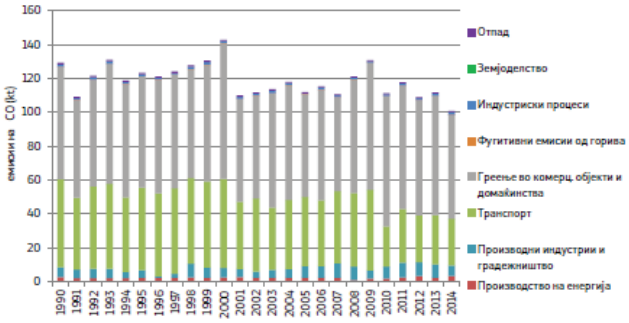


Fig.8 National emissions of CO in the period 1990-2014, by sectors

The limit value for CO (Table 4) is defined in national legislation transposing the air quality directive 2008/50 / EC (EU, 2008).

TABLE 4 ILUSTRATE EU STANDARDS FOR CO

Parameter	Average period	Value	Comments
CO (for Human health)	Maximum daily eight-hour average value	10mg/m3	

In Skopje and other urban areas of the country carbon concentrations monoxide occasionally exceeds the daily limit value.(Fig.9)

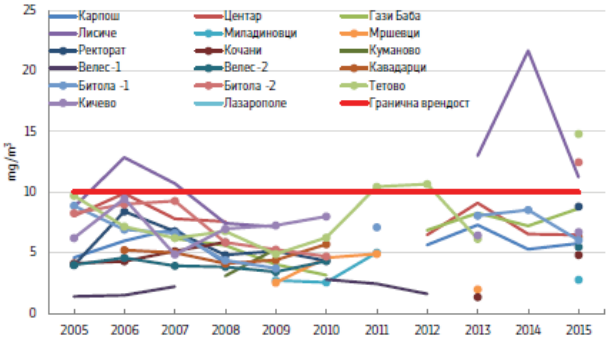


Fig.9 Average daily limit value of CO

Over the past few years, the limit value of CO has been exceeded in Skopje, Tetovo and Bitola, as major cities in the country that have more traffic density.

E. Exceedance of air quality limit values in regions of Macedonia (O_3)

The average ozone concentrations in cities are relatively low due to the presence of other polluting substances that break down ozone from the air. Concentrations of O3 usually increase with an increase in altitude, so the O3 concentrations in monitoring stations at a higher altitude may be higher than with stations stationed at a lower altitude.

Sometimes, in episodes of high solar radiation and temperature, high concentrations of O3 may occur in urban environments [15]. In the national legislation, air quality standards related to O3 are defined in order to protect human health and vegetation

TABLE 5 ILUSTRATE EU STANDARDS FOR O3

Parameter	Average period	Value	Comments
O3 (for Human health)	Maximum daily eight-hour average value	120 µg/m^3	Not to exceed more than 25 days a year
O3(Long - term goal of protecting human health)	Maximum daily eight-hour average value	120 µg/m^3	

The maximum daily eight-hour mean value is 120 µg / m3 and is defined in order to protect human health.

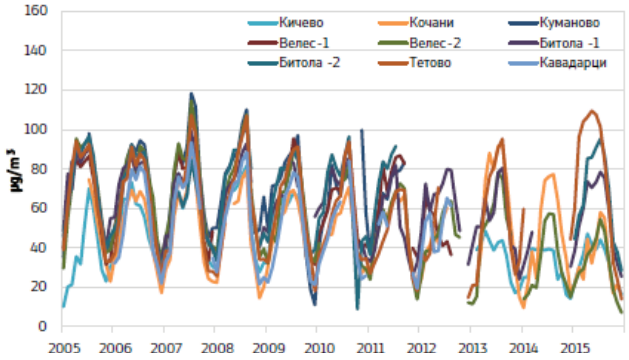


Fig.10. 8-th hour exceedance values for ozone (O_3) in Macedonia.

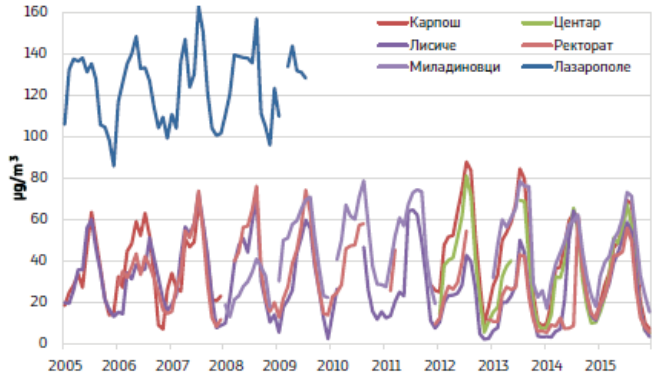


Fig.11. 8-th hours exceedance values for ozone (O_3) in Skopje and Lazaropole.

IV. DISCUSSION AND RECCOMMENDATIONS

Discussion and recommendations given in this section are founded on the results given by the previous maps and measured data obtained from the measuring station.

According to the legislation, it is necessary to implement measures for improvement of the air quality in order not to go beyond the limit values defined for the protection of humans.

The analysis shows a significant downward trend in the concentrations of sulfur dioxide in the ten-year period which was observed. High concentrations of suspended particles represent serious risk to the health of the population.

From the results above, the borders values for suspended particles are exceeded in the whole territory of R. Macedonia. In order to successfully decrease emissions in the air, efforts need to be taken by the central and local authorities, government, the business sector, but also the citizens need to raise their awareness regarding the environment.

IV. CONCLUSIONS

This paper is part of the continuous work performed by the Laboratory for Eco-informatics at the Faculty of Computer Science and engineering at the University "Ss. Cyril and Methodius" Skopje, Republic of Macedonia and represents follow up on the new developments in the EU based on the adoption of the new EU directive.

The ambient air quality in the Republic of Macedonia shows that national plan will be needed for PM_{10} as the limit values are exceeded in most zones.

Based on these research results, not only a review of the locations of all monitoring stations (MoEPP) as part of the overall action plan is needed but also an elaboration of the reasoning behind the selection of each site is more than welcomed solution. In this context a more comprehensive project is needed which might consider modeling the effect of the existing large sources of pollution across the country as a whole and to see what improvement of the ambient air quality this might produce when compared to other sources. This will be an important issue when considering what constitutes the Best Available Technique for each of the major plant and, importantly, the time allowed for the plant to reach new legislative standards.

REFERENCES

[1] Aery V.K., The Human Right to Clean Air: A Case Study of the Inter-American System, Seattle Journal of Environmental Law, Vol. 6, 2016, p.16.
[2] Daniel García S. J., Environmental protection and the European Convention on Human Rights, Council of Europe, 2005, p. 8.
[3] OJ L 152, 11.6.2008, p. 1–44
[4] OJ L 296, 21.11.1996, p. 55–63
[5] OJ L 163, 29.6.1999, p. 41.; OJ L 313, 13.12.2000, p. 12.; OJ L 67, 9.3.2002, p. 14.; OJ L 35, 5.2.1997, p. 14.
[6] Air quality in Europe — 2017 report, European Environment Agency, 2017, pg. 8
[7] Air quality in R. Macedonia — 2017 report, MoEPP, 2017, pg. 4,8
[8] Official Gazette of Republic of Macedonia No. 67/04, 92/07, 35/10, 47/11, 59/12, 100/12, 163/13, 10/15, 146/15)
[9] Official Gazette of Republic of Macedonia No. 50/05 and 4/13.
[10] R. Bojkovska, K. Mitreski, "Ambient air quality in the Republic of Macedonia from the perspective of EU Directives" – Proceedings of the 34-th International Conference ITI -2012, Cavtat, Croatia.
[11] EC Guidance on Assessment under the EU Air Quality Directives Air Quality Steering Group, 1998, p. 26-29.
[12] Horizontal Guidance Note IPPC H1Integrated Pollution Prevention and Control (IPPC) Environmental Assessment and Appraisal of BAT: UK Environment Agency, 2003, p.22-23.
[13] USEPA, 1995b. "User's Guide for the CALPUFF Dispersion Model", Prepared for the Office of Air Quality Planning and Standards, Research Triangle Park, NC, EPA-454/B-95-006,(July,1995).
[14] Air quality in R. Macedonia — 2012 report, MoEPP, 2012,pg13
[15] B. Stoimenovski, K. Mitreski, A. Naumoski, D. Serafimovski., "Analyzing the level of high troposphere ozone during the summer 2014 and 2015 in Skopje, R.Macedonia " International scientific conference - High technologies, business, society 2017 , March13-16, 2017, Bulgaria.

Practical application of data visualization techniques

Miodrag Cekikj, Antonio Antovski, Slobodan Kalajdzhiski
"Ss. Cyril and Methodius" University in Skopje, Faculty of Computer Science and Engineering
"Rugjer Boshkovikj" 16, 1000 Skopje, Republic of Macedonia
cekicmiodrag@gmail.com, toni.antovski@gmail.com, slobodan.kalajdziski@finki.ukim.mk

***Abstract* - Rapid technological development of the Internet and technology for network communications and data transmission have revolutionized nearly all areas of our life. Growth potential of hardware configurations and the development of powerful and sophisticated software processing platforms provide generation of vast amount of data on daily basis that are used in processes that characterize the model of the real world. Early stage in the development of technology and electronic communication meant facing the challenge of preserving the largest possible volume of data. Today we are witnessing the existence of models that successfully manage the acquisition, processing and storage of large amounts of data. The main challenge now is identified as a technique or approach which will allow easier interpretation and access to the data of user`s interest. This paper represents a systematic review and practical usage of techniques and methodologies for graphic interpretation of data and a visual display which allows easier visual perception, interpretation and proper future application.**

Keywords— visual perception; data visualization; graphic representation; visualization techniques; data analysis

I. INTRODUCTION

The visual representation of phenomena, objects and events in the surrounding area can be considered as core of the existence of human civilization. At the very beginning of the development of original communities, lacking the capacity to explain the surrounding processes, people begun to graphically express their thoughts in forms with different characteristics such as color, size and intensity. The ability for intuitive visualization of abstract emotions and phenomena were essential for ensuring a proper communication and cognitive development of human potential.

Historically, the first forms of concrete visual data representation come in the 2nd century BC in Egypt in order to organize astronomical information as a tool for navigation. Although the point was a table that is primarily a textual representation of data, it used the visual attributes of alignment, white space and at times rules to arrange data into columns and rows. The French philosopher and mathematician, Rene Descartes, invented this method of representing quantitative data originally, not for presenting data, but for performing a type of mathematics based on a system of coordinates [1].

Despite this kind of representation, we can use competently cartography maps that have narrative graphic function. One successful example of excellent cartography presentation comes from Charles Joseph Minard, French engineer who visualized Napoleon's Russian campaign in combination of data map and time - series [2]. Although the initial methods and techniques of graphical data presentation originally stem from 1913, however the person who introduced the power of data visualization to us was the statistics professor John Tukey of Princeton, who developed a predominantly visual approach to exploring and analyzing data called exploratory data analysis in 1977 [1].

Basically, the graphic representation of data was always more acceptable form of explanation of some phenomenon, whether it is art or scientific research interpretation. The evolution in the way of acquisition, processing, systematization and organization of data in electronic format is the main trigger for the emergence of the need for more sophisticated techniques of presenting data. It's actually the reason why this area is defined in a separate scientific discipline that today is fundamental to the concept of Big Data and the wording of the specific parameters that are part of large datasets.

The processing starts with the identification of visual perception that enables context modeling and interpretation of human behavior in the environment that surrounds us. It implies the need to define the techniques and methods that result in the visualization that will provide simpler perception, understanding and applying the required information. Systematic concept that leads to clarification and recognition of the specific features through graphical approach diverges as special scientific methodology that is part of the scientific disciplines related to the visualization of data (*II. Visual Perception: Pre - attentive processing*).

The potential for data visualization techniques can best be identified and recognized when the problem to be solved represents a large set of data. The challenge lies in the methodology of isolating, grouping and presenting information and parameters of interest and essential in the process of drawing a specific conclusion. For the purpose of a concrete analysis of the visualization techniques, as well as the criteria for selecting an appropriate display, the exploitation of a large data set for cancer diseases at the level of several continents will be processed. In the context of this, the significance of the appropriate structure and systematization of the data will be briefly explained in order to simplify identification of the

criteria for selecting the appropriate visualization tool (*III. Data Processing: Initial dataset sample*).

The main focus will be on the identification and interpretation of the visual context information in a form that will be important for consuming by the end users. It involves examination of a combination of visual techniques that provide statistical data visualization and analytical ground for understanding and highlighting the desired values. The techniques that will be recognized as the most suitable for practical implementation will be used as a basis for implementation of a software solution useful for referencing by academic and scientific institutes and organizations (*IV. Data Visualization Techniques*).

Finally comes the conclusion and further guidance with specific directions on practical software implementation in the context of presented techniques related to the study of Data Visualization techniques and methodologies. The content and conclusions of this research will be used as a basis for the practical implementation of the processed techniques in order to provide an efficient approach and a transparent way of discovering data of academic and scientific interest. (*VI. Conclusion and Further Guidance*).

II. Visual Perception: Pre - Attentive Processing

Visual perception represents the human ability to monitor, understand and interpret the visual information and circumstances present arounds us. In general, the interpretation is a brain process of giving a meaning to the environment changes in the context of using light in the visible spectrum reflected by the surrounding objects. The complexity of this process results in the obvious fact that our perception is a key aspect in the area of environment visualization [3].

At the base of all visualization techniques underlies the need for representation of objects in a manner that is easy acceptable for human comprehension [4]. The result of representations arising from quantitative factor which is closely related to the duration and the rate of success in the process of recognition. This leads us to the concept of parallel, prompt and effortless process identified as pre - attentive processing, which can be understood as technique that tries to find and explain characteristics that make use of the potential of low - level human vision [5].

Despite serial attentive processing, pre - attentive processing tries to construct exact human perception from a combination of simple visual attributes that are part of the contextual model [6]. The main purpose is mapping and distinguishing differences in objects forms, primary color, length, shape, orientation, but without the need of significant effort, time and focus.

The best way to explain pre - attentive processing is through some basic case scenarios. For example, to count the number of occurrences of specific number in a given long array of numbers it is necessary to scan all the numbers sequentially as presented on Fig. 1.

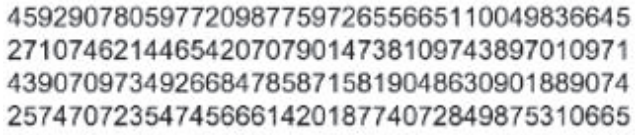


Fig. 1. Sequentially scanning example of attentive processing.

But, according to the scenario presented on Fig. 2., we can notice different visual approach if we have the same array but with a certain change of the color of demanded number.

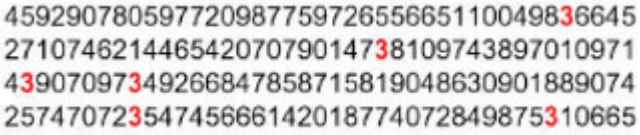


Fig. 2. Parallel scanning example of pre - attentive processing.

The result is even better if we change the shape and style because in that case we get clearly defined pop up from the surrounding [7].

Another interesting example defines three visual cases where pre - attentive processing could help to retrieve better results and confirm the existing of different fields of application. Considering this we can identify Target detection, Boundary detection and Counting and estimation cases [5].

In Target detection, a subject has to detect rapidly whether there is a defined target item. This representation is visible on Fig. 3.

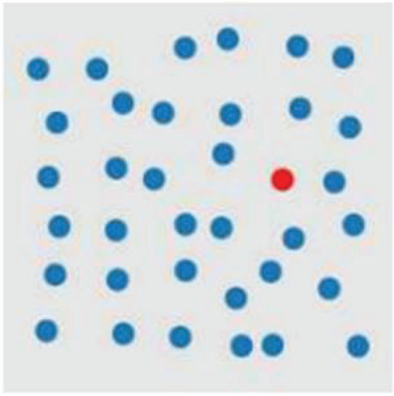


Fig. 3. Target detection.

In Boundary detection, a subject has to detect rapidly and accurately a boundary between two groups of elements, as shown on Fig. 4 (boundary detection: easy to detect in the image on the left; vertical boundary between fifth and sixth column of the image on the right is harder to detect - red circles and blue squares on the left, blue circles and red squares on the right).

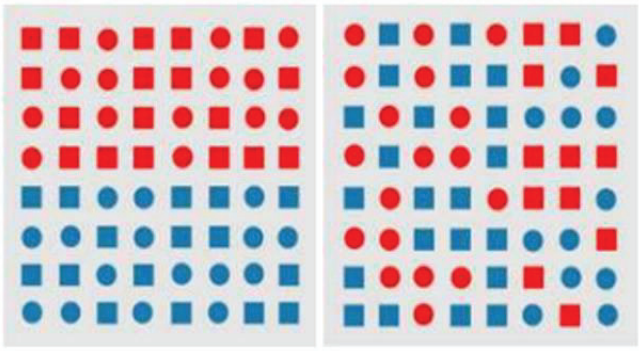


Fig. 4. Boundary detection.

In Counting and estimation, a subject has to count or estimate the number of elements using a common pre - attentive property, presented on Fig. 5 (rapid estimation: the number of red dots can be counted at one glance).

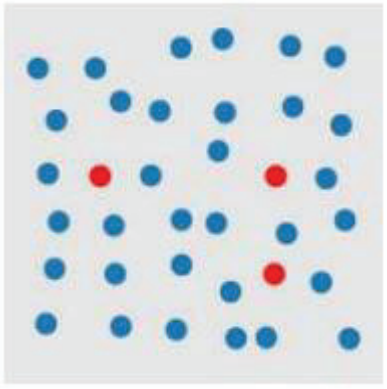


Fig. 5. Counting and estimation.

According to this, in order to make the information easy and efficient to perceive, it should be visually encoded with pre - attentive attributes [6]. This approach is significant part of data visualization and interpretation field and leads to generating content easily affordable for target consumers.

III. Data processing: initial dataset sample

Data visualization or information visualization can be implemented in many different scientific and industrial areas. Through the application of the main concepts of exploration and explanation as well as using visual techniques we are able to visualize statistical data and provide analytical ground for understanding and highlighting the desired values. In this section, we are going to present different techniques and methodologies for data visualization as introduction to our field of interest and presenting current and potential future scope of work.

The data set of interest is a set of publicly available data files for cancer diseases for different geographic regions, i.e. continents. Basically, it's about the Cancer Incidence in Five Continents (CI5) project which is the result of a long collaboration between the International Agency for Research on Cancer and the International Association of Cancer Registries [13].

The available content contains segmented data divided by several attributes: sex, cancer identification number, age, number of cases of cancer divided by age, and number of cancer risk cases. Data from the source is not structured, which means that the records are not sorted and grouped depending on the attributes, their value, and meaning. This form does not allow successful application of any of the visualization techniques due to the fact that it is about unconnected data structures of different nature and character. Due to the inadequate representation of the downloaded format, data preprocessing was required in order to have a clearer representation and to obtain an appropriate structure and scheme for simplifying the grouping, sorting and interpretation of the attributes and their domain of potential and valid values.

The pre - processing procedure consisted of 4 separate phases, which aligned the data structure, linking the inconsistent identification values, as well as segmentation depending on the geographic region and the type of disease. The end result implies obtaining a single file with a total of 10 802 184 records that form the data set that is subject to processing and further visualization. For the future utilization of the DBMS potential and performance, an appropriate ER diagram was created for the data structure, which allowed the storage of attributes in the form of a traditional SQL relational database.

IV. Data visualization techniques

Data visualization is the process or more general the science discipline of displaying the presentation of data in some visual format. In its simpler form, it is defined as information which has been abstracted in some schematic form, including attributes or variables for the units of information [8]. The main goal is to create appropriate visual context of the information so that end users will be able to consume or efficiently communicate the key points of interest. Besides the fact that this process has evolved into a separate science area, the creative techniques and approaches for visual modeling are the actual reason why this area very often is identified as art.

Data visualization or information can be implemented in many different scientific and industrial areas. Through the application of the main concepts of exploration and explanation as well as using visual techniques we are able to visualize statistical data and provide analytical ground for understanding and highlighting the desired values. It is well known that the human brain is simply better at perception, retaining, processing and recalling information that has been graphically presented [9]. There is a large set of data visualization tools and techniques applicable in different areas, but within the purposes of our work we will present the most common, well - known and applicable types of graphic visualizations. In this section, we are going to present different techniques and methodologies for data visualization as introduction to our field of interest and presenting current and potential future scope of work.

Given the nature and scope of the initial dataset sample, we have decided to research and complement the application of multiple visualization techniques through which the various elements related to the semantic interpretation of the data will be highlighted and identified. For this purpose, as one of the most appropriate libraries, we selected the non - commercial and academically targeted version of the amCharts library.

Basically, it is a JavaScript library that supports a number of visualization techniques and tools for data of a different type and structure. From a technical point of view it provides well documented APIs and support for all versions of modern search engines.

In their basic form, the graphical presentations can fall under the 2D area, temporal, multidimensional, hierarchical and network categories [10]. Each of these categories define different visualizations like table, histogram, scatter plot, line chart, bar chart, pie chart, area chart, flow chart, bubble chart, multiple data series or combination of charts, time line, Venn diagram, data flow diagram, and entity relationship diagram [11].

There also additional methods and visualization approaches that are more advanced and enhanced in comparison with the already mentioned conventional methods. These kinds of additional methods are Sankey Diagram, Customer Journey Map, Word Cloud, Gantt Chart, Radar/Spider Chart parallel coordinates, tree map, cone tree, semantic network, etc. [11, 12].

Considering the processed dataset sample and taking into account the fact that it contains data related to different geographical regions, the most logical choice for general visual display of information is two - dimensional map. 2D area types of data visualization are mostly related to the map representation or usually geospatial in the meaning that they relate to the relative position according to the Earth's surface. The most popular representatives of this groups are Cartogram, Choropleth and Dot Distribution Map [10].

All of these visualizations are very popular when we need map to convey the information of an alternative variable, such as population or travel time or when we need measurement of a statistical variable, such as most visited website per country or population density by state. Common to all these approaches is the data systematization or pre - processing procedure with identifying intercorrelation in order to get distilled dataset ready for visualizing.

Within the amCharts library, there are many types of map visualizations. The main difference is of a contextual character meaning that the manner of display depends on the structure of the data, their significance, as well as the expected functionalities for interaction and review. In our case and needs for the most appropriate we consider the representation named as Zooming to Countries Map. This type od visualization is represented on Fig. 6.

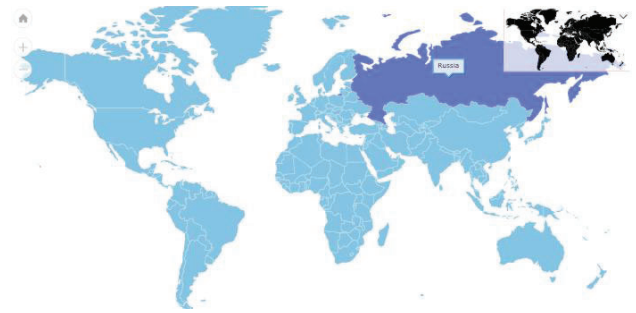


Fig. 6. Zooming to countries map.

The main idea is in addition to the standard interactions with the map (zoom in / zoom out, scroll preview, reset view, initial point), the module can provide the functionality for selecting a region. The selection would involve the opening of a dialogue on a general overview of the number of registered cases, the number of potential cases of cancer, and a period within which the statistical analysis is systematized. These statistical parameters are obtained as a result of the implemented business logic on the server side. In addition to the possibility of selecting a country, the interactive map should also offer the so - called. a hover function that will allow displaying the name of each country that is in focus at the time of the review.

In accordance with our perception, besides the generalized mapping of map information, it is necessary to apply other visualization techniques in order to clearly display all aspects and elements of the attributes that are of contextual importance for the object of interest. This will enable a detailed overview of a specific structure with the ability to filter the display and interactions in accordance with the needs of the data analyst.

Referring to this, successful selection of a geographic region will mean initialization and displaying of detailed percentage analysis of the total number of registered cases in the specific regions at the selected country level. This kind of multidimensional data elements can be designed to represents the target data in two or more dimensions according to the end consumer requirements. This group includes Pie Chart, Histogram and Scatter Plot presentations [10].

Our choice for this type of display will be Pie Chart With Legend representation with advanced functionalities and interactions like selection / de-selection of specific regions, possibility for remarks, preservation as a photograph, etc. The pie chart is displayed on Fig. 7.

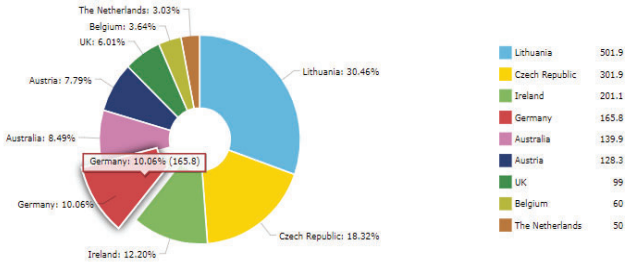


Fig. 7. Pie chart with legend.

Within the same group of visualization graphs, the 3D Bar Chart/Clustered Bar Chart structure can also be applied. Generally speaking, this approach has identical contextual but different visual meaning for the end user and can basically be used for statistical/numerical analysis of the number of potentially diseased people divided in particular regions.

This graphic interpretation is important because the result is a systematic overview of the disease's potential in geographic areas of concern. The 3D Bar Chart is available on Fig. 8.

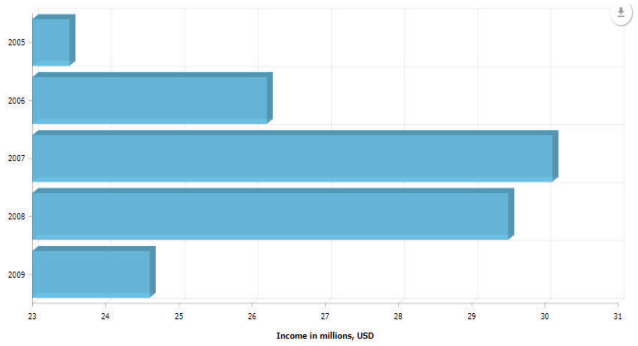


Fig. 8. 3d bar chart.

An interesting subject of reflection and analysis is the detailed comparison of the gender potential. According to the standards and best practices for displaying this kind of data, visualization should be expressed through percentages and numerical values, according to the age limits defined in the initial database sample.

For this purpose, a graphic named 100% Stacked Column Chart appears as the most perfect display. There are other related visualizations, but we consider that the selected selection results in the most obvious parameters preview and the possibility for the fast and simple interaction if needed. The visual representation is displayed on Fig. 9.

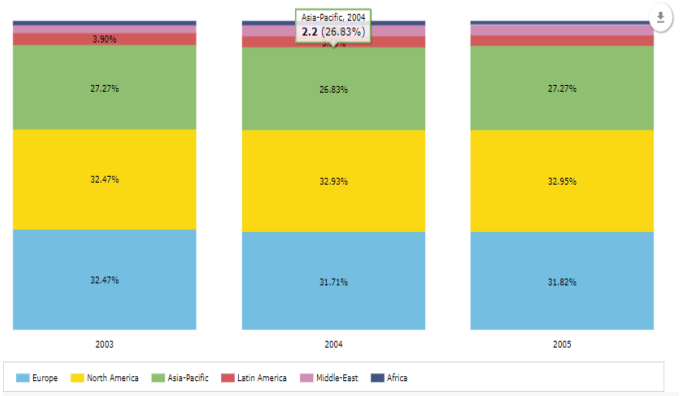


Fig. 9. 100% stacked column chart.

An analysis that we find to be of exceptional importance is the main comparison of the number of patients with the number of potentially ill within the predefined time periods. Such a request is an ideal candidate for applying a technique called Trend Lines. In essence, this is a visual representation that allows modeling of two or more data sets consisting of an identical range of potential values.

The ability of the wave representation is in the possibility of interactive selection of an arbitrary period or value defined by the respective axes.

The end result of such an implementation will almost always mean complete control and precise overview of the requested data sets. Trend Lines graphic visualization is shown on Fig. 10.

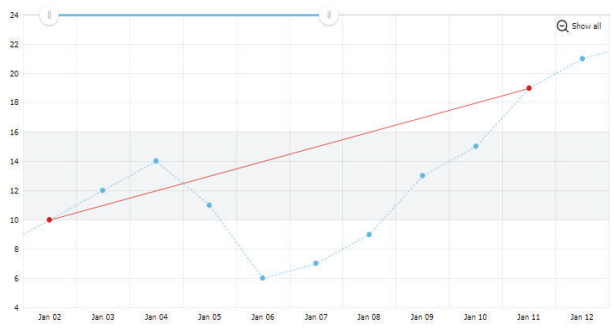


Fig. 10. Trend lines.

V. CONCLUSION AND FURTHER GUIDANCE

In this article, we surveyed the most common data visualization techniques and tools through the interaction and dependencies related to the cancer incidence measured at several different geographical locations. We presented the intrinsic value of visual expression application in order to simplify analysis and interpret large and interconnected initial dataset sample.

Today we are witnesses of growing global trend in development of new types and methodologies for visual presentation of data that will provide acceptable visual perception and cognitive processing by the end users. According to this, the main goal is still the same, to understand the scope of work, define the potential data source, analyze the characteristics and choose correct systematic approach that will result in practical information visualization.

This survey may serve as guide for all researchers involved in the process of interconnecting and visualizing all available sources of information in order to achieve their point of research. In addition, the content represents basis for potential development of a software tool through which all presented graphical visualizations will be practically implemented. In general, the content is divided into separate sections describing the most appropriate techniques and their potential to prepare and visualize big set of data and properties which are closely related to the circumstances and the development of cancer diseases.

One part of the paper is devoted to a review of the initial phase of defining the problem of interest and specific procedures implemented to extract and organize the necessary data. The main challenge pops up when the resulting data set should be analyzed, correlated and structured according some existing research databases that already implement tools that visualize a number of their properties and dependencies.

Our future point of interest will be the concrete application of the techniques for data visualization in practical terms. This implies the fact of using the existing relational data structure in the context of making a web - oriented software solution that will primarily enable the consumption of data that is subject to processing. Processed visualization techniques will be the basis for representation of information and attributes associated with their semantic meaning. The idea is to provide an overview

platform that will offer conditions and modules for related academic or scientific research in the field.

The development of this kind of a platform will mean an open possibility for integration with already existing knowledge databases as well as software tools aimed at collecting, processing, visualizing and manipulating sensitive data and statistical analyzes of this type. The processed content is precisely in this context, successful systematization and the possibility for detailed study of all the significant segments from which additional conclusions can be reached or which may be input parameters in other scientific researches related to human health.

Human health and the causes of modern diseases with severe consequences are fields of processing of many scientific disciplines. A common element for all can be identified in the need for unique and consistent data - based statistical structures, as well as tools for their simple and proper interpretation.

REFERENCES

[1] Stephen Few, January 10, 2007, 'Data Visualization past, present, and future, Perceptual Edge - Visual Business Intelligence for enlightening analysis and communication, [https://www.perceptualedge.com/articles/Whitepapers/Data_Visualization.pdf]

[2] Mária Kmeťová, 'On History of Information Visualization', Department of Mathematics, Constantine the Philosopher University in Nitra, International Scientific Conference on Distance Learning in Applied Informatics - DivaI 2010, May 4 - May 6, 2010, Štúrovo, Slovakia

[3] C. Ware, 'Information Visualization: Perception for Design, Second Edition', Morgan Kaufmann Publishers, Inc., 2004

[4] Stephen Few, 'Data Visualization for Human Perception', The Encyclopedia of Human - Computer Interaction, 2nd Ed., The Interaction Design Foundation

[5] Andreas Albustin, Stefan Bacheitner, Arber Djerdjizi, Bernd Hollerit, 'Pre - Attentive Processing', 5 May 2010, [http://courses.iicm.tugraz.at/ivis/surveys/ss2010/g2-survey-preatt.pdf]

[6] Stephen Few, 'Tapping the Power of Visual Perception', Perceptual Edge - Visual Business Intelligence for enlightening analysis and communication, September 4, 2004, [http://www.perceptualedge.com/articles/ie/visual_perception.pdf]

[7] Colin Ware, 'Information Visualization: Perception for Design, Third Edition', Morgan Kaufmann, May 18, 2012

[8] Michael Friendly, 'Milestones in the history of thematic cartography, statistical graphics, and data visualization', Department of Mathematics and Statistics at York University, August 24, 2009, [http://www.math.yorku.ca/SCS/Gallery/milestone/milestone.pdf]

[9] Ben Collins, 'Visualization Techniques to Communicate Data', Online Behavior - Marketing Measurement & Optimization, April 2017, [http://online-behavior.com/analytics/data-visualization-techniques]

[10] Otto Ottinger, '15 Most Common Types of Data Visualisation', Datalabs Agency, December 21, 2014, [http://www.datalabsagency.com/data-visualization-news/15-most-common-types-of-data-visualisation/]

[11] Lidong Wang, Guanghui Wang, Cheryl Ann Alexander, 'Big Data and Visualization: Methods, Challenges and Technology Progress', Science and Education publishing (SciEP), July 20, 2015, [http://pubs.sciepub.com/dt/1/1/7/]

[12] Dom Brooks, '8 More Common Types of Data Visualization', Datalabs Agency, December 12, 2015, [http://www.datalabsagency.com/articles/more-common-of-types-data-visualizations/]

[13] International Agency for Research on Cancer, 'Cancer Incidence in Five Countries', World Health Organization

Decentralizing The Health Information Exchange Systems – A Blockchain Based Approach

Zlate Dodevski
iborn.net
Skopje, Republic of Macedonia
zlated@iborn.net

Vladimir Trajkovik
Faculty of Computer Science and Engineering
Skopje, Republic of Macedonia
trvlado@finki.ukim.mk

***Abstract*— The main idea of this paper is to analyze the blockchain based approach to the health information exchange systems. Due to the complex nature of health-related data, they are condemned to inertness and rigidity when it comes to their portability and condemned to hermetic isolation and discretion when it comes to the consequences of their abuse. This paper will analyze the possibility and the potential of the blockchain technology to be applicable in the systems dealing with health information exchange. The approach will be compared with traditional implementation model of the current solutions to analyze the architectural transformation and the benefits of decentralization and distributed consensus. Limitations and shortcomings of this approach will be discussed as well, along with areas of improvements.**

Keywords*— *blockchain; electronic health records; health information exchange; smart contracts; decentralization

I. INTRODUCTION

Blockchain represents an implementation of peer-to-peer network that works as a layer on top of the Internet. The underlying architecture was introduced as a concept at October 2008, as a part of the proposal of Satoshi Nakamoto to start virtual and digital currency system [1]. This cryptocurrency system introduced tectonic movements and a lot of discussions and theories in the financial, economic and social structures and departments. The inspiration for a new era in these fields came most because of the characteristics of blockchain systems to be run and organized by distributed authority, as opposite of the nature of existing centralized solutions and involvement of third-party intermediaries. The cryptocurrency called Bitcoin, and the whole cryptocurrency system behind it, is a first implementation of the blockchain technology.

Blockchain in its essence is a distributed system that stores records for specific transactions. It's a distributed ledger of peer-to-peer transaction, built by connected blocks of transactions. Well-defined cryptographic techniques are fundamentals which enable this technology. They are the core principles that enable interaction (in form of storing, exchanging and overview of data) between each participant in the network, bypassing the need for intermediaries and regulatory bodies to acquire trust. In this distributed system, there is no need for central authority, instead of that we have records of transactions which are stored and shared between all involved participants in the network. Every participant must be familiar with every interaction in the network and every transaction needs to be verified by all participants to be successful and valid. The verification of the interactions and the distributed state of the network are the principles that enable the collaboration in system in absence of trust between members, while the global log of transactions is immutable [2][3][4].

Many blockchain enthusiasts that advocate the benefits of this technology will say that we are facing the dawn of technology revolution. This emerging technology is putting strong steps in the well-defined lifecycle, and it's close to reach the plateau of productivity. It constantly surprises the public with the growth of the community and characteristics of mature solution with high potential [5].

The key things that blockchain offers as potential is not in the domain of performance improvement of some business process, it's not some solution of pressing issue in specific industry department, nor some "hazardous" technology which will cast a shadow at some traditional model of digitized business and offer lower costs and workarounds. The potential that offers blockchain as a solution lies in the tectonic movements and reengineering of irreplaceable areas of many industries which don't have characteristic of troublesome and high "seismic activity" [6][7].

The use of the blockchain technology is characterized as revolutionary transformation that changes the face of the digitized world, mostly because of the decentralized approach provided and the established presence of distributed consensus among the participants in the network. These fundamentals backed by cryptography techniques can offer different approach to critical processes related to security, interoperability, chronology of electronic events and privacy. These characteristics add the attribute fundamental next to the blockchain technology, attribute that has the possibility to redefine whole companies and businesses and attribute that is almost a super power in the hands of those who got on the express train of blockchain, giving them right to dream and preach about the dawn of new economic and social systems.

Considering the complex nature of health information, many obstacles appear when it comes to their portability and flexibility. Additionally, the consequences of their abuse transform the process of securing the health-related data into

challenging task. The blockchain technology, even though envisioned as a support of cryptocurrency system, shows great potential for adaptability and wide range of practical use. The purpose of this paper is to analyze which of the characteristic of the blockchain are applicable in the process of manipulation of health-related data [8].

If we look at the novelties that blockchain brought to us in an isolated state, separated from the context, more than sure, our gaze would have stuck on several statements that are part of the mantra of actualized technology [9] [10]. These statements seem to be the long-awaited answer that anyone who has dealt with the field of the health information exchange (HIE) expects. Sharing clinical data in the form of different data elements can be simplified by taking advantage of the benefits that blockchain offers as technology. The question that is currently circulating by the enthusiasts of the fast-growing blockchain technology is whether it is really a solution to the problems as theorists preach or is just another example of a technological hammer in the hands of practitioners who are searching for their own nail [11] [12] [13].

II. Domain

In the health informatics, in the last period, the process of exchange of health information has been more and more active. HIE (Health Information Exchange) represents electronic transfer of clinical and / or administrative information between different (mostly competitive) healthcare organizations. When talking about the actors in this ecosystem, they can be of different size and form, including clinical centers, hospitals, laboratories, insurance companies, pharmacies, emergency centers, nursing homes, public health centers, etc. The data exchanged can differ and can be part of a wide range, moving from a summary of medical examinations, referrals, to laboratory results, and even medical history of specific patient.

This research will be dedicated to the symbiosis of the benefits of blockchain technology and the process of exchanging health information. Hence the purpose of this research is to emphasize the characteristics of this approach that will bring to HIE systems if the blockchain technology is incorporated [14] [15].

One of the key novelty that can be introduced by using the blockchain is the involvement of an independent entity (third person) as a central authority, which in this research we define as a problem of trust.

All forms of health information exchange tend to involve an independent entity that can guarantee trust, conflict resolution and offer implementation of the technical aspects of data sharing within an enterprise, community, country, or region. In addition, each person has an individual right to privacy, which each data exchange system should respect. Hence, the entity that guarantees the exchange should not only satisfy the complex differences between the ecosystem participants, but also take care of the legal and regulatory policies that coming into force to ensure the protection of data. The freshness that blockchain brings to the technological world is that transactions and interactions with the network can occur in a state of absence of trust. All the so-called. third parties may be completely excluded from the systems [16] [17].

This research will also deal with the problem of protecting personal data. In the world of electronic health information, health data is not just private secure piece of data, but also personal data related to specific patient's medical history. Participants in the HIE process that generate health data can have confidence in the infrastructure of the blockchain system because one of the things it brings as a revolutionary technology is the automation of the data integrity. In addition, giving the possession of personal data should also define the decision-making power of who can manipulate with it. A problem that blockchain can overcome [18].

Additionally, Blockchain is an art of secure management and exchange of different types of data through unsecured channels. Blockchain technology, in its first application, provided support for the operation of the Bitcoin cryptocurrency and the whole digital currency system. In other words, blockchain is a distributed database designed to manage unique digital assets within a system composed of multiple parties. Because the Bitcoin exchange does not represent anything but data exchange, it is considered that the infrastructure and the blockchain concept in the exchange of health data is proven and tested [19].

The rest of this paper is organized as follows. Section III introduces blockchain architecture. Section III.I shows typical and traditional model of implementation used in health information exchange systems. Section III.II introduces solution that uses blockchain as an underlying architecture. Section IV summarizes the technical challenges and the recent advances in this area and Section V concludes the paper.

III. Blockchain As Underlying Architecture

The purpose of this research is to justify the decision to use decentralization and distributed consensus introduced by blockchain technology in the domain of the health information exchange systems. This paper is starting point in the process of investigating the usability of the approach and the next step is performing a simulation to consider the practical and technical shortcomings and difficulties. Since most global researches are based on improvements to EHR systems, this research will also investigate the options for sharing health information where we have the absence of an electronic health records system. The blockchain-based approach which is focus of the research should be able to offer a solution to the problem of how to transfer medical data in a safe, controlled and rapid manner to the institutions that need them. The data exchanged can be non-electronic in its original state.

The approach should provide security in the preservation of the scanned documents, it should provide a permissions system that allows the patient to decide who should see the data and should separate the identity of the patient from the documents, although the identity of the patient can be determined uniquely.

The so-called impersonal set of data for which the prototype can uniquely prove their identity, but they do not identify a user (personally non-identifiable information) can be used as training data in the process of conversion from paper to electronic medical records. Hence, a system that uses machine

learning can use that data to facilitate the conversion process from paper to electronic medical records.

At the beginning, we must start with the fundamental things that blockchain as technology is characterized. In its essence, this technology is a distributed system. Distributed systems are built from the so-called nodes that are defined as individual participants in the big picture and can communicate with messages among themselves. Additionally, the nodes are characterized as processor units that have their own processing power along with individual memory. This type of architecture brings enhanced system reliability. With other words, if one of the nodes stop to function as intended from the very beginning, the entire system is not affected by such an occurrence. The system should continue to run without any problems to meet the initial intentions and requirements.

When it comes to electronic health records and health information exchange almost all the solutions are using the client-server approach. In the context of electronic medical records this approach is known as “better safe than sorry” approach. The reason why this architecture is the choice of many EHR systems is the power of the centralized system to more easily deal with the authorization and authentication in its purest form of application. Additionally, we should not neglect the fact that health data has a degree of high sensitivity. Hence, the tradition of storing data sealed in an isolated server component and the strict supervision of which and how to use the data is rather logical and pragmatic. It is a "happy place" for all implementers of systems that use health data as a moving data set, although the needs of the modern world and real problems are striking down its limits and challenges and increasingly impose the need for finding another type of structure and architecture.

The blockchain has been characterized as a technology that can solve the decentralization. The main concept behind the decentralized architecture is that there is no central point of control, nor a centralized authority that is responsible for controlling actions that take place within the system.

When this academic description of the decentralization conflicts with the context that imposed this research particularly, many things arise as doubts and questions. Decentralization and a system that uses health data in the practical world gets an oxymoronic tone and for a longer period it was inconceivable that these two terms be used in the same sentence.

However, blockchain as a technology represents a new era of decentralization where special techniques and algorithms are introduced in order to establish a level of trust between the nodes and reduce the need for the presence of the so-called. third person as authority and mediator.

After introducing significant evolutionary moments regarding the choice of architecture, the next step is to specify the components from which the entire system is composed and to get a closer look and overview.

The challenge to defend the usability of the technology in the domain of health data is probably starting with the use of peer-to-peer architecture. As a representative of the decentralized model, this architecture is ideologically opposite to the Client-Server architecture, mainly because of the absence of a central server that owns all resources. Instead, we get a network of nodes that are capable of independent functioning, which in terms of the privileges and rights in the network are equal. Communication is key in this type of architecture, so a communication channel between all members of the architecture is established, and they are obliged to take one of two roles. Representatives either take the role of servers on a resource or take on the role of a demanded resource.

Fig. 1 Ecosystem participants.

To understand the architecture and choice of component components, we must start with the ecosystem participants and their benefits in an interaction with the system. These same participants will also try to analyze them as an integral part of the "traditional" electronic data exchange model to make the difference. Of course, the data flow and direction can be easily seen if the participants in the global image are involved. The ecosystem consists of the stakeholders presented in Fig. 1:

- Representative of the category of patients. The interaction with the system takes place in the direction of transferring medical records to different types of institutions;
- Different type of institutions (health organizations, government / ministry, regulatory bodies);
- Insurance companies or a representative of another organization whose interaction with the system is aimed at providing data needed for business analysis or building business intelligence;
- Researchers;
- Representative of health institutions. The interaction with the system is aimed at the internal process of

generating medical records within the institution and their propagation to the appropriate places;

III.I Traditional Model Of Implementation

In the traditional model of implementation of the electronic medical records systems the most common is the client-server architecture and having that in mind, the data flow is based on well-defined communication and security protocols. With other words, in such type of architectures, the sensitivity of the data brings increased isolation and centralization, which acts inversely proportionality to the flexibility and simplicity of the transfer of medical records, a concept which is at the heart of this analysis. If we introduce the context of interoperability between these systems, we must not overlook the fact that we are facing a major challenge because of the diversity of performances and the additional layer of customized functionalities within an institution.

We can define the following multi-level interaction:

- within an institution where the patient interacts with a representative of a health institution who uses an electronic system for feeding, generating and manipulating health data within a health institution;
- between health institutions as part of the HIE framework (exchange of health data). This interaction is characterized by a regulatory body, a strict communication protocol and strictly defined message standards and additional infrastructure
- among researchers who use health data for scientific research (whether they are lacking a personal identity or with consent to use) and the system for the exchange of medical records.
- providing data needed for business analysis or business intelligence formation from representatives of relevant institutions.

III.II Blockchain Based implementation

The Fig. 2 shows the components of the blockchain solution [20].

A. Blockchain

Blockchain is a distributed trusted storage space for data. By itself, the blockchain is nothing more than a data structure with which the data are connected to each other and stored. In this type of architecture, it is not used as a storage space for medical records, but for proper assignment of references and indexes to the IPFS data storage space, based on a strict set of rules.

B. Distributed Autonomous Application (DAPPS)

This type of applications as a software solution do not differ from traditional applications in terms of connecting individual users and servers to a given service or resource. However, distributed self-contained applications carry with them a set of features that distinguish them from others. The most obvious difference is the way the code is executed (on a decentralized P2P network).

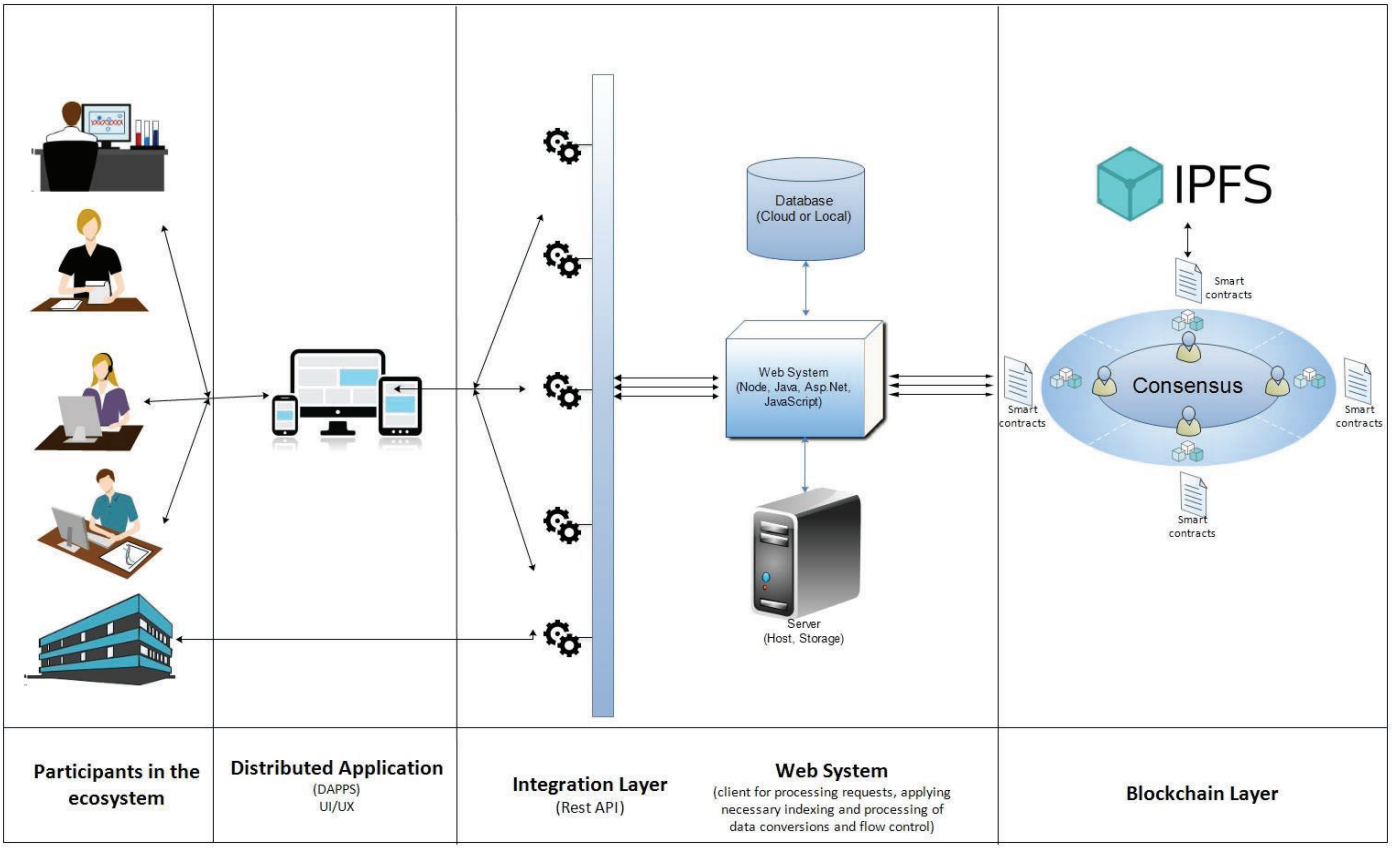


Fig. 2 Implementation of the system by using Blockchain.

C. DAO

When talking about this software solution within the entire system, we must define the so-called decentralized and autonomous organization as a concept.

DAO is an organization whose operation is based on certain rules embedded in infrastructure as computer programs and those rules in the context of blockchain technology are called “smart" contracts. In other words, the systems that are part of this organization have a set of rules that are initially established and due to the digital nature of data in these systems, all participants are digitally forced to obey those rules.

D. Smart Contracts

As mentioned earlier, smart contracts represent small programs with a certain static logic and flow control, and their use by all participants and users of the distributed autonomous organization ensures the reliability established initially in the organization. In the approach using blockchain as underlying structure in HIE systems, smart contracts carry the whole logic that takes care of the proper authorization of certain medical records, as well as their updating and manipulation [21].

E. IPFS (Interplanetary File System)

It is a protocol and a network of users designed to store and share data. It is essentially a decentralized P2P distributed file system that finds application in blockchain based solutions. The purpose of this distributed system is to store data that is not susceptible to modifications and unauthorized alterations and to allow addressing them for their search and access.

IV. Discussion And Challanges

The blockchain architecture as underlying architecture in the health information exchange systems can bring a lot of fresh and new ideas. The core characteristics can be applied to achieve greater data integrity, highly controlled and managed access of data and interoperability improvements, as well. But, the technical and practical realization of this approach is still with limitations and shortcomings [22].

The Fig. 3 presents the main components of the blockchain layer and drills down the services that the offer to the architecture described above. This components overview will help us to better understand the future discussions, limitations and challenges. Each of the component of the blockchain layer can be analyzed as separate module and many improvements were done to meet some expectation in specific implementation.

This limitation of blockchain systems has long been the subject of discussion by developers and academics. Probably the biggest issue is the scalability, since to achieve distributed trust and consensus, each of the node should process all the transaction and to be fully functional it should have the latest state of the transaction ledger [23]. If we analyze the Bitcoin implementation, we have a limitation of 1 MB block size and the network can process approximately 7 transactions per second. We are also dealing with high transaction fees, that needs to be taken into consideration if we analyze the feasibility and the business model of the solution [24] [25].

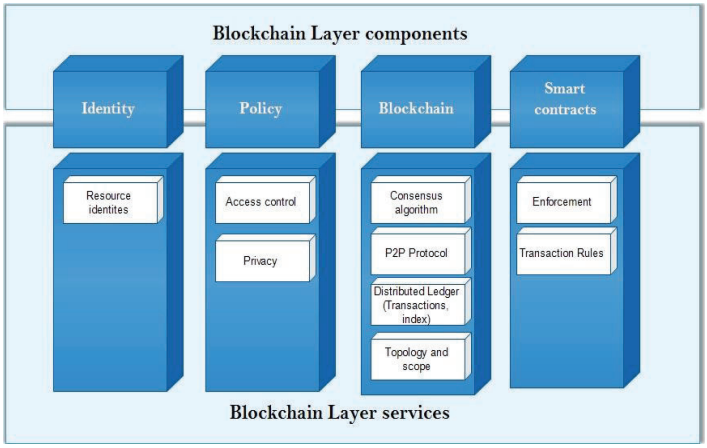


Fig. 3 Blockchain Layer.

The selection of consensus algorithm, plays a big role in the performance and characteristics of the solution. Blockchain implements Proof of Work (PoW) algorithm [1], and while is robust and secure it requires for participants in the network to spend resources on their own, such as computation time, power and energy in order to solve the given puzzle and participate in building new blocks in the backbone. There are a lot of efforts from other organizations to improve and optimize the consensus and network parameters such as Proof of Stake, Hashgraph, Ripple, etc [26].

The clear vision of network topology and taxonomy of blockchain should also be subject of discussion. Every solution that is using the blockchain approach should carefully decide what kind of blockchain configuration will apply to the architectural design and what will be the scope of use [27]. There are three types of networks that can be formed by using blockchain technology [28]. Private, public and consortium network.

By using a public blockchain the solution will have a really wide scope and cover many use cases, but the performance will be put on stake and the transactions will cost much more. Additionally, we need to consider the mining process and what will be the incentive of the stakeholders in the ecosystem.

A consortium blockchain consists of multiple organizations and the difference is that the distributed consensus is managed by specific set of nodes which are authorized to do that. On the other side, in private blockchains the module that covers the permission management is run by one organization and with those characteristics is the easiest to configure. The last two solutions have better cost efficiency, performance and flexibility but they are moving away from the fundamental decentralized concepts of the blockchain approach.

V. Conclusion

In this paper, we analyzed the decentralized approach given by the blockchain technology to the health information exchange systems. The fundamentals of this technology can offer different approach to critical processes related to security, interoperability, chronology of electronic events and

permission control management. To have more clear vision of the approach, we analyzed the traditional approach to the implementation of health information exchange systems and put a comparison with the blockchain-based. Even-though the decentralization of the solutions and blockchain as an underlying architecture can bring many novelties and innovations in health information exchange processes, still the practical and technical part of the idea has many shortcomings and downfalls.

REFERENCES

[1] Satoshi Nakamoto, "Bitcoin: A Peer-to-Peer Electronic Cash System," 2009.

[2] Staff, W. P, "Blockchain for Dummies, " Wiley & Sons Canada, Limited, John, 2017.

[3] Mougayar, W, "The business blockchain: Promise, practice, and application of the next Internet technology, " Hoboken, NJ: John Wiley & Sons, 2016.

[4] Price, M., "Blockchain: The complete guide to understanding Blockchain, " North Charleston, SC: CreateSpace Independent Publishing Platform, 2017.

[5] Lakhani, M. I. (2017, February 17). The Truth About Blockchain. Retrieved September 08, 2017, from https://hbr.org/2017/01/the-truth-about-blockchain.

[6] Swan, M., "Blockchain: Blueprint for a new economy," Beijing: O'Reilly, 2015.

[7] Four genuine blockchain use cases. (n.d.). Retrieved September 09, 2017, from https://www.multichain.com/blog/2016/05/four-genuine-blockchain-use-cases/.

[8] Blockchain Technology in Health Care: Decoding the Hype. (2017, February 09). Retrieved September 09, 2017, from http://catalyst.nejm.org/decoding-blockchain-technology-health/.

[9] Blockchain Use Cases in Healthcare. (2017, May 15). Retrieved September 08, 2017, from https://www.intelligenthq.com/innovation-management/blockchain-use-cases-in-healthcare/.

[10] Nichol, P. B. (2016, December 20). IBM blockchain in healthcare rallies for patients. Retrieved September 09, 2017, from https://www.cio.com/article/3151429/innovation/ibm-blockchain-in-healthcare-rallies-for-patients.html.

[11] IBM Global Business Service Public Sector, "Blockchain: The Chain of Trust and its Potential to Transform Healthcare – Our Point of View, " 2016.

[12] IBM Global Business Service Public Sector, "Healthcare rallies for blockchains: Keeping patients at the center, " 2016.

[13] Project Publications ‹ MedRec: Using Blockchain for Medical Data Access, Permission Management and Trend Analysis – MIT Media Lab. (n.d.). Retrieved September 08, 2017, from https://www.media.mit.edu/projects/medrec/publications/.

[14] Analyst, C. B. (2016, September 30). Blockchain Technology Can Enhance Electronic Health Record Operability. Retrieved September 08, 2017, from https://ark-invest.com/research/blockchain-technology-ehr.

[15] Azaria, A. (n.d.). A Case Study for Blockchain in Healthcare:"MedRec" prototype for electronic health records and medical research data – MIT Media Lab. Retrieved September 08, 2017, from https://www.media.mit.edu/publications/medrec-whitepaper/.

[16] Ackerman Shrier A, Chang A, Diakun-thibalt N, Forni L, Landa F, Mayo J, van Riezen R, Hardjono, T. Organization: Project PharmOrchard of MIT's Experimental Learning "MIT FinTech: Future Commerce." (2016) Blockchain and Health IT: Algorithms, Privacy, and Data

[17] John D. Halamka, MD Andrew Lippman Ariel Ekblaw. (2017, May 18). The Potential for Blockchain to Transform Electronic Health Records. Retrieved September 09, 2017, from https://hbr.org/2017/03/the-potential-for-blockchain-to-transform-electronic-health-records.

[18] Molteni, M. (). Blockchain Could Be the Solution to Health Care's Electronic Record Woes. Retrieved September 09, 2017, from https://www.wired.com/2017/02/moving-patient-data-messy-blockchain-help/. 2017.

[19] Dixon, B. E., "Health information exchange: Navigating and managing a network of health information systems, " Amsterdam: Academic Press, an imprint of Elsevier, 2016.

[20] Antonopoulos, A. M., "Mastering bitcoin: Programming the open blockchain, " Sebastopol, CA: O'Reilly, 2017.

[21] Diedrich H., "Ethereum: Blockchains, digital assets, smart contracts, decentralized autonomous organizations, " 2016.

[22] A. Gervais, S. Capkun, and B. Ford, On the Security, Performance and Privacy of Proof of Work Blockchains. 2016.

[23] M. Scherer, "Performance and Scalability of Blockchain Networks and Smart Contracts," MSE thesis, Dept. Computing Sci., Umeå University, Umeå, Sweden, 2017

[24] Zheng Z, Xie S, Dai HN, Wang H (2016) Blockchain Challenges and Opportunities: A Survey.

[25] M. Vukolić, "The quest for scalable blockchain fabric: Proof-of-work vs. BFT replication", Proc. IFIP WG 11.4 Workshop Open Res. Problems Netw. Secur. (iNetSec), pp. 112-125, 2015, [online] Available: http://www.vukolic.com/iNetSec_2015.pdf.

[26] Z. Zheng, S. Xie, H. Dai, X. Chen, H. Wang, "An Overview of Blockchain Technology: Architecture Consensus and Future Trends", *2017 IEEE International Congress on Big Data (BigData Congress)*, pp. 557-564, 2017

[27] X. Xu et al., "A Taxonomy of Blockchain-Based Systems for Architecture Design", Proc. 2017 IEEE Int'l Conf. Software Architecture (ICSA 17), pp. 243-252, 2017.

[28] V. Buterin, "On public and private blockchains," 2015/08.

Security Patterns for Microservice Account and Identity

Tihomir Tenev
Faculty of Mathematics and Informatics
Sofia University "St. Kliment Ohridski"
5 James Bourchier blvd., Sofia, Bulgaria
tenevtih@gmail.com

Dimitar Birov
Faculty of Mathematics and Informatics
Sofia University "St. Kliment Ohridski"
5 James Bourchier blvd., Sofia, Bulgaria
birov@fmi.uni-sofia.bg

***Abstract*—Microservice based application, used in many cases as cloud solution, has advantages as high degree of scalability and possibility to be split and deployed in many hosts. These benefits itself increase the risk of security problems as tampering of sensitive information and elevation of privilege. Many guides contribute with methods on how to determine vulnerable points, and part of them aim to help developers by advising with proper resolution. However, such guides are either rather common or describe different security problems. In that context, we decided to make a step forward in constructing thorough classification of security patterns, which strongly relates to microservice architecture. Our classification consists of four sections as "Communication", "Deployment and host management", "Data Management" and "Accounts and Identity", and here we represent the fourth section "Accounts and Identity". We briefly describe each of the selected patterns toward microservice architecture style as well as relate them with relevant threat category of the STRIDE model and list of security properties. That approach facilitates users in better finding of appropriate pattern depend on their scenario.**

Keywords—Software architecture; Cloud; Microservices; Security Patterns; Pattern Classification

I. INTRODUCTION

Security is one of the main concern in developing distributed system, since it operates, in many cases, with sensitive information. Therefore, working toward mitigating of security gaps, at earlier phase of application constructing, prevents sequential problems. In that context, we made a classification of security patterns [1], which helps in choosing a proper pattern for applications based on microservice architecture [2]. The subject that we consider here is in regard to using accounts and identity patterns for building a security wall between external clients and microservices as well as among a set of microservices.

Microservice is an architecture style organized by atomic independent components called *microservices*. Each microservice should be as decoupled and cohesive as possible [2]. These constraints lead to splitting an application onto small independent units and spreading them at different places. Such an architecture style might be used in diverse solutions as server-side applications, mobile applications, cloud applications etc., where each of them can be built as a set of components. In that paper, we decided to work toward mitigating security issues with cloud solution.

Cloud system is a functional part of a distributed system [3] and has the ability to grow with contemporary activities, which rapidly increases its complexity. The National Institute of Standards and Technology (NIST) [4] categorizes the cloud context in meaning of three service models: Software as a Service (SaaS), Platform as a Service (PaaS) and Infrastructure as a Service (IaaS). The significant models, that we consider here, are Platform as a Service and Software as a Service. PaaS provides platform for deploying customer application, and excludes the support of any hardware assets. SaaS is the highest layer and provides complete application, on which a consumer is not given possibility to rearrange the code as well as manage the infrastructure.

One of the major prerequisites for application developing is to has possibility for external interaction by using User Interface. Following microservice architecture style there are two options: either fragmenting the UI to call each microservice directly [2] or using API Gateway [5] as mediator, which prevents direct access (Fig. 1). The first option requires rearranging microservices behavior and refers to PaaS layer, while the second option requires building API Gateway for calling and conveying some information to a microservice and therefore refers to SaaS layer.

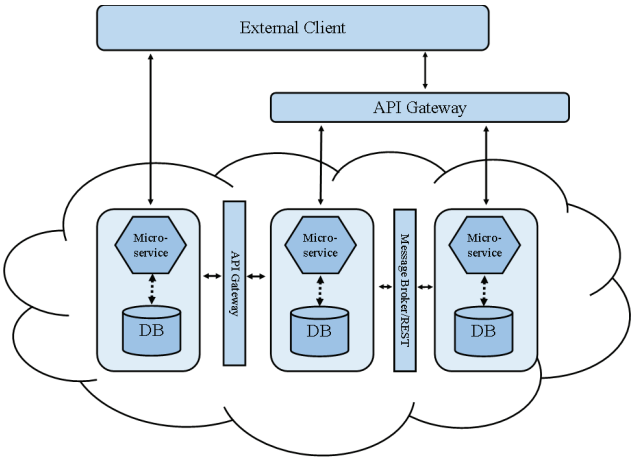


Fig. 1. Cloud application based on microservice architecture style

Depend on the application architecture, some of the requests need involving at least two microservices for computing the result. Hence, the microservices should communicate one another internally and such an activity mainly happens via either REST [6] or Message Bus [5] [7]. In that paper, we present a manner on how to secure microservices from the external users as well as the interaction among a set of microservices by using security patterns.

Essentially, security patterns [1] are described as a solution for a problem, which occurs frequently in a certain case. Furthermore, each security pattern should consider certain software constraints to implement its resolution accordingly.

There are many classifications [8] [9], but we could not find any toward microservice architecture style. For that reason, we decided to make a step forward in building a security pattern classification based on microservice architecture style. The whole classification consists of sections as “Communication”, “Deployment and host management”, “Data Management” and “Accounts and Identity”. In that paper, we present the fourth section “Accounts and Identity”.

For better understanding, we describe each of the selected patterns toward microservices and then relate them with relevant threat category of the STRIDE [10] model whose acronym stands for *Spoofing*, *Tampering*, *Repudiation*, *Information Disclosure, Denial of Service* and *Elevation of Privilege*. In additional, to increase the comprehensibility, we represent each of the STRIDE threat category to corresponding security property as *Authentication*, *Data Integrity, Non-repudiation, Confidentiality, Availability* and *Authorization*. We describe the STRIDE and the relation with these security properties in next section.

The paper is structured as follows: Section 2 explains in detail how we filter proper security patterns and how we construct our classification. Section 3 shows a set of security patterns collected in meaning of microservice accounts and identity. Section 4 shows other papers in that matter. Section 5 derives conclusion and place the next steps for future work.

II. Classification

By design, Security patterns originate from plenteous experience of developers, engineers and researchers. All policies and methodologies that they offer are used as a template in modelling of a secure system. Moreover, such patterns can be enforced with microservices in building completed cloud solution.

Each security pattern has its own structure. In many cases, it consists of several sections - *Intent*, *Context*, *Problem*, *Solution*, *Structure*, *Implementation*, *Consequences* and *Known uses*. However, we decided to narrow them to only three points as *Context*, *Solution* and *Known uses*. They mostly reveal the essential of each pattern and can help in better selecting. *Context* describes the nature of a situation, which includes domain assumption and expectation of a system environment. *Solution* guides a consumer how to solve a problem by providing a decision. *Known uses* gives example on where a certain pattern is already implemented.

After estimating an appropriate reading method, we had to filter suitable patterns, which mostly relate to accounts and identity perspective. Based on that, we collected a set of security patterns, which we present in Table I “Security Patterns for Accounts and Identity”.

The next step, after filtering of security patterns, is classifying them in threat models approach. Such a combination gives more clarity to a security pattern, which in turns helps in better understanding.

In essential, threat modelling facilitates security experts by estimating vulnerable points. In this regard, binding vulnerable points with the most suitable pattern is a good way for building a secure application without eavesdropping or tampering of sensitive information.

After thorough research, we decided to use threat modelling approach based on STRIDE [10]. This approach advises readers to split a software onto several components for identifying the types of attacks that they may encounter.

TABLE I. Security Patterns for Accounts and Identity

Patterns	Spoofing / Authentication	Tampering / Integrity	Repudiation / Non-repudation	Information Disclosure / Confidentiality	Denial of Service / Availability	Elevation of Privilege / Authorization
A Pattern for WS-Trust	x		x			x
Access Control List	x				x	
Account Lockout	x					
Actor and Role Lifecycle Pattern			x	x		
Administrator Objects	x					x
Authenticated Session	x					
Authenticator	x					
Authorization						x
Biometrics Design Alternatives	x					
Capability	x				x	
Client Input Filters	x					x
Credential Delegation				x		
Directed Session		x				
Grant-Based Access Control Pattern				x		
Password Design and Use						x
Privilege-Limited Role				x		
Role-Based Access Control (RBAC)	x					
Session-Based Attribute-Based Authorization	x					x

Here is a more detailed description of the six threat categories:

- *Spoofing* is a type of fraud where a violator tries to gain access to a user's system or information by pretending to be the user. For example, when a certain user waits for an event, and the event is triggered by another user with malicious purpose, usually this leads to irrelevant proceeding. The best way to prevent such a threat is to build guard method. Furthermore, that method closely relates to *Authentication* and implies using of security patterns in this regard.
- *Tampering* means making illegal changes of data flow when a user should not. In many cases, tampering leads to adjusting the content of a file, memory unit or data, transferred over network. Nobody wants to receive or read information, which persists with incorrect data. To prevent such a threat, keeping data in consistent state with correct content is a must. Therefore, *Tampering* closely relates to security property called *Data Integrity*.
- *Repudiation* is in meaning of rejecting to accept something. An example of such a situation is when a

user did something, but claims he didn`t touch it. This in many cases can lead to less responsibilities. Such threat can be handled with logging of each activity went over a system. Enforcing *Non-repudiation* related patterns imply preventing of repudiation claims.

- *Information Disclosure* is in situation when an unauthorized user can see information, which is forbidden for disclosing. The source of data might come from a running process, storage or data flow. The best way to mitigate that risk is to enhance the *Confidentiality*.
- *Denial of Service* mostly relates to exhausting of a certain part within a system. This kind of attack works against memory, CPU or data store and mostly leads to software inaccessibility. Moreover, it may fill up network bandwidth and inflict high degree of time responding. The option to mitigate *Denial of Service* is to look how to increase *Availability*.
- *Elevation of Privilege* means allowing user to execute a command without required access rights for that. Nobody, except Administrator user, should operate with major processes. Otherwise, somebody may bring a system in corruption state. *Authorization* is the property, which gives someone permission to do or to own something and sup-ports preventing *Elevation of Privilege*.

As can be seen, we extend each STRIDE category by linking it with a certain security property, which mostly fits in that context. The properties are *Authentication*, *Data Integrity*, *Non-repudiation*, *Confidentiality*, *Availability* and *Authorization*. Such an approach implies injection of greater perceptiveness in selecting appropriate security pattern and leads to a more intuitive classification.

Part of the patterns run into more than one STRIDE category, however, this can only facilitate the right choice. For example, Access Control List [12] protects a microservice against *Spoofing* and *Daniel of Service*.

III. Enforcing Security Patterns

As part of distribution architecture, microservice has several drawbacks as: tough deployment process, data management with many microservices, complexity of testing process, external communication among microservices etc. We decided to focus on how to enhance the security of microservices to prevent at least tampering and disclosure of sensitive information from external clients. To solve such kind of threats, we suggest using “Accounts and Identity” security patterns.

In next several paragraphs we explain each pattern from Table I toward three aspects: microservice architecture style, what communication method covers and what STRIDE abbreviation deal with:

A Pattern for WS-Trust [11] shows a manner on how using a token can authenticate two microservices to admit transferring of a sensitive information. As in Fig 2 is shown sequence diagram in process of accepting a token along two microspecies. That secure relation works against three STRIDE points: *Spoofing*, *Repudiation* and *Elevation of Privilege.*

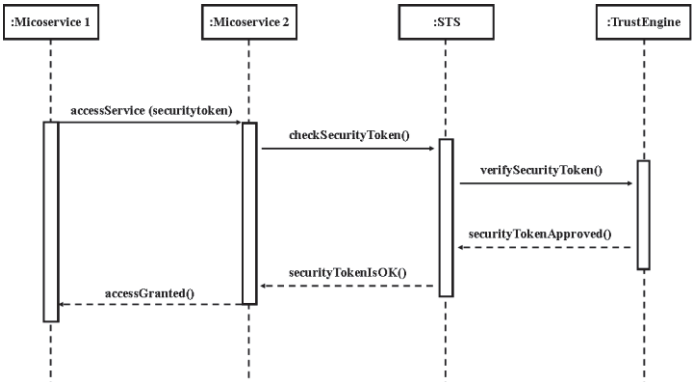


Fig. 2. Sequence Diagram for accessing a microservice using a token

Access Control List (ACL) [12] describes how implementing an access matrix and associating each microservice to a certain user and list of activities can enhance the security. That pattern might be implemented at API Gateway to block wrong authorization and unappropriated exploitation of a microservice. It prevents *Spoofing* and *Denial of Service*.

Account Lockout [13] pattern denies attempting to disclose a password by using, for example, password-guessing tools. The pattern sets a number of attempts for password entering. That approach can be implemented at API Gateway to works toward *Authentication*.

Actor and Role Lifecycle Pattern [14] is associated to the employee life cycle. Depends on changing of responsibilities either from promotion or change of a role, a certain employee usually fulfills different tasks. Changing access rights dynamically can be enforced at API Gateway layer to works against *Repudiation* and *Information Disclosure*.

Administrator Objects [15] gives a solution on how to manage the assignments of a user-role and a role-privileges. It creates a separate list of administrator roles with different privileges to manage users and roles. Therefore, that pattern can be applied at API Gateway as well as at each microservice to prevent further problems with *Authentication* and *Authorization*.

Authenticated Session [13] or so called “Single Sign-On” gives possibilities to a user, who is already authenticated in opening a restricted page, to proceed accessing other protected pages without re-authenticating. It is applicable either for API Gateway or for accessing microservices directly to ensure *Authentication*.

Authenticator [1] verifies that a user or a microservice is the one that claims it is. As in Fig. 3 is shown sequence diagram, where two microservices authenticate one another. That pattern can be applied in API Gateway and within a microservice to prevent *Spoofing*.

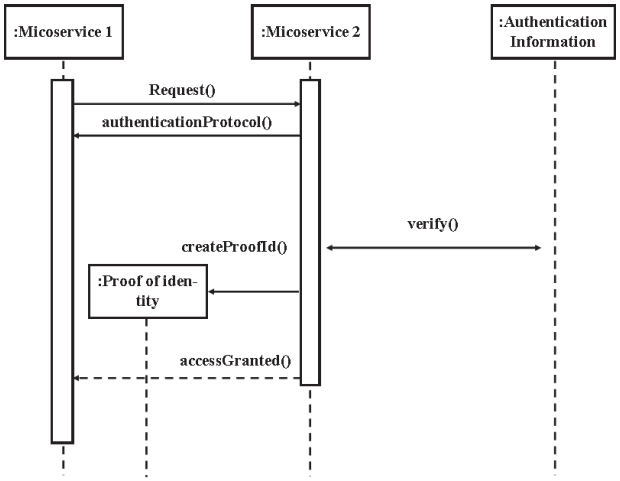


Fig. 3. Sequence Diagram for Authentication

Authorization [1] describes whether a user or a microservice is authorized to communicate with a certain microservice. That pattern works against *Elevation of Privilege* and can be combined with Authenticator [1] for complete solution.

Biometrics Design Alternatives [1] considers mechanisms for finger prints, iris recognition, retinal scanning etc. It authenticates a user and can be applied at API Gateway for preventing *Spoofing*.

Capability [12] gives a solution in case there is an enormous list of users and using Access Control List [12] is not appropriate due to long processing time. As in Fig. 4 is shown sequence diagram, on which *Microservice 1* requests access to *Microservice 2* by using the *Capability.* The request is handled by Policy Enforcement Point (*PEP*) for indicating the identity, the object, and the access type of *Microservice 1*. The access is given after accepting the request by Policy Decision Point (PDP).

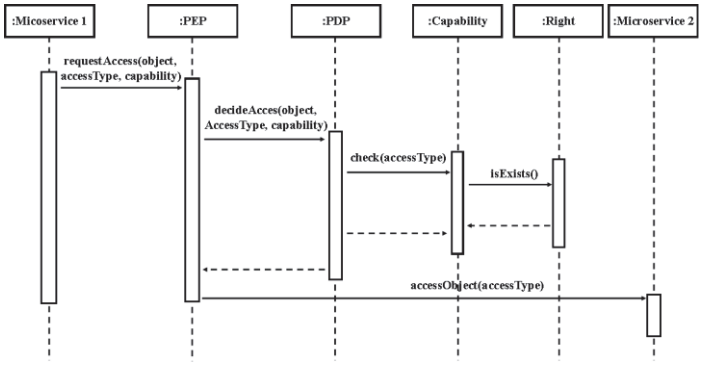


Fig. 4. Sequence Diagram for Capability

Client Input Filters [13] describes how untrusted clients may tamper the input flow and grant more access. Filtering the request, which came either directly to a microservice or through API Gateway implies security enchantment toward *Spoofing* and *Elevation of Privilege*.

Credential Delegation [17] addresses an issue, where a microservice has two instances: one in operational state and another in standby state [5]. The first one, for some reason, is stopped and the second one works on behalf of the first. Using signed certificates (proxy certificates) allows switching communication flow in secure way. As in Fig. 5 is shown sequence diagram with two instances of one microservice and another standalone microservice. The pattern works toward *Information Disclosure.*

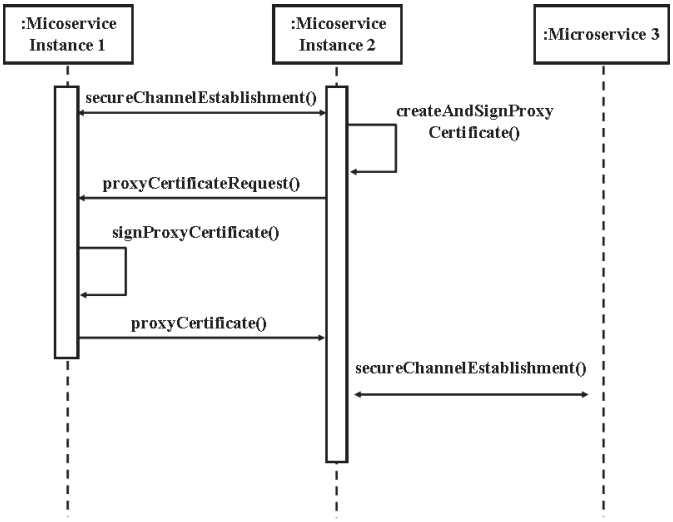


Fig. 5. Sequence Diagram for Credential Delegation

Directed Session [13] ensures that a user calls several microservice in appropriate sequence without skipping some as well as not providing session data to malicious users. The pattern can be implemented at API Gateway to save the *Integrity*.

Grant-Based Access Control Pattern (GBAC) [18] describes a situation, where a user is given access to a microservice except ciphered data that the microservice owns. Cryptographic material for data decrypting is allowed after providing the credentials to API Gateway. The pattern covers *Information Disclosure*.

Password Design and Use [1] provides a list of requirements, which helps in designing, creating and managing passwords for authorization purposes. Applying it at API Gateway contributes to *Elevation of Privileges* point.

Privilege-Limited Role [15] advises how API Gateway admits users to observe the staffs/pages on which they have access. The rest are hidden to prevent *Information Disclosure*.

Role-Based Access Control (RBAC) [1] describes how users are assigned to roles, and each role is given appropriate rights for accessing microservices. API Gateway can enforce it to prevent *Spoofing*.

Session-Based Attribute-Based Authorization [16] gives access to a resource based on a list of provided attributes. For example, a certain user may see different information depends on his location. Therefore, API Gateway should collect that information to decide what microservices to call in case of a user request. The pattern gives solution for *Spoofing* and *Elevation of Privilege*.

IV. Related Works

Hafiz et al [8] show a similar method, however, their categorization presents an approach, which follows "one pattern per one STRIDE point". Here is not the same, because a pattern may participate in more than one STRIDE point.

The focus in [19] is mainly against securing of web applications. Its author distinguishes security patterns in two directions: procedural and structural. Structural patterns can be applied in an already completed product, while procedures are aimed in phase of planning and writing software.

Fern et al [20] consider only three security patterns: Authorization, Role-Based Access Control, and Multilevel Security. They argue that these are the only three basic patterns that can be applied at each level of entire system. However, there are many scenarios, which require specific patterns. For instance, Authenticated Session [13] describes how a user can reach many microservices without re-authenticating.

V. Conclusion

There are many publications, which guide with approach on how to enhance security [10], however, current paper is slightly different. It aims to show how security patterns can help in developing secure cloud system, since that system can be built via microservice architecture style. The point that we consider here relates to securing microservices from external clients by applying accounts and identity related patterns.

Our next step was to look for an appropriate way of reading security patterns. Initially, they consist of several sections and we decided to use only three of them: *Context*, *Solution* and *Known uses*.

After estimating the appropriate reading method, we filtered a list of security patterns, which closely relates to microservice *Accounts and Identity* aspect. Additionally, we categorize it toward the threat model entitled STRIDE. This in turn led to a more intuitive classification.

For simplicity, we associated each category of STRIDE with six security properties: *Authentication*, *Data Integrity*, *Non-repudiation*, *Confidentiality*, *Availability* and *Authorization*. This implies understandability of the current approach and facilitates developers in choosing appropriate security patter in building cloud decision.

Further works is considered in researching of more security patterns, which match the rest aspects: *Deployment and host management*, *Communication* and *Data Management*. The entire classification will help in enhancing the security at earlier phase of building a cloud decision based on microservice architecture style.

References

[1] M. Schumacher, E. Fernandez-Buglioni, D. Hybertson, F. Bushmann, and P. Sommerlad, "Security Patterns Integrating Security and Systems Engineering", 2006

[2] S. Newman, "Building Microservices", O'Reilly Media Inc. 2015

[3] M. Richard, "Software Architecture Patterns", O'Reilly Media Inc., 2015 , Page 27 – 35

[4] P. Mell and T. Grance, "The NIST Definition of Cloud Computing", Special Publication 800-145

[5] C. Richardson and F. Smith, "MICROSERVICES From Design to Deployment", NGiNX Inc 2016, pp 15-20

[6] R. Fielding, "Representational State Transfer (REST)", 2000, Chapter 5

[7] G. Hohpe and B. Woolf, Enterprise Integration Patterns, 2003

[8] M. Hafiz, P. Adamczyk and R. Johnson, "Growing a pattern language (for security)", Onward! 2012, Pages 139-158

[9] A. Motii, B. Hamid, A. Lanusse and J. Bruel, "Guiding the selection of security patterns based on security requirements and pattern classification", ACM Transactions on EuroPLoP 2015, July 2015

[10] A. Shostack, "THREAT MODELING: Designing for Security 1st Edition", John Wiley & Sons, Inc, 2014

[11] O. Ajaj and E. B. Fernandez, "A pattern for the ws-trust standard for web services," in Proceedings of the Asian Conference on Pattern Languages of Programs, 2010, pp. 1–11

[12] N. Delessy, E. B. Fernandez, M. M.Larrondo-Petrie, and J. Wu, "Patterns for access control in distributed systems," in Proceedings of the Conference on Pattern Languages of Programs. New York, NY, USA: ACM, 2007, pp. 1–11.

[13] D. M. Kienzle, M. C. Elder, D. Tyree, and J. Edwards-Hewitt, "Security patterns repository, version 1.0," 2003

[14] B. Elsinga and A. Hofman, "Control the actor-based access rights," in Proceedings of the European Conference on Pattern Languages of Programs. UVK - Universitaetsverlag Konstanz, 2002, pp. 233–244

[15] S. R. Kodituwakku, P. Bertok, and L. Zhao, "Aplrac: A pattern language for designing and implementing role-based access control," in Proceedings of the European Conference on Pattern Languages of Programs. UVK - Universitaetsverlag Konstanz, 2001, pp. 331–346

[16] E. B. Fernandez and G. Pernul, "Patterns for session-based access control," in Proceedings of the Conference on Pattern Languages of Programs. New York, NY, USA: ACM, 2006, pp. 8:1–8:10

[17] M. Weiss, "Credential delegation: Towards grid security patterns," in Proceedings of the Nordic Conference on Pattern Languages of Programs, 2006, pp. 65–70

[18] A. Cuevas, P. E. Khoury, L. Gomez, and A. Laube, "Security patterns for capturing encryption-based access control to sensor data," in Proceedings of the International Conference on Emerging Security Information, Systems and Technologies. Los Alamitos, CA, USA: IEEE Computer Society, 2008, pp. 62–67

[19] S. Romanowsky, "Security Design Patterns Part 1", Morgan Stanley Online, September 2001

[20] E. Fern and R. Pan, "A pattern language for security models", PLoP 2001 Conference, 2001

Securing a Home Network by Using Raspberry Pi as a VPN Gateway

Bogdan Jeliskoski, Biljana Stojcevska, Adrijan Bozinovski
School of Computer Science and Information Technology
University American College Skopje
Skopje, Macedonia
bogdan.jeliskoski@live.com, { stojcevska, bozinovski }@uacs.edu.mk

Abstract— The need of home Internet of Things (IoT) applications is rapidly increasing as the ease of access and diversity of technologies is growing. Securing the network is the first thing to consider before connecting any IoT device on the Internet. The central theme in this paper is providing protection for IoT devices in a home network and their safe connection to the Internet. The security threats and risks of using IoT devices on the Internet are outlined and discussed and a solution for security in an IoT home network is presented. The protection used in the IoT home network is based on a Raspberry Pi device that is installed as a Gateway on the internal network, with all IoT devices connecting to it as their gateway, or via VPN tunnels. The paper explains the problems, challenges and the role the gateway plays in protecting the devices on the network and presents an analysis of the results of the testing of the presented solution.

Keywords— Internet of Things; Raspberry Pi; network security

I. INTRODUCTION

The modern way of life would be inconceivable without quick and easy access to information and would be difficult to imagine without computer networking systems that connect different or similar devices in one whole. The need for network connectivity naturally arises primarily from the need for continuous exchange of information.

Today, there is a large number of devices that are used for Internet connectivity, and, according to purpose, they are commonly divided into consumer, infrastructure, and business applications [1]. It is inevitable to mention Internet of Things (IoT) in the field of network connectivity and information transfer. IoT is seen as the next stage of an information revolution in which there should be a connection to everything in the world. IoT contributes to the modernization of services that a person can receive through the network connection, or the Internet [2] [3].

IoT represents interconnection of devices that incorporate software, electronics, actuators and sensors which are connected in a network in order to achieve easier and faster information exchange. The idea of IoT is to primarily overcome the gap between the physical and the computer world, where physical objects and devices would be integrated into a seamless information network and physical objects could be active participants in the processes of information exchange. Such advanced technology offers the opportunity to connect people, systems and processes and finds huge application in all spheres of modern life.

IoT devices are used in modern cars with built-in sensors, in modern medicine (for example, for monitoring the work of the heart), the chipping of animals, city transport, video surveillance, environmental monitoring, various field operations that help firefighters save people, or in modern households, to name a few applications. But the emergence of more applications and their availability for users makes them more vulnerable to a wider range of security weaknesses [4] [5].

II. IOT AND SMART HOMES

A. Internet of things

The term Internet of Things was first used in 1998 to distinguish itself from other related terms, but also to present the vision of IoT is a network of networks, through which a large number of objects, devices, objects and sensors are connected while requiring minimal human intervention. IoT enables physical devices, vehicles, home appliances and other objects that are characterized by electronic, software, sensory, actuarial, and connectivity to exchange information [4] [6].

A feature of each IoT device is that it has a built-in unique computer system that is capable of working internally within the existing Internet infrastructure. The number of such devices on the Internet is constantly increasing. According to the assumptions of experts and researchers in the field, by 2020 the number of such devices and facilities will be as much as 30 billion, and the global market value of IoT will reach 1.7 trillion dollars [1] [3] [6] [7] [8].

IoT enables devices and facilities to be managed and controlled remotely via an existing network infrastructure, creating greater opportunities for direct integration of the physical and computer world. All this is done to facilitate and improve the efficiency and accuracy of reduced human intervention, as well as economic benefits. IoT is also an example of a more general class of cyber-physical systems that encompass technologies applicable from smart networks, virtual centers, smart and modern homes and cities, and intelligent information transport [6] [9] [1] [10].

IoT has a huge and wide range of uses in almost all areas of contemporary living. Because of this, IoT devices are characterized as a mixture of information, hardware, software and service. However, the main function of IoT devices is the collection and dissemination of useful information through different technologies among different devices [1] [10].

B. *IoT application in homes*

The application of IoT technology and the solutions resulting from the use of such devices in the home make the place of living more comfortable and pleasant. IoT technology and devices occupy a large part of modern homes, primarily in the field of automation. This system provides its users an opportunity to control all devices in the home including lighting, heating, air conditioning, security systems, home maintenance etc. The benefit of IoT devices is based primarily on ease of use, connectivity and functionality. In addition, they contribute to long-term benefits and create an ecological home by automating some functions such as switching off lights or electronics. According to a large number of people, one disadvantage associated with the use of this technology in homes is certainly the high price [11] [12] [13].

IoT devices can be used to monitor and control mechanical, electrical and electronic systems used in different types of homes (for example, houses, apartment buildings or institutions of different types). Some of their areas of application are:

- integration of the Internet with energy management systems in order to create energy efficient and smart homes managed by IoT,

- applications for monitoring energy consumption and monitoring the behavior of visitors in the home,

- integration of smart home appliances with future applications,

- providing assistance in the operation of people with disabilities and older people, etc.

Devices that are connected in the system can be managed even from a distance, so, for example, the air conditioner in a home or office can be switched on to operate at a certain temperature through the mobile phone before the person arrives there. The idea is that through these applications users can do more things at the same time. IoT technology thus contributes to providing consumers with a greater quality of life [13] [14].

Smart technology of this type is in fact a generic platform that consists primarily of hardware devices, sensors and software applications. Information is collected through the sensors and injected into the applications, from which the appropriate action and solution originated. For example, a water sprayer placed in the yard for irrigation of grassy surfaces can recognize the rain and switch off to save energy.

There are many examples of IoT applications and we will mention some of them in the remainder of this section.

Much of the IoT technology refers to virtual help. The Ubi application acts as a voice activated by the computer, and can perform tasks related to reading, voice memos, performing notifications on certain events, audio calendar, e-mail, and so on. The application uses a microphone and speakers and also has sensors for monitoring the environment (temperature, lighting, air pressure and humidity).

The Netatmo application determines the air quality and offers solutions for smart homes. In order to determine the air quality, it collects information about the temperature, humidity and amount of CO_2 in the air. This application sends a warning to the user when needed [4].

WeMo is another type of application from the wide range of IoT technology used to turn on or off electronic devices from anywhere. For this purpose, it uses Wi-Fi, but it can also be set up automatically (for example, turn on devices at sunrise) [4].

The Lockitron application is used to lock the door through the phone, and the user can authorize members of the family so they can unlock them via their phones using the Internet [4].

Blufitbottle is a drinking water bottle that records the user's hygiene habits but also remembers the hydration [4].

III. PROTECTION OF SMART DEVICES WITH VPN AND IOT GATEWAY

According to many researchers, the biggest disadvantage of IoT technology is the lack of technical standards (hardware and software variations among devices that are connected). The IoT's amorphous computational nature is also a security issue, since operating system kernel errors often deter users. Other researchers believe that IoT is becoming a "powerhouse" tool, which creates a supervisory society where technology is routinely used. According to them, people are gradually losing control of their own lives and are driven by sensors and self-managing devices. In addition, the American Civil Liberties Union has expressed concerns about the IoT, believing it impedes people's control over their own lives [15] [16] [17] [18].

However, for IoT users, the greatest threat is privacy protection. While smart technology can reduce or eliminate human intervention, it also increases the potential for hacking or major crashes. Certain problems can also arise by generating a large number of unnecessary information that will make smart devices produce misguided conclusions [17] [18]. The overall understanding of IoT is essential to the basic security of users. A growing number of surveys are focused on the IoT invasion and the threats arising from the application. Many consumers are willing to give up smart technology to preserve their privacy and security [19] [20] [21].

The great concern is that the communication channel in IoT technology is not only realized between the person and the machine, but also between the machines. And, in these circumstances, the guarantee of control, access, authorization, privacy and protection is a major problem. Accordingly, data security in the devices itself is necessary, but also security when transmitting messages from one device to another [6].

Security challenges can generally be divided into three groups [4]:

- system security;

- network security;

- security of the IoT application.

The research presented in this paper is concerned about network security and proposes a solution that is based on a VPN (Virtual Private Network) gateway.

The main reason for choosing an open VPN over PPTP is in this project is that a PPTP protocol is not completely secured, meaning it is limited to:

- MPPE-128 encryption, using RC4 encryption with a 128 bit key
- MS-CHAPv2 authentication-using SHA-1
- strong passwords (minimum 128 bits of entropy)

The cryptosystem used for VPN encryption is RSA 2048, using certificate file.

A VPN gateway is a type of network device that connects two or more devices or networks to a shared infrastructure. Its role is to enable connection (communication) among multiple devices or networks located in different locations. In other words, it can be said that VPN gateway is a virtual private network that provides a private and encrypted channel through which communication is accomplished, as can be seen in Fig. 1. With the development of technology, this type of configuration develops, becoming more accessible for every type of computer, phone or tablet. VPN-enabled applications provide users with secure Internet access. VPN is used for secure connectivity by creating a cohesive network, avoiding space constraints and security-related issues in the transmission of information [5] [22] [23].

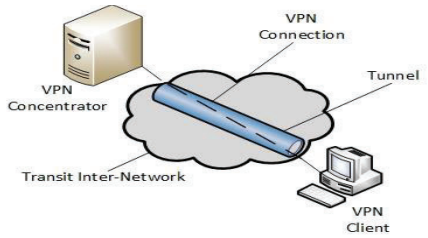


Fig. 1. A VPN network

VPN gateway can be a router, server, firewall, or similar device with data transfer capabilities. The VPN device is easily applicable to most operating systems on any computer, phone or tablet. Most often this device is a router [23].

The VPN system is created by establishing a point to point connection through the use of virtual protocols and tunneling links or encrypted data traffic. The part of the link in which private messages are entered is called a tunnel. Messages transmitted through the tunnel are encrypted, a process that is professionally referred to as a virtual private network connection. This system provides a number of benefits and access from a distance. Computer designers are constantly working to perfect the system to enable full customer support [24].

Although the VPN cannot completely save the user's anonymity, it can increase privacy and security. For this purpose, it uses authentic remote access using tunneling protocols and encryption techniques. The VPN security system provides [24]:

- confidentiality through encrypted data,

- authentication of the sender so that unauthorized users cannot access the network,

- detection of all intrusions through direct messages to the user.

The configuration line for VPN setup are as follows:

```
ca /etc/openvpn/ca.rsa.2048.crt
auth-user-pass /etc/openvpn/login
crl-verify /etc/openvpn/crl.rsa.2048.pem
```

Generation of certificate and private key for the server is issued with:

```
./build-key-server server
```

The VPN server is using a PKI (public key infrastructure). There are two important things to be addressed for the PKI:

- Public key, (a separate certificate), and a private key for the server and client.
- Master Certificate Authority (CA) certificate and key used to sign certificates for the server and client.

The server also supports a two way authentication, which is based on certificates. The client will be authenticated with the server's certificate, while the server must authenticate the client's certificate.

The server and the client, both will authenticate the other one with verification that the presented certificate was previously signed by the "master certificate authority" (CA), and then testing information in the authenticated header will be conducted.

The clocks on the server and a client must be in the same sync, or certificates will not work.

This security model has a number of desirable features from the VPN perspective:

- The server owns a certificate or key.
- The server will accept connection only on whose certificates are signed by the master CA certificate. The server can sign verification without access to the CA private key. This gives freedom to store CA key on separate server.
- If in some way, a private key is compromised, its certificate can be added to a "certification revocation list" (CRL). With this, a CRL will reject the compromised certificates.
- Client specific access right can be used, based on embedded certificate fields.

Therefore, the design of the VPN meets most security objectives: confidentiality, integrity and authenticity [24].

Using LZO compression is optional with the VPN server. This compression is fast, and 1 byte per packet for incompressible data may be added. The default setting for the VPN server is adaptive compression. With this compression, the VPN server will sample the compression process in some predefined time for measuring its efficiency. If the data is

already compressed, the efficiency of the compression will be low, giving the server option to disable the compression for a period of time, until the next cycle of testing.

DNS security can be used for enhancing the VPN server, but not as a primary protection, because the client could use own DNS settings, and bypassing the secured DNS, placed on the router or internet gateway.

IoT gateways are compact, smart and secure products that are part of the network connection. An IoT gateway contributes to bridge the gap between smart devices in the home and the place (user equipment such as phone, computer, tablet) where information is manipulated and stored and they can use wireless or wired local networks (LAN, WiFi, 3G, Zigbee and RF) [25].

IV. Raspberry Pi as a VPN Gateway

Raspberry Pi is a very small (credit card sized) computer that plugs into a computer monitor, keyboard and a mouse. This device is using Raspbian, which is a Debian-based computer operating system especially designed for Raspberry Pi [26]. For security reasons, applying security patches and updates is necessary.

In Fig. 2, a Raspberry Pi model 3B is presented. The pin-out mapping in Raspberry Pi 3B model is shown in Fig. 3.

Fig. 2. Raspberry Pi model 3B

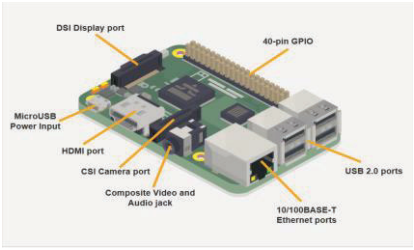


Fig. 3. Raspberry Pi 3B pin-out mapping [26]

For better security, it is recommended to create a separate user –working with "root" privileges is not recommended. A newly created user `iotuser` is added to the list of users with root privileges with the following command:

```
iotuser ALL=(ALL) ALL
```

The network topology designed is displayed in Fig. 4, where all network elements are shown.

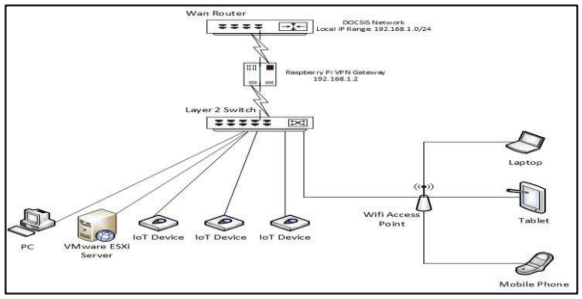


Fig. 4. Network topology

The local network setup of the DOCSIS (Data Over Cable Service Interface Specification) router used in the project is as follows:

- WAN Router IP: 192.168.1.1
- Subnet Mask: 255.255.255.0
- DHCP Range: 192.168.1.2-192.168.1.254

For easier setup, the Raspberry Pi is given a static address, right after the WAN router, 192.168.1.2.

For the VPN encryption to work properly, the NTP (Network Time Protocol) [27] is needed. If the time is not correct, the server will reject the client. The checking of synchronization with NTP servers is done with the command ntpd –p.

Next, other network devices are configured such that their default gateway is set to be the Raspberry Pi. The IP address of the gateway is the same as the IP address of the Raspberry Pi, i.e., 192.168.1.2, and public DNS servers from Google 8.8.8.8 and 8.8.4.4 are used. Other secure DNS servers can be used, ex. Comodo Secure DNS [28]. Also, 192.168.1.2 can be used in DNS servers as well.

It is recommended to install dnsmasq [29] in the Raspberry Pi device to ensure that all DNS traffic goes through the VPN. With dnsmasq, the Raspberry Pi device will accept DNS requests from all local LANs and then will forward requests to the external DNS servers.

Another security measure that can be used is blocking of all local LAN Internet access if the VPN goes down, so no device could have inbound or outbound traffic. Specifically, the traffic will not be routed through the existing Internet connection if it is unprotected and thus exposed to security risk. This is accomplished by using iptables [30]. The commands bellow illustrate how the iptables configuration is performed:

```
sudo iptables -A OUTPUT -o tun0 -m comment --comment "vpn" -j ACCEPT

sudo iptables -A OUTPUT -o eth0 -p icmp -m comment --comment "icmp" -j ACCEPT

sudo iptables -A OUTPUT -d 192.168.1.0/24 -o eth0 -m comment --comment "lan" -j ACCEPT

sudo iptables -A OUTPUT -o eth0 -p udp -m udp --dport 1198 -m comment --comment "openvpn" -j ACCEPT

sudo iptables -A OUTPUT -o eth0 -p tcp -m tcp --sport 22 -m comment --comment "ssh" -j ACCEPT
```

```
sudo iptables -A OUTPUT -o eth0 -p udp -m udp --
dport 123 -m comment --comment "ntp" -j ACCEPT

sudo iptables -A OUTPUT -o eth0 -p udp -m udp --
dport 53 -m comment --comment "dns" -j ACCEPT

sudo iptables -A OUTPUT -o eth0 -p tcp -m tcp --
dport 53 -m comment --comment "dns" -j ACCEPT

sudo iptables -A OUTPUT -o eth0 -j DROP
```

Connecting one Raspberry Pi with another would improve the IoT network security even more. This is very convenient if there are two remote locations that need control of IoT devices. For example, a user that owns a summer house would benefit from interconnecting the home network with the summer house network. A VPN tunnel can be opened among the VPN gateways, connecting the two Raspberry Pi devices. With that, full access control can be specified on both locations, thus enabling secure communication among network devices. This can be achieved with static public IP addressing of the routers.

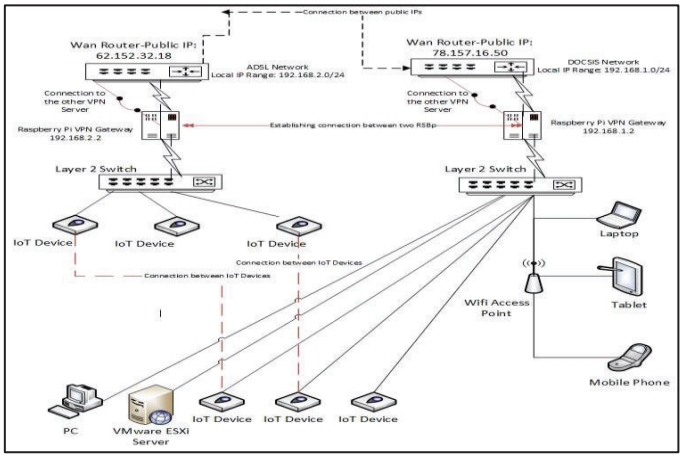


Fig. 5. A Concept of interconnecting two VPN Gateways using Raspberry Pi devices

Mainly, the performance of the VPN server depends on the hardware used. The throughput is limited to the speed of the CPU. High performance CPU e.g. multi-core CPU can handle more multiple VPN request at the same time.

Raspberry Pi is a cheap solution for setting up a VPN server. This solution can be limited, but it will be sufficient for most home and small business needs, where users will feel no difference that they are using VPN instead of a direct connection.

V. CONCLUSIONS

The ability to quickly and easily transmit information is an important component of human life. The functionality of IoT is of paramount importance. The goal is to create a "better world" in which the physical world will know what we want, what we like and what we need without further instructions.

The smart home can be defined as a space equipped with computers and information technology that meets human needs, which contribute to greater comfort and fun. Despite all the benefits offered by the "smart" home, it is still a rare phenomenon. But, IoT technology is an inevitable part of life in the future.

There are many reasons for which users would hesitate to accept the concept, security and privacy being among the foremost. One way to increase the security and privacy of users is by using a VPN and IoT Gateway, which was the method applied in this project. The main outcome is the understanding of the security threats and risk of using IoT devices on the Internet, the needs of such a device, its performance and the role it plays in protecting the devices in the network.

Further work will focus on detailed testing of the system in order to determine whether and how much the offered solution with Raspberry Pi as a gateway to a virtual private network will provide security protection.

REFERENCES

[1] Vermesan, O., Friess, P., "Internet of Things: converging technologies for smart environments and integrated ecosystems", ISBN: 978-87-92982-73-5, River Publishers, 2013.

[2] Haller, S., Karnouskos, S., and Schroth, C., "The Internet of Things in an enterprise context", https://doi.org/10.1007/978-3-642-00985-3_2, Springer, Berlin, Heidelberg, 2009.

[3] Lochab, K., Yadav, D. K., Singh, M., and Sharmab, A. "Internet of Things in cloud environment: services and challenges", International Journal of Database Theory and Application Vol.10, No.5, 2017, pp.23-32.

[4] Perera, C., Liu, C. H., and Jayawardena, S., "The emerging Internet of Things marketplace from an industrial perspective: a survey", IEEE transactions on emerging topics in computing, arXiv:1502.00134v1, 31 Jan 2015.

[5] "An Internet of Things", accessible at https://www.postscapes.com/internet-of-things-examples/ (last accessed on October 4, 2017).

[6] Vongsingthong, S., Smanchat, S., "Internet of Things-a review of applications & technologies", accessible at https://www.researchgate.net/publication/308711274_INTERNET_OF_THINGS_A_REVIEW_OF_APPLICATIONS_AND_TECHNOLOGIES (last accessed on March 18, 2018).

[7] Hsu, C. L., Lin, and J. C. C. Lin "An empirical examination of consumer adoption of Internet of Things services: network externalities and concern for information privacy perspectives", doi:10.1016/j.chb.2016.04.023, Published online April 2016.

[8] Kang, W. M., Moon, S. Y., and Park, J. H., "An enhanced security framework for home appliances in smart home", doi:10.1186/s13673-017-0087-4, Published online: March 2017.

[9] "Internet of Things: science fiction or business fact?" accessible at http://www.locate-now.com/tags/Harvard%20Business%20Review.pdf (last accessed on March 18, 2018).

[10] Mattern, F., Floerkemeier, C.,"From the Internet of Computers to the Internet of Things", https://doi.org/10.1007/978-3-642-17226-7_15, Springer, Berlin, Heidelberg, 2010.

[11] "Internet of Things (IoT)", accessible at http://www.gatewaytechnolabs.co.uk/internet-things (last accessed on September 27, 2017).

[12] Harper, R., "Inside the smart home", ISBN 1–85233–688–9 Springer-Verlag, London Limited, 2003.

[13] Demiris, G., Hensel, B. K., "Technologies for an aging society: a systematic review of smart home applications", IMIA and Schattauer GmbH, 2008, p.33-40.

[14] Mulvenna, M., Hutton, A. Coates, V. Martin, S. Todd, S., Bond, R.,and Moorhead, A., "Views of caregivers on the ethics of assistive

technology used for home surveillance of people living with dementia", doi: 10.1007/s12152-017-9305z, Published online: January 2017.

[15] La Diega, G, N., Walden, I., "Contracting for the 'Internet of Things': looking into the nest", Research Paper No. 219/2016. SSRN 2725913, February 2016.

[16] Thomas, D. R., Beresford A. R., and Rice, A., "Security metrics for the android ecosystem", accessible at https://www.cl.cam.ac.uk/~drt24/papers/spsm-scoring.pdf (last accessed on December 3, 2017).

[17] "Panopticon as a metaphor of the Internet of Things – Why not? But if it were the opposite?" accessible at https://www.theinternetofthings.eu/sites/default/files/Rob%20van%20Kranenburg/Panopticon%20as%20metaphor%20for%20the%20IoT_GS%20Dec2011.pdf (last accessed on September 27, 2017).

[18] "The societal impact of the Internet of Things", accessible at http://www.bcs.org/upload/pdf/societal-impact-report-feb13.pdf (last accessed on November 2, 2017).

[19] "Disruptive civil technologies", accessible at https://www.hsdl.org/?abstract&did=485606 (last accessed on April 30, 2017).

[20] "Igniting growth in consumer technology", accessible at https://www.accenture.com/_acnmedia/PDF-3/Accenture-Igniting-Growth-in-Consumer-Technology.pdf (last accessed on September 18, 2017).

[21] Aleisa, N., Renaud, K., "Privacy of the Internet of Things: a systematic literature review (extended discussion)", arXiv:1611.03340, Available online: September 2016.

[22] McEwen, A., Cassimally, H., "Designing the Internet of Things", ISBN 978-1-118-43063-7, John Wiley and Sons, Ltd, 2014.

[23] "The ABC's of VPN Configuration", accessible at https://www.techopedia.com/2/30433/networks/the-abcs-of-vpn-configuration (last accessed on January 3, 2018).

[24] "Virtual Private Networking: An Overview", accessible at https://docs.microsoft.com/en-us/previous-versions/windows/it-pro/windows-server-2000/bb742566(v=technet.10) (last accessed on March 18, 2018)

[25] "Internet of Things gateway solutions", accessible at http://www.supermicro.com/products/system/compact/ (last accessed on January 26, 2018).

[26] "Raspberry Pi 3 model B", accessible at https://raspberry.piaustralia.com.au/raspberry-pi-3-model-b (last accessed on August 29, 2017).

[27] "Network time protocol-NTP", accessible at http://www.ntp.org/ (last accessed on March 5, 2018).

[28] "Comodo secure DNS", accessible at https://www.comodo.com/secure-dns/ (last accessed on February 7, 2018).

[29] "Dnsmasq", accessible at http://www.thekelleys.org.uk/dnsmasq/doc.html (last accessed on March 14, 2018).

[30] "IPtables", accessible at http://ipset.netfilter.org/iptables.man.html (last accessed on January 13, 2018).

Identity Provider Management System for University Services

Kostadin Mishev, Aleksandar Stojmenski, Goran Petrevski, Boro Jakimovski, Vesna Dimitrova, Ivan Chorbev and Ivica Dimitrovski
"Ss. Cyril and Methodius" University in Skopje
Faculty of Computer Science and Engineering
"Rugjer Boshkovikj" 16, 1000 Skopje, Republic of Macedonia
{kostadin.mishev,aleksandar.stojmenski,goran.petrevski,boro.jakimovski,vesna.dimitrova,ivan.chorbev,ivica.dimitrovski}
@finki.ukim.mk

***Abstract*—The need of central user identity management repository is emerging in all university that represents multiple functional communities, especially in providing e-services that facilitate the administrative work. In this paper, we will provide an overview of the process of implementation and setup of the Identity Provider Protocol in the Ss. Cyril and Methodius University in Skopje. We will summarize the challenges and difficulties that we overcome in implementation and integration with existing software platforms that are already established in the University.**

Keywords: University Authentication System, Eduroam, User Repository, Cloud services.

I. Introduction

System integration is a complex process where a cohesive platform is created from components that were not specifically designed to work together. Components of an integrated platform are often stand-alone systems that operate on different computer environments. This paper describes the platform integration of various applications and their interconnection and inter-dependability. In order to collaborate between each other, these applications need to specify suitable protocols for exchanging data, as well as protocols for flow control. Usually, these application are sharing the same user roles and same authentication credentials. This paper gives an overview of the system for user management between the multiple applications for faculty services. For this purpose, different design patterns are used to facilitate the communication between systems and to harmonize endpoint data formats that this paper examines.

Different problems and their possible solutions are presented, regarding systems lifecycle, architecture, process, interface, synchronization and security. Each application endpoint exports security protocols functionalities for system authentication and authorization. Following the principles of the identity provider, each server and the eduroam access service authenticates the users using bearer tokens. The eduroam is a secure, world-wide roaming access service developed for the international research and education community. In our implementation of eduroam, a technical infrastructure which issues and revokes credentials and certificates is setup. Moreover, a Radius server is setup which verifies access credentials and subsequently grants access to eduroam and to the other applications within our network.

II. Background work

Although a lot of work and progress has already been done in the area of web services in the past years, efforts have been mostly focused on service description models and languages, and on automated service discovery and composition [1]. The term Web services is used frequently nowadays, although sometimes it is very ambiguous. Existing definitions of the terms vary from generic to specific and restrictive. One definition is that a Web service is seen as an application accessible to other applications over the Web [7]. This is a very open definition meaning that anything with a URL address is a Web service. It can include a CGI script or refer to a program accessible over the Web with a stable API, published with additional descriptive information on some service directory. A more precise definition is provided by the UDDI consortium, which characterizes Web services as "self-contained, modular business applications that have open, Internet-oriented, standards-based interfaces" [5]. This definition is more detailed, placing the emphasis on the need for being compliant with Internet standards. In addition, it requires the service to be open, which essentially means that it has a published interface that can be invoked across the Internet. In spite of this clarification, the definition is still not precise enough. For instance, it is not clear what it is meant by a modular, self contained business application. A step further in refining the definition of Web services is the one provided by the World Wide Web consortium (W3C), and specifically the group involved in the Web Service Activity: "a software application identified by a URI, whose interfaces and bindings are capable of being defined, described, and discovered as XML or JSON artifacts.

It is a standard requirement that the web application require some information about the users. This information is provided by the Service provider [3] and it is typically used to provide persistent user profiles, but can be also useful for cross application authentication and accounting. The full identity information about the user is not always required. Another type of identifier or token is mandatory, which remains the same for

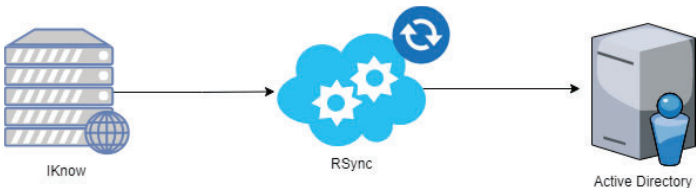


Fig. 1: Synchronization process between iKnow and UKIM Active Directory

any user among the multiple authentication endpoints. In [4], the authors present a system for user management that solves the IdM problem which is stated in [2]. Different models are proposed including an agent-based delegation model for secure web services in ubiquitous computing environments based on Security Assertion Markup Language (SAML)[8]. To develop our faculty identity service provider, we took into consideration the state-of-art techniques and methods applied in the solutions above.

III. System Architecture

The team from The Faculty of Computer Science and Engineering extended its strives in development and adaptation of framework for development of e-services for students and staff for faculties and institutions within the University Ss. Cyril and Methodius in Skopje [6] [9]. The University does not implement any unique user identification provider nor unique database user management solution for students and staff. Due to these necessities, it is considered that it should be established a system that will provide services for central user management which can be reused as identity provider. Such identity provider can provide facilities for user authentication and authorization of appropriate e-services which will be able to act as a complete federation of services. As user management database, it is used Microsoft Active Directory that is hosted in Faculty of Computer Science and Engineering. For each faculty (institution) it is created appropriate organization unit (OU) that manages its users, students and staff separated in different groups. The biggest challenge that we had to overcome was an implementation of consistent user synchronization job process. Such job updates the user profile information about currently active users in all institutions and organization units. iKnow, as University platform that keeps information about the staff and students for most of the faculties that are part of the University, may be used as a human resources system. User synchronization should be done periodically due to the frequent change of the status of the users. For this purpose, it is developed an additional service called RSync Fig. 1 that runs periodic jobs just to synchronize the state of the iKnow users database with the user profiles created in Active Directory. First at all, it checks the new users that may be created meanwhile in iKnow database. Afterwards, it deactivates the users in AD that have been deactivated in iKnow. At the end, it updates the changed attributes of the users.

The attributes that are exchanged between AD and iKnow are the following: name, surname, affiliation, e-mail and

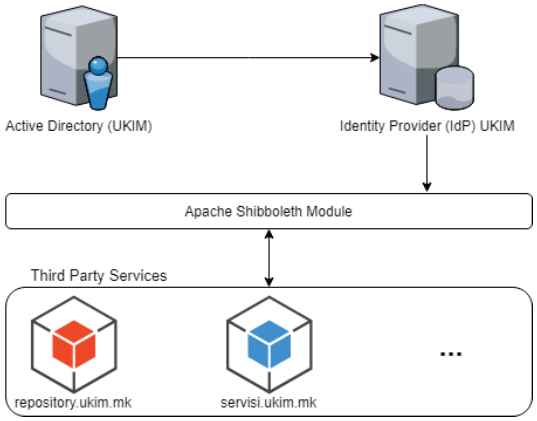


Fig. 2: IDP Architecture

activity status. Passwords are kept in encrypted format in Active Directory. Also, they are not transferred over the synchronization process. Due to this, the password should be updated directly by the user by using appropriate password reset form. An identity provider (IdP) is a system entity that creates, maintains, and manages identity information for principals while providing authentication services to relying party applications within a federation or distributed network. It offers user authentication as a service which can be reused as a Single sign-on (SSO) module for third-party applications used within the University 2. SSO enhances the user experience by reducing the password fatigue phenomena by increasing the user security and providing one profile per user for all services provided by the University, reducing the problem to remember an excessive number of passwords.

IdP uses the users' database provided by the synchronized Active Directory server. It acts as an additional layer whose role is to approve or decline the user that tries to authenticate. The University IdP is configured to use user's e-mail and password as authentication attributes. After the successful authentication, it provides access to the name, surname and affiliation attributes of the user to the service provider.

First at all, the user should determine his e-mail address, in this case the University e-mail, by using the appropriate form embedded in the main login page of the IdP. Afterwards, he has to reset the password just to receive a link to his e-mail address whereas he will be able set the password. After this step, he is able to use the IdP authentication service to login to the services that implement the IdP authentication.

The authentication process between the IdP service and service providers is enabled by using the Shibboleth software package. The Shibboleth software implements widely used federated identity standards, principally the OASIS Security Assertion Markup Language (SAML), to provide a federated single sign-on and attribute exchange framework. To facilitate its usage by third-party service providers, the Shibboleth Apache module is installed as an edge server. It authenticates the user that tries to enter to the services provided by the University. After the successful authentication, the user is redirected to the requested service providing the default attributes required by the service provider. Currently, the first

two services that work within the University use the described authentication method are:

- repository.ukim.mk – Central repository for scientific and art work management intended for University Ss. Cyril and Methodius
- servisi.ukim.mk – Central dashboard for e-services availability management intended for students and staff within University Ss. Cyril and Methodius.

The user management provided by using the appropriate AD within the University, provides simplification in integration of the EduROAM as a service within all faculties and institutions within the University. FreeRadius server provides simple integration with AD and supports all EAP flavours commonly used for user authentication in eduroam (EAP-PEAP, EAP-TLS, EAP-TTLS-PAP, EAP-TTLS-MSCHAPv2). The user can easily authenticate to EduROAM by using the same user credentials as described in the previous section 3.

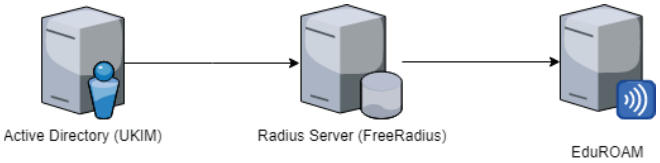


Fig. 3: UKIM EduROAM Architecture

IV. Conclusion

Faculty of Computer Science and Engineering continues the development of multiple platforms for student and staff services by providing new e-services and their adaptation with machine interfaces to the central data repository. Such services provide simplification and acceleration of the Faculty administrative workflow by providing easy-to-use interfaces avoiding congestion and bottle-neck scenarios.

In this paper, we present the usage of identity management providers as a part of the global faculty e-platform, interconnections and interoperability between various applications. The complete architectural design and implementation is covered in this paper. Also, a complete scenario is described which explains the inter-service communication among multiple data service providers including process of identity user management, user verification and reusability of the authentication tokens which are enabled by using the proposed faculty service architecture.This service architecture is able to address the Identity Management problem among the faculty applications and show how it can be successfully applied to manage the authentication and roles needed among different applications.

References

[1] B. Benatallah, F. Casati, D. Grigori, H. R. M. Nezhad, and F. Toumani, "Developing adapters for web services integration," in *International Conference on Advanced Information Systems Engineering*. Springer, 2005, pp. 415–429.

[2] E. Bertino, F. Paci, R. Ferrini, and N. Shang, "Privacy-preserving digital identity management for cloud computing." *IEEE Data Eng. Bull.*, vol. 32, no. 1, pp. 21–27, 2009.

[3] A. Busboom, R. Quinet, M. Schuba, and S. Holtmanns, "Method for authenticating a user to a service of a service provider," Mar. 9 2006, uS Patent App. 10/513,212.

[4] A. Celesti, F. Tusa, M. Villari, and A. Puliafito, "Security and cloud computing: Intercloud identity management infrastructure," in *Enabling Technologies: Infrastructures for Collaborative Enterprises (WETICE), 2010 19th IEEE International Workshop on*. IEEE, 2010, pp. 263–265.

[5] U. Consortium *et al.*, "Uddi executive white paper, nov. 2001 http://uddi.org/pubs," *UDDI_Executive_White_Paper. pdf*.

[6] K. Mishev, A. Stojmenski, I. Dimitrovski, V. Dimitrova, and I. Chorbev, "Cloud services for faculty workflow automatization," 2017.

[7] M. P. Papazoglou, "Service-oriented computing: Concepts, characteristics and directions," in *Web Information Systems Engineering, 2003. WISE 2003. Proceedings of the Fourth International Conference on*. IEEE, 2003, pp. 3–12.

[8] N. Ragouzis, J. Hughes, R. Philpott, E. Maler, P. Madsen, and T. Scavo, "Oasis. security assertion markup language (saml) v2. 0 technical overview," *Committee Draft*, vol. 2, 2008.

[9] A. Stojmenski, V. Nikolovski, I. Chorbev, and V. Dimitrova, "Cross platform system integration using web services," 2016.

Can a Blended Learning Environment Increase the Quality of Learning?

Maja Videnovik
Center for innovations and digital education DIG-ED
Skopje, Macedonia
majavidenovik@gmail.com

Gjorgjina Dimova
Center for innovations and digital education DIG-ED
Skopje, Macedonia
gina_mkd@yahoo.com

***Abstract*—Blended learning environments combine different pedagogical approaches, integrating the best aspects of face-to-face and online collaboration using ICT. Innovative learning environments increase students' interest and motivation for achieving learning outcomes. This paper elaborates how a blended learning educational scenario can contribute to the quality of learning. Quality of Learning is a combination of quality of experience while learning and results of achieved learning outcomes. The quality of experience while learning is determined with students' attitudes toward learning and achieving learning goals. Results of the achieved learning outcomes are determined using students' grades. The methodology used in the paper combines qualitative and quantitative analyses in order to determine the influence that blended learning environment has on achieved Quality of Learning. The blended learning environment is created using the educational networking site – Edmodo. The blended learning environment was supported by development of a blended learning educational scenario for engagement, guiding students in the learning process and achieving educational goals.**

Keywords—learning environment, blended learning, technology acceptance, social network in education

I. INTRODUCTION

Contemporary education is trying to develop students' 21st century skills necessary for living in this constantly changing world. The traditional education where space and the time were usually strictly controlled and restrained is slowly fading away.

Education is an interactive process during which students gradually adopt knowledge, develop intelligent reasoning, judgment and skills. The student-centered educational approach emerged as an alternative to the traditional teacher-centered education, by bringing the student's engagement with the educational activities adapted to their individual abilities and interests in the center of the learning process. At the same time, students motivation to collaborate in order to achieve certain educational goal is increased. This approach promises to produce multiple benefits regarding academic performance, as well as behavioral gains such as students' retention in the education, responsibility, and development of transferable skills like collaboration, communication, problem solving etc. The gradual shift from the traditional learning environment towards student-centered learning environment stimulates understanding above pure memorization of the educational content leading to increased quality of educational outcomes.

If we want to empower our students' skills for the 21st century jobs we must put them in the center of the teaching, encouraging them for active participation in the classroom, collaboration and getting information on their own which can be used in their further learning. The teacher should just guide students' activities and continuously monitor students' work giving them constructive feedback and direction for their further learning. On this way students will acquire a lot of knowledge, skills and attitudes that are needed to function independently in a world of change.

Schools must promote "learning to learn", acquisition of knowledge and skills for continuous learning over lifetime. The different kind of pedagogical approaches should be implemented in order to raise students' interest and motivation and to make them active participants in the learning process. The learning must happen according to students' needs and opportunities, exceeding the classroom boundaries; it. Today's students are complex, energetic and tech-savvy individuals. Technology is an integral part of their lives, so if we want to motivate and inspire them the same surroundings should be used. Educators must recognize the potential for improving student engagement in the classroom using different technologies [1].Technology must be used as a supplement of face-to-face learning in order to create innovative learning environments where students would like to be active participants.

Blended learning has become a "hot topic" in the educational world in the recent years. The combination of face-to-face instructions and online teaching allows individualization, flexibility and greater student success. In blended learning, technology is used to serve multiple learning styles or needs, engage the learners, and prepare the students for life after school. There are some choices of platforms that can be used for blended learning in the learning process at school, which among others are Edmodo, Moodle, Kelase, etc. [2].

Edmodo as a blended learning environment supports the learning and networking of teachers and students. This free and secure learning platform provides a simple way for teachers and students to connect and collaborate in a virtual class. This social networking geared towards the needs of students could have a profound impact on how students collaborate and learn in their world, rather than the school setting their teachers grew up in [3].

In the paper, the development of blended learning educational scenario is elaborated and the benefits from this pedagogical approach are described. The main purpose of this study is to investigate the Quality of Learning achieved using

blended learning environment created with Edmodo online platform. For that purpose, the quality of learning is defined as a combination of the quality of experience while learning, and the results of achieved learning outcomes. The quality of experience while learning is determined with the students' attitudes toward learning and achieving learning goals. The results of achieved learning outcomes are determined using students' grades.

The paper is organized as follows: next section, Section 2 presents the related work. The used methodology is present in Section 3, while results and discussion are given in Section 4. Section 5 concludes the paper.

II. Related Work

Today's education is trying to overcome traditional teaching, where there is one classroom, one teacher, one class and where students are passive listeners. The education must adapt to students and their way of learning, empowering students 21st century skills. The gap between the skills people learn and the skills people need is becoming more obvious, as traditional learning falls short of equipping students with the knowledge they need to thrive [4]. Students want to be challenged and inspired in their learning, to collaborate and work with their peers, to incorporate technology they love into their classroom experience as much as they can.

There is a constant urge in the educational community to promote new technologies that will add additional value to education [5]. The learning process should be supported with carefully created synchronous and asynchronous learning events [6]. The main benefit of using distance learning educational systems and different learning management systems is enabling teachers and students to be virtually "present" in an environment that they cannot physically reach because of lack of different resources [7]. In addition, different real [8] and virtual resources [9] can be made available for students. The impact and effect of the media on the student's satisfaction and engagement in the learning process is subject of many studies [10], [11], [12]. The authors conclude that the asynchronous rich media presentations increase the quality of student's learning experience.

The ability of the distance education system to adapt to the learning goals of individual student, leads to improved quality of learning experience [13]. The educational gains from this approach can be measured and compared to traditional classroom teaching concept in order to evaluate the results and determine the directions for further improvement and development of the distance educational systems [14], [15].

Blended learning environments enable strategic and systematic approach to combining times and models of learning, integrating the best aspects of face-to-face and online interactions for each discipline, using appropriate ICTs [16]. In a blended course, conventional learning is supplemented with the use of proper learning technologies that create innovative learning environments that enable instructors to organize their teaching in a more efficient way [17]. Blended learning can remove deficiencies found in the learning process such as time limit of face-to-face classroom. Previous studies reveal positive results of blended learning on learners' performance [18], [19]. With the advent of technology and the ease of access that it affords, the unique environment for blended learning can be more easily and successfully created [20].

Researches show that Edmodo changes traditional classroom and transforms it into blended learning. In traditional classes, students usually work using paper and schoolbooks. By Edmodo all the information can be reached online without time and space limitation. When teacher uses Edmodo the students can do their individual work and group work online and the teacher can check the work, and give feedback to the students online [21]. While the traditional classroom may be limited to words spoken and written by instructors, and can be trapped in hard copies of the materials distributed for reference, blended learning via Edmodo goes far and beyond [22]. Shared multimedia information can be reached multiple times, according to students' pace of learning. The personal communication is what students benefit in face-to-face interaction in the classroom. But Edmodo continues the work online and it ensures that knowledge is no longer contained in places and defined by geography. The implementation of Edmodo for creation of a blended learning environment could become a powerful medium that extends responsible learning environment beyond the classroom [23].

III. Methodology

The study was carried out with total of 160 primary school students from 8 different classes.

The methodology used combines qualitative and quantitative analyses in order determine the influence that the blended learning environment created using Edmodo, has on achieved Quality of Learning. Qualitative analyses were used to determine quality of experience while learning, while quantitative analyses were used to compare results of achieved learning outcomes.

A simple blended learning educational scenario is used for the purpose of the study. Virtual groups where students and the teacher exchanged materials, information and products connected with the curriculum were created for each of the classes. The aim was to see could we use those virtual groups as a "virtual classrooms" which would allow learning to take place outside the classroom, but monitored by the teacher. The collected data were qualitative, consisting of posts on Edmodo (students' self-reflection, feedback or discussion) and the focus discussions conducted with the students. Quantitative data from students' grades were considered as results of achieved learning.

A. Creating a blended learning environment

Edmodo is a free and secure learning platform that looks similar to Facebook, but is much more private and safe because it allows teachers to create and manage accounts for their students. Once account is created, student receives a group code and than he can register, access and join the group. Students without group cannot participate in any of the group activities. The Edmodo site provides a simple way for teachers and students to connect and collaborate in a virtual class.

For each class separate group on this platform was created. The aim was to see if we could we those groups as a "virtual

classrooms" which would allow learning to take place outside the school, but monitored by the teacher.

In those groups the students and the teacher exchanged different multimedia contents (materials, information and different things of interest connected to the curriculum). Teaching materials were posted in the groups regarding the previous or the next lesson. Different tasks with the description of the activities that should be carried out, objectives that should be achieved and evaluation criteria, were shared with the students. Different quizzes and surveys were set, too.

Students created their own profiles - accounts on the corresponded group. Each student uploaded information and has been involved in the implementation of the activities given by the teacher. For each activity, students uploaded the product of their work and using given outcomes and success criteria had to reflect on the learning process and write a post about this. The teacher encouraged students in these posts to describe their work, activities that they carried out, to explain their strengths, the difficulties during the work, used ways for improvement and what else they would like to learn. In this way, students create their own e-portfolio as a reflexive diary of their learning. Their e-portfolio consisted of the products from their work according to the teacher's guidelines, their reflection of the achieved learning outcomes and ways of improvement. Teacher analyzed students' posts, their works and gave effective feedback to each student, focusing on student learning. At the same time the teacher motivated students to assess each other's work and give each other a feedback. The students had enough time to act according to the received feedback.

During group work the teacher created small groups inside every class where students could collaborate and work together on the assignment. Students' were encouraged to asses each other work and to start discussions about the level of achievement and what could be improved. Discussions were carried out both online and during the classes.

The collected data were qualitative and consist of students' posts on Edmodo, whether it refers to reflection, feedback or discussion. The performed analysis is descriptive and focuses on students' attitudes toward use of Edmodo. The students were trained for self-reflection, as a way for identification of their strengths, weaknesses and further learning directions as a base for developing lifelong learning skills.

The posts from the students were analyzed together with students' focus groups discussions in order to determine students' engagement, guidance in the learning process and their perception of achievement of educational goals. These elements were used to determine students' quality of experience while learning.

The results of the achieved learning from the students that participated to the study was compared with the results of the achieved learning from students that learned using traditional learning environment.

The quality of experience while learning and the results of achieved learning outcomes can be used to determine the quality of learning. In our approach we consider both factors to contribute equally to the quality of learning.

IV. Results and Discussions

All students successfully created their user accounts and with little or no efforts started using all Edmodo features. They used Edmodo for exchanging information with the teacher and other classmates. All students regularly logged on Edmodo, during and after lessons. Most of the students continued their learning online using the materials the teacher posted on the network. Collaboration with their peers and the teacher continued online. Even shy students participated in online discussions and gave their feedback on other students' work.

Following teacher's guidelines and success criteria all students created their own e-portfolio. Most of the students had qualitative reflections with a detailed analyzes on their learning.

The feedback from the teacher was motivating for the students and inspired them to modify their work. Students were persistent to achieve goals when they saw that the teacher monitor and guide their work. Because they had the opportunity to correct their work, they were not afraid to take risk and explore new possibilities for achieving the goals. Each student worked individually, according to the direction that he got, according to his abilities, independently from the other students.

Focus discussions carried out with students showed students' satisfaction from using Edmodo in everyday learning. They find it easy to reach teaching materials in Edmodo. Furthermore, the materials helped the majority of the students in successfully achieving of learning outcomes. The feature of uploading their work on Edmodo was considered as very useful by the students.

Students believe Edmodo to be a useful social network platform for communication between the teacher and students during the learning process and reflection on students' work. They feel very positive about the possibility of commenting on peers' posts, but also value the feedback they were given by their peers and their teacher.

One of the biggest benefits, according to students, is that they can use Edmodo outside the classroom, at a time that suits them. This is a huge benefit for learning, too. Students spend more time on platform learning, which means that they were more persistent in achieving the learning outcomes. In that way, they clearly preferred the blended learning environment over traditional classroom.

The comments from the students were very interesting: "It's like Facebook, I felt comfortable using it", "All my work is on one place and I can reach it any time I want", "I always have proof for the work I've done", "Now nobody can use my homework as his own", "Everything that is not in the book, I can find it on Edmodo, where my teacher posted it", "I like it because I can work from home, too".

Achieved results by the students that used Edmodo blended learning environment were increased by 20% compared to the achieved results from the students that took traditional classroom learning environment that stayed the same.

At the same time, the quality of experience while learning was much better for the students that participated in the study.

As conclusion, the quality of learning was much better for the students that participated in the study (used blended learning environment created by Edmodo) compared to the quality of learning among students in the traditional classroom.

V. CONCLUSION

The development of lifelong learning skills inevitably imposes the necessity of ICT integration into daily teaching as a tool for communication and collaboration that exceeds time and space limitation of traditional classroom.

By using Edmodo, blended learning environment was created and students got the possibility to progress according to their abilities, through continuously cooperation with the teacher. Students were able to explore, create, reflect on their work, organize and direct their own learning.

The results from the students' work on this platform and discussions with them showed that students find Edmodo to be user-friendly social learning platform that enables them to achieve learning outcomes on an online class. They consider that Edmodo is a useful tool for finding materials and accessing them with their own pace. They like the possibility to have evidence for their own learning and to upload their work which can be accessed and evaluated after classes. As a largest advantage student indicate flexibility in space and time when using it. They have used Edmodo for communication, cooperation, sharing information, reflection, self-evaluation and directing their own learning, by which lifelong learning skills were developed.

Achieved results and the quality of experience while learning was much better for the students that participated in the study compared with students that used traditional learning environment. As conclusion, the quality of learning can be increased using blended learning environment

REFERENCES

[1] C. Holland, and L. Muilenburg, "Supporting student collaboration: Edmodo in the classroom," Society for Information Technology & Teacher Education International Conference. Association for the Advancement of Computing in Education (AACE), 2011.

[2] I. M. Ardana, I. P. W. Ariawan, and D. G. H. Divayana, "Development of decision support system to selection of the blended learning platforms for mathematics and ICT learning at SMK TI Udayana," International Journal of Advanced Research in Artificial Intelligence, vol. 5(12), pp. 15-18, 2016.

[3] B. K. Gushiken, "Integrating Edmodo into a High school service club: To promote Interactive online communication," 18th Annual TCC online conference 2013, Hawai, USA.

[4] World Economic Forum, New Vision for Education: Fostering Social and Emotional Learning through Technology. Switzerland, (2016). Retrieved from http://www3.weforum.org/docs/WEF_New_Vision_for_Education.pdf

[5] P. J.-H. Hu, T. H. K. Clark, and W. W. Ma, "Examining technology acceptance by school teachers: a longitudinal study," Information & Management, vol. 41(2), pp. 227-241, 2013.

[6] T. Malinovski, M. Vasileva, T. Vasileva-Stojanovska, and V. Trajkovik, "Considering high school students' experience in asynchronous and synchronous distance learning environments: QoE prediction model," The International Review of Research in Open and Distributed Learning, vol. 15(4), pp. 91-112, 2014.

[7] L. Breslow, D. E. Pritchard, J. DeBoer, G. S. Stump, A. D. Ho, and D. T. Seaton, "Studying learning in the worldwide classroom: Research into edX's first MOOC," Research & Practice in Assessment, vol. 8, pp. 13-25, 2013.

[8] G. Kimovski, V. Trajkovic, and D. Davcev, "Negotiation-based multi-agent resource management in distance education," IASTED International Conference on Computers and Advanced Technology in Education including the IASTED International Symposium on Web-Based Education, pp. 327-332, 2003.

[9] G. Kimovski, V. Trajkovic, and D. Davcev, "Resource manager for distance education systems," Advanced Learning Technologies, Proceedings, IEEE International Conference on IEEE, pp. 387-390, 2001.

[10] P. A. Havice, K. W. Foxx, T. T. Davis, and W. L. Havice, "The impact of rich media presentations on a distributed learning environment," Quarterly Review of Distance Education, vol. 11(1), 2010.

[11] S. H. Liu, H. L. Liao, and J. A. Pratt, "Impact of media richness and flow on e-learning technology acceptance," Computers & Education, vol. 52(3), pp. 599-607, 2009.

[12] N. Koceska, and V. Trajkovik, "Quality of experience using different media-presentation types," Information Technology Based Higher Education and Training (ITHET), 16th International Conference on, IEEE, 2017.

[13] T. Malinovski, M. Lazarova, and V. Trajkovik, "Learner−content interaction in distance learning models: students' experience while using learning management systems," International Journal of Innovation in Education, vol. 1(4), pp. 362-376, 2012

[14] M. Allen, E., Mabry, M., Mattrey, J., Bourhis, S., Titsworth, and N. Burrell, "Evaluating the effectiveness of distance learning: A comparison using meta-analysis," Journal of Communication, vol. 54(3), pp. 402-420, 2004.

[15] T. Vasileva-Stojanovska, M. Vasileva, T. Malinovski, and V. Trajkovik, "An ANFIS model of quality of experience prediction in education," Applied Soft Computing, vol. 34, pp.129-138, 2015.

[16] G. Saliba, L. Rankine, and H. Cortez, "Fundamentals of Blended Learning," Sydney:University of Western Sydney, 2013.

[17] S. Wichadee, "A development of the blended learning model using Edmodo for maximizing students' oral proficiency and motivation," International Journal of Emerging Technologies in Learning, vol. 12(2), pp. 137–154, 2017.

[18] M. Jou, Y. Lin and D. Wu, "Effect of a blended learning environment on student critical thinking and knowledge transformation," Interactive Learning Environments, vol. 24(6), pp. 1131-1147, 2016.

[19] G. Motteram, "Blended education and the transformation of teachers: A long-term case study in postgraduate UK higher education," British Journal of Educational Technology, vol. 37(1), pp. 17-30, 2006.

[20] W. O'Leary, Blended Learning: Engaging 21st Century Students.

[21] M. A. S. Enriquez, "Students' perception of the effectivness of the use of Edmodo as a supplementary tool for learning," DLSU Research Congress 2014, De La Salle Iniversity, Manila Philippines.

[22] F. Al-Kathiri, "Beyond the classroom walls: Edmodo in Saudi secondary school EFL instruction, attitudes and challenges," English Language Teaching, vol. 8(1), 2015.

[23] K. Balasubramanian, V. Jaykumar, and L. N. Fukey, "A study on student preference towards the use of Edmodo as a learning platform to create responsible learning environment," Procedia-Social and Behavioral Sciences, vol. 144, pp. 416-422, 2014.

Introduction of 21st Century Skills in Primary Schools: Case Study Macedonia

Maja Videnovic
DIG-ED NGO
Skopje, Macedonia
majavidenovik@gmail.com

Aleksandar Karadimce
Faculty of Computer Science and Engineering
University St. Paul the Apostle
Ohrid, Macedonia
akaradimce@ieee.org

***Abstract*—The current educational system in primary schools is mostly knowledge oriented. This approach is commonly found in STEM (Science, Technology, Engineering, and Mathematics) courses, where teacher deliver the learning material and students repeat the knowledge they gain. There has been an effort to modernize the primary school education with the introduction of the Cambridge study program, but teachers still use the knowledge-oriented delivery, which does not bring improvement.**

The project activities 21st-century schools are consistent with the educational practices at the national level, as well as with the implementation of new educational standards in primary schools. Moreover, this project was realized in accordance with the current educational curriculum and the framework for 21st Century Learning. The main objective of this framework is to provide the necessary skills, knowledge, and expertise that students must gain to succeed in work and life. Therefore, the main goal of this project was to introduce new innovative teaching methods and ways to improve the specific skills, expertise and digital literacies of young people. This project covers primary schools from the Western Balkans, focusing on students aged 10 to 14 years. Currently, this project has delivered core skill for problem-solving and critical thinking, followed by training on digital literacy and practical usage of Micro: bit digital device for teachers of primary schools. In this paper, we will describe the used training methodology, the follow-up practical activities for students in primary schools and teachers satisfaction survey from the training.

***Index Terms*—primary school, critical thinking, digital literacy, problem-solving, Micro: bit**

I. Introduction

The classical traditional educating where space and the time were usually strictly controlled and restrained is slowly fading away. These traditional based concepts on the knowledge acquisition and the repetition model has provided a lot of doubt about the quality of primary education [1]. Today's modern society occupied with tech-devices and virtual social interaction has extended student's opportunities for creative learning across time and space. The primary school education approach from teacher-lead and knowledge oriented is shifting towards students-engaged and skills development. In this manner, within the twenty-first-century learning framework, teacher role is more of a facilitator who encourages the class to think and question the world around students, than educator that leads the class from the classroom front.

The introduction of the 21st-century learning framework is planned to provide the necessary skills, knowledge, and expertise that students need to survive in work and life [2]. The introduction of this framework will transform the lesson plans and will include the development of thinking skills, provide examples of applied thinking, and adaptations for diverse student needs. Teachers help students to develop higher order learning skills through the scaffolding concept. The scaffolding approach is a dynamic intervention of teacher giving students support at the beginning of a lesson and gradually requiring students to practice the skills independently [3]. This way of learning changes students towards higher-order thinking abilities to work in teams or individually and become leaders while being accountable and adaptable, making them a socially responsible. Furthermore, the higher order thinking can be practiced by examples of open-ended questions that encourage students to analyze the known facts in order to make a concrete conclusion independently. This will support students when making choices, team brainstorming, finding solutions and practicing interpersonal and self-directed skills [4].

These global effects have certainly affected the educational system in primary schools in Macedonia. There has been an effort to improve the primary education with the introduction of the Cambridge Primary curriculum framework [5]. However, teachers in our country still use the traditional knowledge oriented delivery that does not stimulate student's higher-order thinking skills. With the improvements in educational practices at the national level, as well as the development of new educational standards, the project 21st-century schools have been initiated for primary schools in the Western Balkans.

The main goal of this project is to introduce new specific core skills and improve digital literacy in the existing student's primary school curriculum. It is projected that the implementation of 21st-century learning skills stimulates logical, reflective, meta-cognitive, creative and critical thinking [6], will have their biggest impact when learning science, technology, engineering, and mathematics (STEM) courses. In this way, the student's engagement in the STEM learning activities in long term will improve the development of many soft skills, such as communication, courtesy, flexibility, integrity, interpersonal skills, positive attitude, teamwork, responsibility, and work ethic [7]. The hands-on practice for students will stimulate a process of analyzing the evidence collected in problem-solving, in order to evaluate their higher-order thinking abilities.

In this paper, Section II provides the background information on using the core skills in primary education. Next, Section III describes the used methodology in the 21st-century skills project. The project implementation and delivery of the training in Macedonia is described in Section IV. Finally, Section V describes the potential benefits of using this project in primary schools.

II. Related work

Today's education is moving from the traditional based concepts of teaching, where there is one classroom, one teacher, one class, and one subject at a time [1] toward the 21st-century learning approaches. The authors Crowl et al., in [4], have shown that biological development of the children, together with enhanced instructional techniques affects their cognitive development stages. The cognitive functions together with the knowledge structures help the assessment of student's domain-specific problem-solving ability [8]. This has been reflected by the Bloom's three taxonomies (cognitive, affective, and psycho-motor), causing the lower levels to provide a base for higher order levels of learning [9]. More exactly the higher order thinking is defined as a concept that considers critical, logical, reflective, meta-cognitive, and creative thinking [6]. It's important to note that the critical thinking and problem-solving is not something new and it has been built into the educational curriculum since the 1960s [10] but was not regularly practiced in-class.

The biggest emphasis in primary school education is given towards learning multiple STEM (science, technology, engineering, and math) courses as well as other interdisciplinary abilities. To introduce the STEM education professors have introduced the Computer Programming Principles (CPP) module within the IT subject [7]. This module has been used by 24 pupils in the 5th-grade curriculum. Authors Kintsakis and Rangoussi, in [7], have confirmed that the first group that was taught both a way of thinking to solve a given problem and a tool (programming language Scratch) to code the algorithmic solution was capable of using of both the moodle and Scratch. On the other hand, the second group that was instructed face-to-face in class at the end of the module was not ready to code programs in Scratch. However, this research has not provided detailed comparison what effect had the e-learning and the face-to-face group approaches on the learning outcomes. Another similar research has confirmed that nowadays children learn more through connection and collaboration, facilitated by technology that leads to knowledge creation [11]. The augmented and virtual reality experience has become so popular that will reshape the thought of creating digitally literate students who use technology to create and discover information [10]. This leads to a conclusion that technology will play an important role in students' improved educational success and assessment on their pedagogical thinking.

The process of learning through connections and collaborations facilitated by technology had played an important role in teaching students mathematics in primary school [11]. Different ways of thinking that include the skills for problem-solving; creativity and critical thinking have supported the development and increased children's interests in STEM subjects and careers [12]. Critical thinking and problem-solving approach within digital learning environments have an associated relationship in the learning process. Critical thinking is a meta-cognitive process that evaluates information through exploration of validity and produces logical conclusions to arguments or solutions and achieves resolutions [13]. This skill is becoming very important for education with the large quantity of information and resources made available with the Internet and connected society. Effective critical thinking skills within a digital learning environment will help students become more adaptable, flexible and better able to cope with the rapid development of ever-evolving information [13].

The ever-increasing range of technology tools available to support learning in the classroom enables students and teachers to use digital tools to personalize learning and promote creative thinking within a connected learning classroom [14]. The digital revolution brings many benefits for education and makes students be able to engage in fact-finding, understand bias and validity testing. Furthermore, most of the employers expect that today's graduate students are tech savvy and know how to use technology in their future careers. The digital literacy today means that a person understands computing technologies, programming, and computational concepts, which has become a core skill for an informed participant in modern society [15]. Therefore, it is very important to engage students with the computing concepts in the primary schools.

III. The 21st-century skills in primary schools

Educational innovations in elementary schools are necessary with a strong pedagogical focus on student-centered and increasingly student-directed didactic approaches facilitated by ICT, whereby teachers should play more of a coaching role. Schools must promote "earning to learn", acquisition of knowledge and skills that make possible continuous learning over a lifetime. The 21st-century teaching is characterized by active learning rather than a passive acquisition of knowledge, collaboration in groups through a variety of communication technologies, learning from authentic situations, research rather than memorizing, analyzing the information, critical thinking, problem-solving and using technology. The 21st-century skills encompass a wide range of skills, knowledge, and competencies essential for learning, life, work and participation in an increasingly globalized and connected world. They include skills and competencies such as collaboration, communication, critical thinking, problem-solving, leadership, research and analyses and creativity, see Fig. 1.

A. Importance of the Core Skills

Digital technologies are increasingly pervading all aspects of our lives, including education. As they do, it is very important for teachers and students to not only understand how they might utilize technologies effectively but to be able to critically evaluate the digital content they encounter and

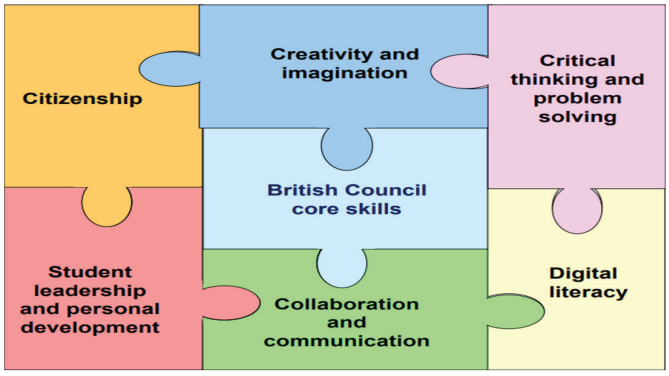


Fig. 1. The 21st-century core skills.

understand the purpose for which it has been developed, to create new information and resources, keep safe and secure on-line etc. Digital literacy encapsulates the range of knowledge, skills and behaviors demonstrated when utilizing a broad range of digital devices, software and applications and in a range of on-line environments, communities, and networks. It's not just about how to use technologies but rather how to understand, find, evaluate content and apply this in a meaningful and responsible manner. Schools and teachers need to consider how they should support the development of digital literacy among the students.

Critical thinking has been a buzzword in education for many years. Students must develop self-directed thinking that produces new and innovative ideas and solves problems, reflecting critically on learning experiences and processes, making effective decisions and finding solutions to some problems. It is more than necessary to taught critical thinking and problem solving, as integral part of the context of subject matter and to train the teachers when to use different critical teaching strategies and how to use them successfully. Students must develop these skills and competencies through solving non-routine problems and questions, considering different perspectives on issues, evaluating evidence for and against different positions and understand the deep structure of issues. Non-routine questions as questions for which there is not a predictable, well-rehearsed approach or pathway explicitly suggested by the task, task instructions, or a worked-out example [20], are the best approach for developing students skills for critical thinking and problem-solving. For that purpose, we need to have trained teachers that will practice different teaching strategies for developing students deeper thinking.

B. Micro:bit

In the digital era, there are plenty of small size digital devices that can support children to study computational concepts, such as Arduino, Galileo, Kano, littleBits and Raspberry Pi [16]. However, each of these digital tools uses a specific coding environment, which is very complex to be learned for the targeted group of students. To gain wider dissemination the project goal is to provide each student with a programming digital device, which puts emphasis on the cost of these devices. The British Council together with the BBC foundation has decided to introduce the Micro:bit device in the given project.

The BBC micro:bit is a pocket-sized, codable computer, designed to allow children to get creative with technology [16]. With the BBC micro: bit computing platform, creating ubiquitous computing applications is very easy to be done by the students [17]. This device has a possibility to connect with other devices, sensors, kits and objects, and is intended as a companion rather than a competitor to other devices (Arduino, Raspberry Pi and etc.) acting as a springboard for more complex learning [16]. The BBC micro: bit is a device with a size of 4cm by 5cm and its design is intended to make it fun and easy to use. It is powered by an ARM Cortex-M0 Processor and has 256K non-volatile flash (for a program and static data) and 16K volatile RAM (for stack, heap) [16]. Considering that BBC Micro: bit has a small device that can be placed in a pocket, its production is low-cost, and because of the compact size, its transport is easy and it can easily become accessible. The chip is self-contained with sufficient on-board sensors and buttons for input, as well as LEDs acting as an output. This makes it a creative tool for primary school students to become more familiar with concepts of algorithmic thinking, science, especially with coding, programming, game development and robotics [18]. The key features that makes this device portable, simple and flexible include:

- 25 red LEDs to light up and flash messages;
- Two programmable buttons to provide input;
- An on-board motion detector or 'accelerometer' to detect forces acting on the device;
- A built-in compass or 'magnetometer' to sense direction;
- Blue-tooth Smart Technology to connect other Micro:bits and devices, kits, phones, tablets, cameras and so on;
- Built-in light sensor on the front side;
- Five Input and Output (I/O) rings to connect the Micro: bit to devices or sensors.

Microsoft has developed an enhanced version of their popular Touch Develop web application and hosted on the Microsoft Azure service [18]. The major benefit for students in the primary schools is that uses web-based MakeCode environment, which does not require installing dedicated software. This environment supports both the JavaScript Blocks editor that makes it easy to program the BBC Micro: bit in graphical coding blocks and JavaScript syntax-directed editor [19]. This editor is very useful because it teaches children basic coding concepts, such as variables, types, procedures, iteration, and conditionals. The visual simulator in the MakeCode environment enhances the student interactive experience by providing a real-time feedback for every change in the workspace, see Fig. 2. The process of the code delivery from the student's laptop to the actual BBC Micro: bit device is done using a USB cable connection. The drag-and-drop feature has simplified as transferring a file from laptop to the Micro: bit device. All of these features of the MakeCode environment provide a solid educational base for learning students about computational

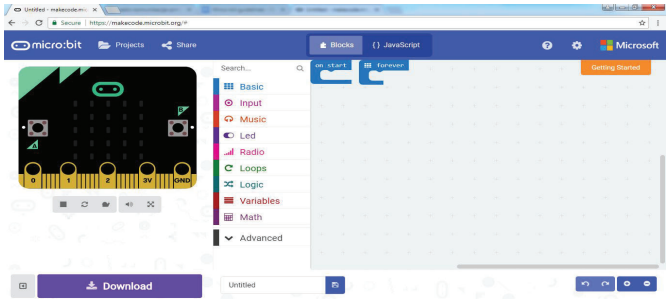


Fig. 2. The web-based MakeCode environment.

concepts [15]. For more advanced users it has a Python editor that is perfect for those who want to push their coding skills further [19].

The BBC Micro: bit device will allow students to think computationally by first formulating a problem and conceptualizing a solution. The computational thinking is a process that involves formulating a problem and expressing its solution(s) in a way that a student is able to communicate with a digital device effectively. Teaching children at an early age the fundamentals of computing will provide them with the programming skills and the computational thinking skills they will need to join the 21st-century workforce. By solving STEM problems with the BBC Micro: bit students will be able to easily understand the computational thinking skills, such as abstraction, decomposition, pattern matching, algorithm design, and data representation. Early learning of the programming concepts enables students to choose more appropriate ICT (Information, Communication, and Technology) resources, as well as to adjust to the way of problem-solving capabilities of these ICT assets [18].

The 21st-century school project elaborated in this work is proposed by the British Council and realized in accordance with the current educational primary school curriculum. The methodology for delivering this project is following the framework for 21st-century learning. It considers two days of training per skill for the teachers, then dissemination and practice for 8 weeks in the schools. During this period, they should adjust the learning material to implement students learning with the specific skill. Finally, the teachers that have delivered the introduction of the new skill will provide a reflection in one day. This project covers primary schools from the Western Balkans, focusing on students aged 10 to 14 years. The first stage of the project realization in Macedonia is intended to support ten primary schools to improve student's capacity and offer higher-order thinking abilities.

IV. Project Implementation and Delivery of the Training in Macedonia

The first phase of the project implementation included training for teachers from ten selected schools and counselors of the Bureau for the Development of Education in the Republic of Macedonia. The realization of the training for teachers within the framework of the 21st-century school project took place at the Hotel Continental, Skopje, on January 10-11-12 and January 15-16, 2018. In order to prepare for the core training skills, we have adopted the British Council facilitator guide and presentations. However, some of the provided examples were not applicable to the targeted audience. Therefore, for the core training skills, we have tried to give national examples and problems that we face in Macedonia.

The training for digital literacy skills has been introduced first. Using a discussion and teamwork we have engaged teachers to share and present good practices of using digital tools in the classroom. Next training was dedicated to critical thinking and problem-solving skills. During this training teachers and counselors have been introduced to the basic concepts of the skills. They have been challenged with different on-site activities, puzzles, and teamwork to practice the skills for critical thinking and problem-solving. The training has the aim to empower teachers with the necessary knowledge and skills for developing digital literacy among the students. It was focused on supporting participants for developing their own interventions and inquiries so they can drive the changes in their schools. The idea was to demonstrate to the teachers the key elements of digital literacy and to show them examples how to implement digital literacy in-school practice.

The whole training was organized by the discussions and sharing ideas, knowledge, skills and good practices among the trainers and teachers - participants at the training. In the beginning, they all agreed that digital literacy is very important and was introduced with the UNESCO (2011) ICT Competency Standards Framework for Teachers. The lack of digital competencies framework in Macedonia was stressed as an important issue for further teachers' professional development of this field. Discussions about what the term digital literacy means and which elements are consisted of lead to the broader definition of this term and better understanding from the participants. The most important conclusion was that digital literacy is not only searching on the Internet or making a presentation, it includes finding, analyzing and evaluation of information, creating their own digital contents, keeping safe on-line etc.

The connection between digital literacy, pedagogy, and learning was emphasized and what pedagogical approaches can be used in order to develop students' digital literacy skills. Different strategies for engaging children with digital literacy were presented: creating educational applications for tools children are already familiar with, enabling children to engage with a broader audience, encouraging children to create digital artifact's, integrate digital literacy into children's research skills. Participants were encouraged to discuss different applications that can be used in the schools. Also, examples how digital literacy look in practices and which activities can lead to successful development of these skills were presented. It was highlighted that digital literacy is something that must be achieved through practice and the participants must take the responsibility for being change agents.

In order to train the teachers to develop students' critical thinking and problem-solving skills, participants were put in

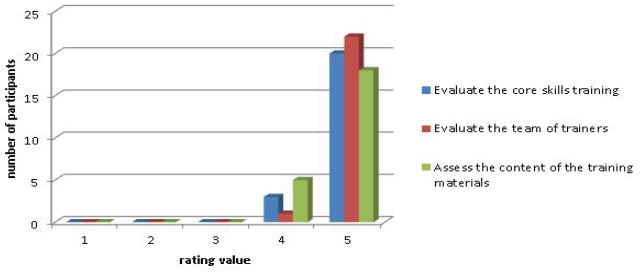


Fig. 3. Evaluation for delivered core skills training.

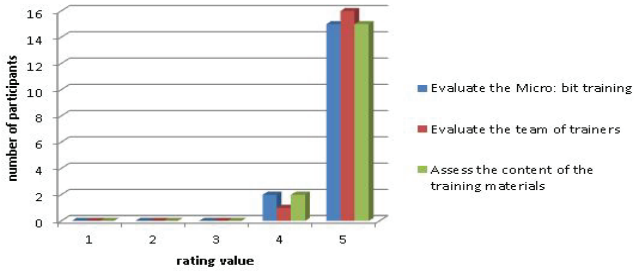


Fig. 4. Evaluation for delivered Mirco: bit training.

the similar situation, which empowers their critical thinking and problem-solving skills. They were introduced to three different teaching strategies: asking questions to develop deeper thinking and check for understanding, modeling how to think critically and solve problems and providing systematic feedback and corrections. Consideration of different perspectives was a good starting point to think analytical about some problems and situations, to see them from different angles and to take into account another opinion before making conclusions. Discussions among participants about the theme lead to a conclusion that teachers didn't put much attention to these activities in the classroom and different strategies for doing that were presented. Open questions and giving feedback are the most used pedagogical approaches in developing deeper learning skills, but because of the different reason, they are not implemented correctly.

In order to effectively develop these skills, students must be asked to do the evaluation of evidence, to clarify why they make a decision or conclude something, to explain the future step of some activities etc. Different strategies for asking questions and ways for delivering effective feedback were demonstrated to the participants. Through discussions and group work participants were put in a situation to discuss their approaches in the classroom and to share examples and good practices. After delivering the training for the core skills, we have made an evaluation for the delivered training, given in Fig. 3. The training for the core skills has been successfully delivered for teachers and counselors.

Last, it was the delivery of the Micro: bit training. The major challenge that we have faced in this training was the lack of time to deliver thorough and practical training. The aim of the Micro: bit device is to inspire future generations of students to have a creative approach to coding, programming and application of digital technologies. It is a very simple computer that accepts input instructions, processes this according to a previously recorded program and then generates an appropriate output. Students can programme the Micro: bit device using another digital device (computer, smartphone or tablet) to write the program, which then switches to the Micro: bit. In general, has been used to draw attention to digital creativity as never before and to help build the digital skills of the nation. The possibility to be connected to external devices and sensors is lowering the barrier to entry into programming. In the same time, students by using the BBC Micro: bit can independently create notions in benefit of computational concepts and become familiar with the basics of programming. Its low-cost price, easy and simple web-based interface and variety of sensors will enable students to continue developing computational thinking skills independently.

During the training, we have organized teachers and counselors to work in teams for more efficient practice with the Micro: bit devices. The training for Micro: bit has engaged all the trainers because it was required to describe Micro: bit features and to support teachers with their practice. The ICT teachers have a crucial role to scaffold their STEM colleagues in the process of practical training to work with Micro: bit. For future Micro: bit training we would recommend more days and more trainers to be dedicated. The training for the Mirco: bit has been positively accepted by attending teachers and counselors, see Fig. 4.

The dissemination on school level was done after the training concerning core skills, especially digital literacy, critical thinking and problem-solving. Participants deliver learned lessons to the other teachers in the school, with the compulsory presence of STEM teachers. Different approaches were presented and ideas, examples of good practices were shared. Most of the teachers implemented the learned lessons in their classrooms in the period of two months after the training. Activities including open questions, effective feedback and use of different digital tools were presented in their reports.

The dissemination of the using Micro:bit was conducted with the teachers on a school level, too. After that, in each of the pilot schools, Coding Club was organized. Students age 11 to 14 years participate in these Clubs, with the aim to empower their coding skills. Teachers' reports show that the students show great interest in using Micro:bit. They are very enthusiastic and motivated to learn to code from one side, and open for finding different examples and idea how can Micro:bit used in other subjects. The interest, motivation, and involvement of the students in the activities is the evidence about the success of implemented activities in the schools, see Fig. 5.

In order to measure the satisfaction of the end users-students from the project, two surveys are going to be carried out (one before starting the work of coding clubs and one after the achieved training). They have the aim to gather data about their

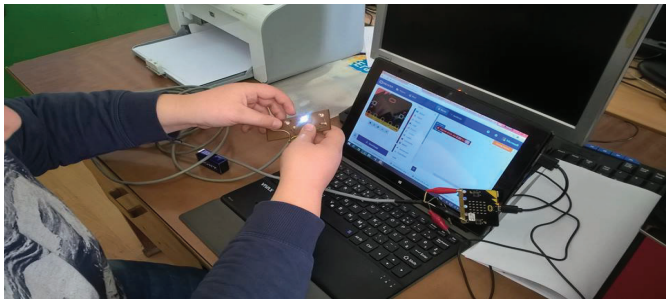


Fig. 5. Student practicing in the Coding Club.

knowledge of digital literacy, critical thinking and problem solving, activities developing these skills in the classroom concerning STEM subjects, their programming skills, before and after the training. The results from these surveys are going to be described in our future work.

V. Conclusion and Future Work

Different reforms were implemented in the Macedonian educational system in the last few years in order to modernize teaching process and to develop students' 21st-century skills. The project 21st-century school is consistent with the improvements in educational practices at the national level, as well as with the development of new educational standards in primary schools. The main goal of this project is to introduce new innovative teaching methods and ways to improve the specific skills, expertise and digital literacy of young people. This project covers primary schools from the Western Balkans, focusing on students aged 10 to 14 years. The first stage of the project realization in Macedonia is to support 10 pilot primary schools to develop capacity and offer their students skills essential to positively contribute to a 21st-century culture and economy.

Results from the used trained methodology, follow up activities and teachers' satisfaction survey show that there is good acceptance of the approach in the primary schools in Macedonia. Teachers agree with the need for the changes in the teaching process and applying different strategies and approaches that will develop students' core skills. In the period of two months after the training, they have implemented gained knowledge and skills in their classroom and made some changes at the school level. In all pilot schools, Coding clubs were initiated and have started with the developing coding skills using Micro:bit. Plans for future activities are made, also. Established network of teachers, results from the survey and conducted follow-up activities are a good start point for developing the 21st-century skills for students.

Acknowledgment

The purpose of the project 21st-century schools was to provide means, directions, and examples how to implement the problem-solving and critical thinking, digital literacy skills. To acknowledge that we had excellent communication and received professional support for organizing the project from the British Council in Macedonia.

References

[1] K. Kumpulainen, A. Mikkola, and A. Jaatinen "The chronotopes of technology-mediated creative learning practices in an elementary school community," Learning, Media and Technology V.39, n.1, pp.53–74, Routledge, 2014. doi:10.1080/17439884.2012.752383

[2] C. L. Scott, "The futures of learning 2: What kind of learning for the 21st century?," ERF Working Papers Series, No. 14, UNESCO Education Research and Foresight, Paris, 2015.

[3] J. van de Pol, M. Volman, and J. Beishuizen, "Scaffolding in teacherstudent interaction: a decade of research," Educational Psychology Review, vol. 22, no. 3, pp. 271–296, Sep. 2010.

[4] T. K. Crowl, S. Kaminsky, and D. M. Podell, "Educational psychology: Windows on teaching," Madison, WI: Brown and Benchmark, 1997.

[5] Cambridge primary curriculum framework. [Online]. Available: http://www.cambridgeinternational.org/programmes-and-qualifications/cambridge-primary/cambridge-primary/curriculum/ ¡accessed: 03.2018¿.

[6] S. Watson, "Higher order thinking skills (HOTS) in education." Thought Co, Apr. 30, 2017. [Online]. Available: https://www.thoughtco.com/higher-order-thinking-skills-hots-education-3111297 ¡accessed: 03.2018¿.

[7] D. Kintsakis and M. Rangoussi, "An early introduction to STEM education: Teaching computer programming principles to 5th graders through an e-learning platform: A game-based approach," in IEEE Global Engineering Education Conference (EDUCON), pp. 17–23, Athens, 2017.

[8] B. Sugrue, "A theory-based framework for assessing domain-specific problem-solving ability," Educational Measurement: Issues and Practices, vol. 14, no. 3, pp. 29-36, 1995.

[9] S. B. Benjamin, "Taxonomy of educational objectives; The classification of educational goals," New York: Longmans, Green, 1956.

[10] R. Ralph et al., "Metrics for evaluation of educational experiences: Will virtual reality have impact?,' 2017 IEEE Pacific Rim Conference on Communications, Computers and Signal Processing (PACRIM), Victoria, BC, pp. 1–6, 2017.

[11] F. Kolyda and V. Bouki, "School essentials for 21st century learning: Connect, collaborate, create, contextualise," International Conference on Information Society (i-Society 2013), Toronto, ON, pp. 207–211, 2013.

[12] T. Bekker, S. Bakker, I. Douma, J. van der Poel, and K. Scheltenaar, "Teaching children digital literacy through design-based learning with digital toolkits in schools," International Journal of Child-Computer Interaction, vol. 5, pp. 29-38, Sep. 2015. DOI: 10.1016/j.ijcci.2015.12.001.

[13] C. Dwyer, M. Hogan, and I. Stewart, "An integrated critical thinking framework for the 21st century," Thinking Skills And Creativity, 12, pp. 43–52, 2014.

[14] K. Robinson and L. Aronica, "Creative schools: revolutionizing education from the ground up," Australia: Penguin UK. 2015. ISBN: 9780141978574.

[15] A. Schmidt, "Increasing computer literacy with the bbc micro:bit," IEEE Pervasive Computing, vol. 15, no. 2, pp. 5–7, Apr 2016.

[16] T. Ball, J. Protzenko, J. Bishop, M. Moskal, J. de Halleux, and M. Braun, "The bbc micro: bit coded by microsoft touch develop," Microsoft Research, 2016.

[17] Y. Rogers, V. Shum, N. Marquardt, Z. Lechelt, R. Johnson, H. Baker, and M. Davies, "From the bbc micro to micro: bit and beyond: a british innovation," Interactions, vol. 24, pp. 74–77, 2017.

[18] D. Stanojevic, A. Rosic, B. Randelovic, and Z. Stankovic,"Use of BBC Micro:Bit in teaching technical and IT education, " Sinteza 2017 - International Scientific Conference on Information Technology and Data Related Research, Belgrade, Singidunum University, pp. 231-235, 2017.

[19] Micro:bit Educational Foundation. [Online]. Available: https://microbit.org/ ¡accessed: 03.2018¿.

[20] J. Woodward, S. Beckmann, M. Driscoll, M. Franke, P. Herzig, A. Jitendra, K.R. Koedinger, and P. Ogbuehi, "Improving mathematical problem solving in grades 4 through 8: A practice guide (NCEE 2012-4055)," Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education. 2012.

The Need for an Automated Model which makes Matching between Job Market Demand and University Curricula

Ylber Januzaj, Artan Luma, Azir Aliu, Besnik Selimi, Bujar Raufi, Halil Snopce
Faculty of Contemporary Sciences and Technology
South East European University
{yj16535, a.luma, azir.aliu, b.selimi, b.raufi, h.snopce}@seeu.edu.mk

***Abstract*— The need for a balance between university curricula's and labor market demands has always been high. With the increase of graduating students, there is also a growing need for an automated model which makes adaptation between university curricula and labor market demands. The current need for such a model is based on the large number of study programs that are currently offered by the universities in Kosovo as well as the huge labor market demands in the region. In our paper we present an analysis of the current situation of universities curricula's and market demands in the field of technology in Kosovo. In this paper we also identify the ways how universities are announced with job vacancies in order to keep informed students and make easier to find a related job. Finally we extract information's from specific websites which publishes job vacancies in the field of technology. We make this extraction by applying open source framework which extracts information's and save them in a specific format.**

Keywords—university curricula, market demands, data-mining, vacancies, employment, higher education.

I. INTRODUCTION

Nowadays the importance that high education holds all over the world is high, having a direct impact on the state's income [1, 2, 3]. On the other hand, we also know the problems that the labor market is facing every day. On the one hand, we have a great variety of study programs that are offered in the field of technology, but on the other hand we also have large market demands that are not being met by these graduate students. These requirements, although they are at a high level, they also are increasing day by day. Realistically, meeting labor market demands is currently a very active topic, and a huge research and investment is underway in such projects [4, 9, 10]. So far research has gone into this area that it has been predicted if market demands will be filled by graduate students in the future. Our research gives special importance to the field of technology, both in terms of labor market requirements but also in terms of curricula currently offered by universities in Kosovo. Knowing that Kosovo is not really good in terms of unemployment in the field of technology, it is reasonable to look at such an analysis and research, which will push us to the creation of an automated model which makes a comparison between market demands and university curricula offered in Kosovo. In our paper we are able to see the forms that students and universities are announced about latest job vacancies in the field of technology. Also in our paper we can see how much importance universities give to such an automated model that will be able to make comparisons between labor market demands and curricula offered by these universities. And finally, in our paper we can see how much universities in Kosovo are able to support such a model. Next we show the higher education situation in Kosovo, by showing total number of Higher Education Institutions in Kosovo, including private and public institutions. Also we show the current situation of accredited program in the field of technology.

II. HIGHER EDUCATION IN KOSOVO

For a qualitative education, it is necessary to have an agency that controls the quality of a country's education [7]. Kosovo, the place where our research has been made, possesses such an agency, which accredits study programs almost every year. The number of Higher Education Institutions that are accredited in Kosovo is considerable, based on the number of that currently operate as public universities and private colleges. Next, we show the accredited Higher Education Institutions in Kosovo.

TABLE I. Higher Education Institutions in Kosovo

Higher Education Institutions in Kosovo	
Sector	**Number**
Private	21
Public	9
Total	30

As we can see the number of Higher Education Institutions which are accredited in Kosovo is 30, being divided into 9 public universities and 21 colleges. The areas that students can attend are diverse, ranging from social to scientific and technical. In order to have a deep research, we present the number of programs accredited by these institutions.

TABLE II. Higher Education Institutions Accredited Programs in Kosovo

Study programs				
Sector	**BA**	**MA**	**PhD**	**Total**
Private	124	73	0	**206**
Public	130	108	24	**262**
Total	254	181	24	**468**

In Table II we can see the list of study programs accredited by KAA[1] in Kosovo. As we can see, in total we have 468 programs accredited by public universities and private colleges that currently operate in Kosovo. Given the high unemployment rate in Kosovo and the large number of study programs offered by these higher education institutions, we can easily conclude that there is a discrepancy between the study programs offered by these institutions and market demands. As we said in the beginning, our research will be focused in the field of technology, so next we present the number of Technology Faculties in Kosovo.

TABLE III. Number of Technology Faculties in Kosovo

Technology Faculties				
Sector	**BA**	**MA**	**PhD**	**Total**
Private	7	3	0	**10**
Public	5	4	1	**10**
Total	12	7	1	**20**

As we can see, we have in total 12 faculties that offer Bachelor level studies in the field of technology, of which 7 are offered by private colleges and 5 of them from public universities. While at the Master level there are 7 faculties that offer technology study programs, of which 3 of them are offered by private colleges and 4 of them from public universities. While at the PhD level it is only one faculty where technology programs are offered. Below we present the number of study programs offered by these faculties in Higher Education in Kosovo.

TABLE IV. Number of Study Programs in Field of Technology in Kosovo

Study programs in field of Technology				
Sector	**BA**	**MA**	**PhD**	**Total**
Private	16	8	0	**24**
Public	22	18	4	**44**
Total	38	26	4	**68**

Table IV shows the number of technology study programs offered by Higher Education Institutions in Kosovo. In total, we have 68 study programs in the field of technology, 38 of which are Bachelor level, 26 are of the Master level and 4 of them are Doctoral level. If we make a comparison between public universities and private colleges, we can see that there is a drastic change in the technology programs offered in Kosovo. Public universities are the ones that are in advantage in number of study programs in the field of technology. One of the main reasons why we have such a big difference between these programs offered is exactly the fact that the academic staff is missing in order for these programs to be accredited [6, 8].

III. RESULTS

As we mentioned in previous, the processing of results that have been derived from public universities and private colleges in Kosovo is very necessary. The problems that we wanted to solve through these results were different, all of them important to implement such a model that would make comparisons between market demands and curricula offered by universities in the field of technology. Some of the problems that we wanted to solve are:

- Forms of how universities in Kosovo are announced on recent vacancies that are open as a job opportunity for their students.
- How much universities in Kosovo think it is necessary to create a model that will make a comparison between market demands and university curricula.
- How much universities are able to support such a model.
- How much model will meet market requirements.
- How much the model will be able to mitigate the unemployment situation in Kosovo.

In the following we show the form of how public universities and private colleges in Kosovo are announced about latest job vacancies.

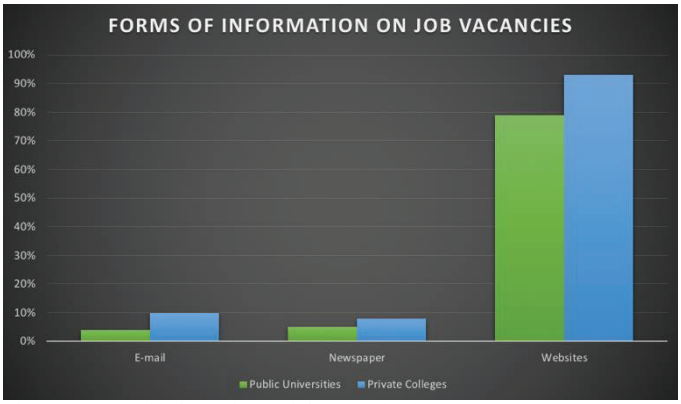


Fig. 1. Forms of information on job vacancies in Kosovo

In Fig. 1 we can see the forms that universities are announced about latest job vacancies. As it is shown graphically we have 3 forms: through e-mail, through newsletters and through web pages. So these are 3 general forms of how universities in Kosovo are notified of open job vacancies. In Fig. 1 we can see that only 5-10% of the notifications are through e-mail and through newsletters. While the most noticeable form of announcements is the form through the websites, where over 90% of the announcements that universities take over job vacancies take through this form. Next, we show how much universities think that such a model is necessary to be implemented.

[1] http://www.akreditimi-ks.org/new/index.php/en/

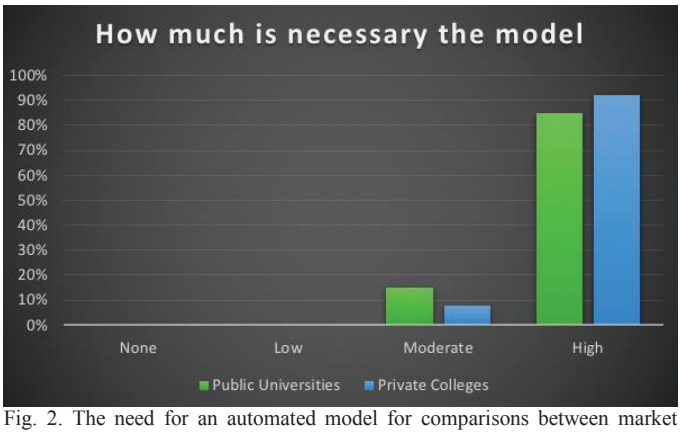


Fig. 2. The need for an automated model for comparisons between market demands and university curricula

In Fig. 2 we graphically present the need for the implementation of an automated model that will make a comparison between market demands and curricula offered by universities. From the answers received from Higher Education Institutions in Kosovo we see that over 90% of them think that the need to implement such a model is very high, while about 10% think that the need to implement such a model is moderate. So with such a percentage that universities have expressed then we can easily conclude that such a model is very necessary to be implemented. Next we show how much such a model is supported by universities in Kosovo.

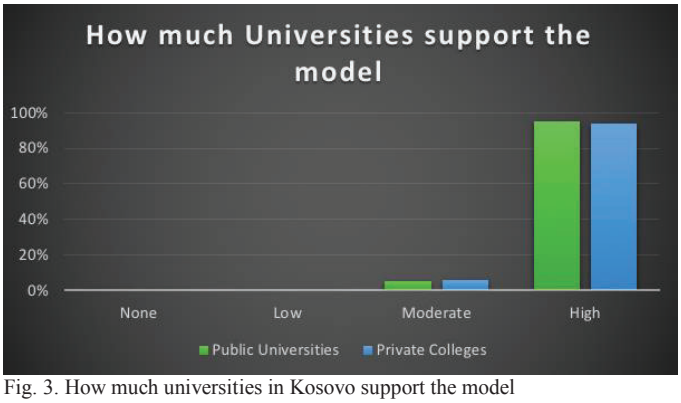


Fig. 3. How much universities in Kosovo support the model

In Fig. 3 we can see graphically how much such a model will be supported by universities in Kosovo. According to the results obtained and processed, we can see that over 97% of the Higher Education Institutions in Kosovo have expressed high readiness to support such a model, and 3% have expressed a moderate readiness to support such a model. When talking about support, here is the readiness of Higher Education Institutions to provide open access to their data as well as at the same time applying such a model from each institution in order to have a high level of adjustment of their curricula. In the following, we show how much universities in Kosovo think that such a model will be able to meet the current labor market demand in the field of technology.

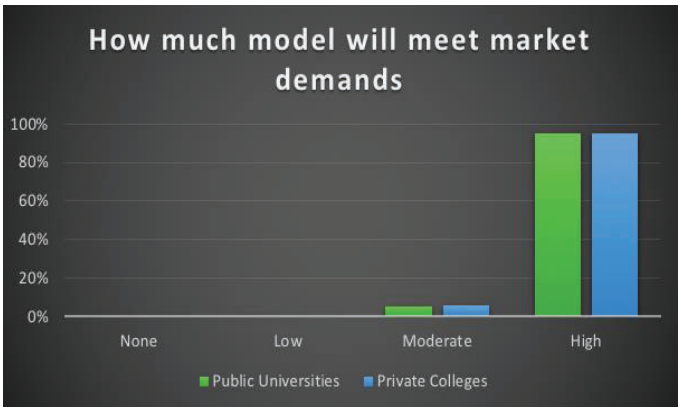


Fig. 4. How much model will meet market demands in the field of technology

In Fig. 4 we can see that approximately 95% of universities in Kosovo think that such a model that will make comparisons between labor market demands and university curricula will meet the market demand at a very high level high. While about 5% of them think such a pattern will be able to meet the labor market demand at the moderate level. When we talk about meeting the needs of the labor market in the field of technology, we mean that the market in Kosovo is currently facing more difficulties in finding prepared staff. In the following, we show how much such a model will mitigate the level of unemployment in Kosovo.

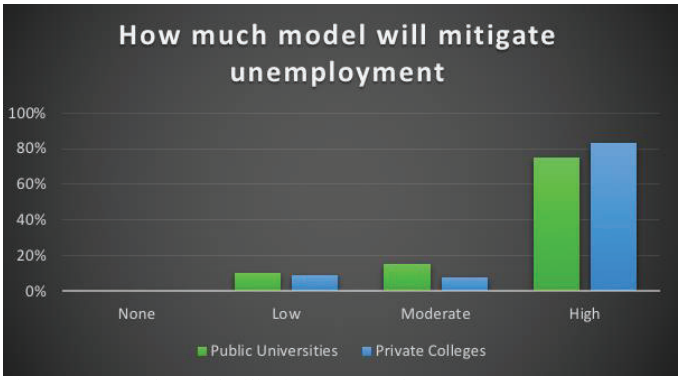


Fig. 5. How much model will mitigate unemployment in Kosovo

Calculating the high level of unemployment in Kosovo [5], we have also thought that such a model will be able to mitigate the situation of unemployment. In Fig. 5 we can see that approximately 80% of universities think that such a model will mitigate the level of unemployment at a high level, 15% of them think that such a pattern will be able to mitigate the level of unemployment at moderate level, and 5% of them think that unemployment will be mitigated at a low level. Therefore, also calculating the results of universities for such a model then in this case we conclude that such a model is very needed and important to be implemented.

IV. JOB VACANCIES

The form of how job vacancies are published in the field of technology is very important in our research. As we mentioned in previous up to 90% of universities get information about job vacancies through web sites. Based on this, this is the form that will be used by us in order to implement our model. In order to do this, we need to apply some techniques to extract website information about job vacancies in the field of technology. Based on our experience the most appropriate framework for extracting data from a specific website is Scrapy[2]. The next step is the identification of the websites that publish information on job vacancies in Kosovo. The most functional and usable website for job vacancies in Kosovo is Portalpune[3], which publishes the latest vacancies in different fields. Next we show the code which is executed in order to extract the website content, this code is used from Scrapy Tutorials[4] from official website and modified and adapted to our needs.

```python
import scrapy


class QuotesSpider(scrapy.Spider):
    name = "quotes"

    def start_requests(self):
        urls = [
            'hhttp://www.portalpune.com/other.php?Filter=&Category=54&Place=-1'
        ]
        for url in urls:
            yield scrapy.Request(url=url, callback=self.parse)

    def parse(self, response):
        page = response.url.split("/")[-2]
        filename = 'quotes-%s.html' % page
        with open(filename, 'wb') as f:
            f.write(response.body)
        self.log('Saved file %s' % filename)
```

Fig. 6. Source code of Scrapy framework

In Fig. 6 we can see the source code with is used in order to extract information from specific website. As we can see the source code is divided in three parts, each of them with different roles. The first part is used to nominate the project and to identify the spider, in our case is used the name “quotes”. Second part is used to define the url that we want to visit and extract, in our case we used the link of the website that publishes job vacancies in the field of technology. Finally we have the third part which is a parse method. This method is used to hold the information’s which are extracted from website, and these information’s are later saved in html file. Depending on the url we possess, such files will be saved. Next we show extracted information which is saved in html file.

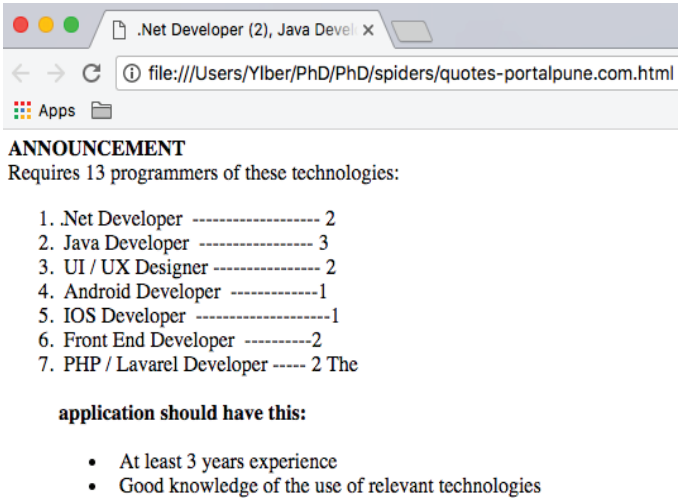


Fig. 7. Extracted information from a specific link

In Fig. 7 we can see the information’s which are extracted from the website. As we can see we have at least 13 positions in the field of technology including .Net and Java Developer, Android and IOS Developer, UX Designer, Front End and PHP Developer. Based on this we have all the information’s to make comparisons between these positions and university curricula.

V. CONCLUSION

In this paper we analyze the need for a model which makes comparisons between market demands and university curricula. We started with reviewing the current situation of Higher Education Institutions, including the number of public universities and private colleges in Kosovo. First we express the total number of accredited programs which actually are functional, to continue later with the total number of faculties which actually are accredited in Kosovo. As we can see we have in total 20 technology faculties which are functional in Kosovo. Second we express the number of accredited programs in the field of technology in public universities and private colleges. Based on the information which are published in their websites we have in total 68 accredited programs in the field of technology in Kosovo. As a part of our paper was also the analysis of the need for a model which makes comparisons between job market demands and university curricula. Based on the results which are taken from public universities and private colleges in Kosovo we can see that the form that they are announced about job vacancies is by specific websites. We see that up to 90% of universities and colleges thinks that the need for such a model is very high, and also the support that they offer in this direction is high. Also we can see that up to 90% of universities and colleges thinks that such a model will meet market demands in high level, and up to 80% think that such a model will mitigate the unemployment situation in Kosovo. Finally we used a framework to extract information’s from a

[2] www.scrapy.org

[3] www.portalpune.com

[4] https://docs.scrapy.org/en/latest/intro/tutorial.html

specific website which publishes latest job vacancies in the field of technology. Based on this we conclude that the need for a model which makes comparisons between market demands and university curricula is very high, and we also conclude that such a model will solve many problems of Higher Education Institutions, making them to adapt their curricula to the market demands.

REFERENCES

[1] L. Anastasiu, A. Anastasiu, M. Dumitran, C. Crizboi, A. Holmaghi, M. Roman, "How to align the university curricula with the market demands by developing employability skills in the civil engineering sector". In Education Sciences, pp. 2-7, September 2017.

[2] J. Andrews and H. Higson, "Graduate employability, "soft skills" versus "hard" business knowledge: a European study". In Higher Education in Europe; No. 4; Routledge Francis and Taylor Group: Oxford, UK, pp. 1-4, 2008.

[3] N. Tiraieyari and J. Hamid, "Employability of leadership development program graduates: career identity, social capital, psychological capital, employability, graduates". In Saarbrücken, Germany, pp. 46-63, 2012.

[4] M. Agaoglu, "Predicting instructor performance using data mining techniques in Higher Education". In IEEE Access, pp. 4-6, May 2016.

[5] Unemployment Statistics. Available online http://ask.rks-gov.net/

[6] P. Stokes, "Higher Education and employability: new models for integrating study and work". In Harvard Education Press: Cambridge, MA, USA, pp. 78-94, 2015.

[7] C. Mocanu, A. Zamfir, S. Pirciog, "Matching curricula with labour market needs for Higher Education: State of Art, obstacles and facilitating factors". In Elsevier, 1877-0428, pp. 1-2, 2014.

[8] F. Eyraud, D. Marsden, J. Sylvestre, J.-J, "Occupational and internal labour markets in Britain and France". In International Labour Review, 129, 501-517, pp. 8-12, 1990.

[9] W. Bartlett, M. Uvalic, N. Durazzi, V. Monastiriotis, T. Sene, "From University to employment: Higher Education provision and labour market needs in the Western Balkans synthesis report". ISBN 978-92-79-64428-3 doi:10.2766/48413, pp. 51-73, 2016.

[10] E. Corominas, C. Saurina, E. Villar, "The match between University Education and graduate labour market outcomes (education-job match)". Catalunya, pp. 21-38, 2010.

Analysis of Using Information and Communication Technologies in Mathematics Courses in the Republic of Macedonia

Mirjana Trompeska
Institute for Advanced Composites and Robotics – Prilep
Faculty of Information and Communication Technologies-Bitola, "St. Kliment Ohridski" University – Bitola
Republic of Macedonia
mirjanat@iacr.edu.mk

Blagoj Ristevski
Faculty of Information and Communication Technologies-Bitola
"St. Kliment Ohridski" University – Bitola
Republic of Macedonia
blagoj.ristevski@uklo.edu.mk

***Abstract*– The current advances in global economy require worldwide unceasing changes in the education. In the past fifteen years, the education in the Republic of Macedonia was also subject to reforms in order to initiate development of the skills for digital literacy, inventive thinking, effective communication and high productivity. This paper presents the results obtained from research about using information and communication technologies (ICTs) in the classroom of the teachers' point of view. This research aims to explore how the mathematics teachers in the Prilep region respond to these fast changes, what are the advantages and disadvantages of the education reforms, as well as to detect the major problems they faced during the reform implementations and what can be done in order to overcome these obstacles. For that purpose, we proposed a new educational portal.**

Keywords*– *e-learning, information and communication technologies, learning management systems, ICT in mathematics education.

I. INTRODUCTION

Nowadays, tremendous technological changes and globalization lead to development of a new global economy followed by the unceasing lifelong learning. Thereby, the education cannot be indifferent for those changes which lead to a transition from industrialization-based society to information-based society. Every participant in the global society beside the basic reading and writing skills has to develop skills for lifelong learning: skills for digital literacy, inventive thinking, effective communication and high productivity skills [1].

Therefore, under the influence of the globalization and the rapid technological development, the global educational process in the past years is subject to continuous changes. In general, these educational reforms have a tendency to employ the technology in the teaching and learning processes. Teachers had started to prepare their teaching materials in electronic form, while pupils and students had started to do their homework and projects in the same form. These changes led to the student's achievements and examination to be monitored by using ICT, printed books are replacing with electronic books, traditional boards are replacing with smart boards etc.

ICTs in a very short period of time became important part of the contemporary society, because in many countries their usage is considered as basic knowledge and skills, like reading and writing. For properly use of the term ICT, it has to be noted that the term isn't only used for computers and computer activities, but it has more general meaning: Internet services, information and communication equipment and services, computer networks, mass media and broadcasting, libraries and documentation centers [2] [3].

Numerous studies about using ICT in the education are conducted and almost all of them have shown the gain of their application. The use of ICT in the education has potential to speed up the learning and teaching processes, to give an opportunity for wider and deeper knowledge, to motivate pupils and students for active learning process, to help them for easier memorizing and understanding lectures and to encourage their ambitions for self-study and self-research [2][1].

Based on the research carried out in 2002 in the Republic of Macedonia, regarding the state of the education, it was concluded that the education in Macedonia wasn't following the development of the new global economy engendered by the rapid technology development and globalization. Therefore, Macedonian Ministry of Education and Science authorities had decided that the education needs to be reformed in order to promote the development of the skills needed in the 21st century. Starting from 2005, in the next ten years the education in Macedonia, according to several programs that were approved, was subjected to respective changes that are detailed in the next section.

The goal of this paper is to explore how teachers react to those changes after implementation of the educational reforms. The targets of this research were teachers that teach Mathematics in the primary schools. The research intends to identify the advantages of the use of ICT in the Mathematics education, to identify the problems that Mathematics teachers

encountered during implementation of ICT in education, and to offer appropriate solutions for some of the noticed problems.

The rest of the paper is organized as follows. In second Section, the conditions of the education in Macedonia is detailed, during the implementation of the educational reforms since 2005, while Section 3 details the methodology of the research. Section 4 gives the results gained from the research followed with brief discussion. Next section provides a proposal to overcome some of the identified problems that faced during the implementation of ICT in the classes, while the final section provides concluding remarks.

II. THE STATE OF THE EDUCATION IN MACEDONIA

In order to fulfil the reform in the primary and secondary education, in 2005 several programs, such as "*National Program for Educational Development (2005-2015)*", "*Draft Program for Development of ICT in Education (2005-2015)*", "*National Strategy for Development of Informational society (2005 - 2015)*", "*National Strategy for e-content (2010-2015)*" and "*National short-range ICT Strategy (2016-2017)*" were introduced [4][3] [5] [6] [7]. The reform for digitalization and computerization of the education started in 2005, but its impact started to be perceived in the schools two years later with the realization of the project "*Computer for every child"*. With this project in every classroom of the 366 primary schools and 93 secondary schools 17.818 personal computers, 98.710 LCD monitors, keyboards and mice, 53.00 portable computers for the students in the first, second and third grade, and 22.000 portable computers for the teachers were installed. With the installed computer equipment two thirds of the students in Macedonia had a possibility to access to computer and educational applications during their time spend at school. The installed computer equipment works under the Edubuntu platform, which includes additional 46 educational applications intended for particular application of ICT in the following subjects: Computer Science, Mathematics, Physics, Chemistry, Music and Latin language.

For successful implementation of the educational reforms, several trainings and workshops about using installed hardware and software, aimed to teachers, were organized by the Ministry of Education and Science and the Ministry of Information Society and Administration. The trainings and workshops started in 2009 and were intended to introduce the teachers with the installed computer equipment, operating system Edubuntu, software packages OpenOffice and another 46 additional educational software applications.

In 2010, an educational portal containing 513 learning objects was created, whose learning objects were composed of simulations, multimedia contents and notes. This portal was intended to assist in the learning process for the following subjects: Mathematics, Biology, Chemistry and Physics.

As a final phase of the reform for development of ICT in education, a web-portal as a Learning Management System was created. This web-portal was intended to make available the online learning, connections among all students/pupils and all teachers involved in the educational process, as well as to control and monitor the student's achievements, sharing of learning materials, etc. But, although the implementation of this reform in the education is finished, this Learning Management System was not yet fulfilled completely.

The Bureau for education development has changed the educational programs by introducing new subjects in order to support ICT literacy and ICT skills, such as: "Working with Computers" (facultative subject in first, second and third grade), "Computer Science" (mandatory subject in sixth and seventh grade) and "Projects from Computer Science" (facultative subject in seventh, eighth or ninth grade). Also the Bureau for education development has made annexes of the existing syllabuses regarding the methodical directions (recommendation and praxes) for all subjects, in order to make the use of ICT obligatory process in the formal education in Macedonia. According to those directions at least 30% of the classes that are held for one subject yearly, have to be implemented using ICT tools. It means that for the subject Mathematics in primary school, which is represented with four hours weekly or annual fund of 144 hours, teachers have to implement ICT at least in 44 hours or once in every school week. The teachers have completely freedom in the selection for the ICT tools that have to use in their classes.

In the meanwhile, the subject Mathematics in 2013 was subjected to another reform called "Cambridge Educational System". Then the Ministry of Education and Science signed an agreement with the Center for International Education of Cambridge to implement its programs and to use its books. The Center for International Education of Cambridge is the largest provider of international educational programs and qualifications for students from 5 to 19 years that are accepted in 9000 schools in 160 countries. This reform in the primary school was achieved in three time stages or in three successive school years. The "Cambridge Educational System" was implemented in three successive grades. The new reform had an intention to develop the critical thinking among students.

According to the described state all teachers in Macedonia in a very short period of time had faced many changes in education. In order to investigate how they handle with the numerous changes, particularly with the implementation of the ICT in their classes, in 2015, a study that is described in the next section was carried out.

III. METHODOLOGY

The observed target group was the Mathematics teachers from the primary schools in the Prilep region. Thirty Mathematics teachers were questioned, which is 80% from all Mathematics teachers that realized classes in the Prilep region.

As an instrument for data collecting a questionnaire, which contained 36 questions in open and closed form, was used. The questions were organized in four categories: (1) basic data for the target group, (2) implementation of ICT in education, (3) usage of ICT tools and (4) problems in implementation of ICT in education.

The first category had three questions needed to make teachers' profiles. The second category contained eleven questions in order to gain the data for implementation of the ICT in education, like the level of implementation of the ICT in their classes, advantages of implementation of ICT, the time

needed to prepare the lessons, the usage of existing ICT tools in their classes etc.

The third category was consisted of eleven questions in order to gain data about ICT tools used in the Mathematics classes. The fourth category of the questionnaire was also composed of eleven questions, asked in order to identify the problems, which teachers encountered during the implementation of ICT in their Mathematics classes.

IV. Results and discussion

According to the data collected from the first category of the questionnaire, 10% of the target group belong to the first age group (from 20 to 30 years), 43% of the target group belong to the second age group (from 30 to 40 years), 27% of the target group belong to the third age group (from 40 to 50 years), while 20% of the teachers were aged between 50 and 60 years.

20% of the target group were male teachers, while 80% of the target group are female teachers. This gender imbalance isn't surprising because in Macedonia most teachers are female. 37% of the target group realized classes in the rural areas, whereas 63% in the urban area of Prilep.

The second category of the questionnaire aimed to collect general data about the implementation of ICT in the Mathematics classes. The results for some similar questions are presented in one bullet, as follows:

- 17% from the Mathematics teachers implement ICT almost every day, 53% from the Mathematics teachers implement ICT once or twice per week, while 30% of Mathematics teachers implement ICT once or twice monthly. Hence, generally, 70% from the participants of the target group follow the instructions from the Bureau for education development, and at least once per week implement ICT tool in their classes. Furthermore, analysis related to the teachers' age groups has shown that most participants who do not follow the instructions from the Bureau for education development belongs to the first age group (from 20 to 30 years) and to the fourth aging group (between 50 and 60 years). As a main reason for this results it can be considered the lack of working experience among younger teachers, as well as the refusal of work with computer among older teachers.
- From the implementation of ICT in their classes, they noticed several advantages like better student's motivation, interest and activity during classes, almost all students do their tasks quickly, improve exploratory and creativity skills, gain easy access to data needed for the classes etc.
- The time needed for preparing the activities and materials for Mathematics classes with implementation of ICT is three times bigger than the time for preparing the activities and materials for Mathematics classes in the standard manner. The average time for preparation activities and materials for Mathematics classes by implementing ICT in the class with just one ICT tool is 110 minutes, while in the case of combining of more ICT tools in one class is 142 minutes.
- 70% of the participants declared that the class duration (40 minutes) isn't sufficient to realize the whole prepared activities when the teacher implements ICT in the Mathematics class. The main reasons about this are the technical troubleshoots occurred on the installed computer equipment during the class.
- 40% of the participants answered that they deal individually with the technical problems of the installed computer equipment, 13% of them get help from the Computer Science colleagues, 23% of them get help from the computer equipment maintainers, 10% of the them get help from the students, 7% of them get help from the students' teams for computer equipment maintain, while 7% of them answer that nobody deals with the technical problems of computer equipment. From their responses, it can be seen that teachers have to deal with the technical problems of the computer equipment individually or to get help from students and the Computer Science colleagues, because the computer equipment maintainers at the same time work in 4 or 5 different schools and usually they should be present in one school in one particular day in the week. In the last two years this time of their presence was reduced to once or twice per month.
- 63% of the teachers implement ICT only within the Mathematics classes. 37% of the teachers implement ICT during the Mathematics classes and as well as for preparing homework and projects. The big percentage difference is because there are few electronic, public and free educational contents that entirely match to the Mathematics curriculum and student's tasks.
- 13% of the teachers rarely use ICT tools, 47% of them sometimes use ICT tools, 27% use ICT tools often, while 13% of the teachers always use ICT tools that are prepared by somebody else. This fact is not surprising if the time needed to prepare a lesson with implementation of ICT tools has been taken into account.
- According to the responses given by the participants, more than 80% of the electronic contents prepared by somebody else can be entirely or with few changes used to their classes. The problem is the way they obtain those ICT tools, because most of the teachers get them from other colleagues. Also, more than 70% of the teachers answered that electronic contents, which can be used as ICT tools, are free and publicly available on Internet, are not in Macedonian language.
- The final question from the second group is an open question where the Mathematics teacher has to write some examples of lessons where ICT is implemented and from that responses can be seen that the teachers more often implement ICT in Geometry classes, than in the Algebra classes.

The results from the third category of the questionnaire aimed to collect the data from the Mathematics teachers about the ICT usage in the classes and in their everyday lives. In the questionnaire, there were listed 10 types of different programs and the participants had to answer how often they use the particular type of program in the classroom and in their everyday usage. The results are shown in Fig. 1.

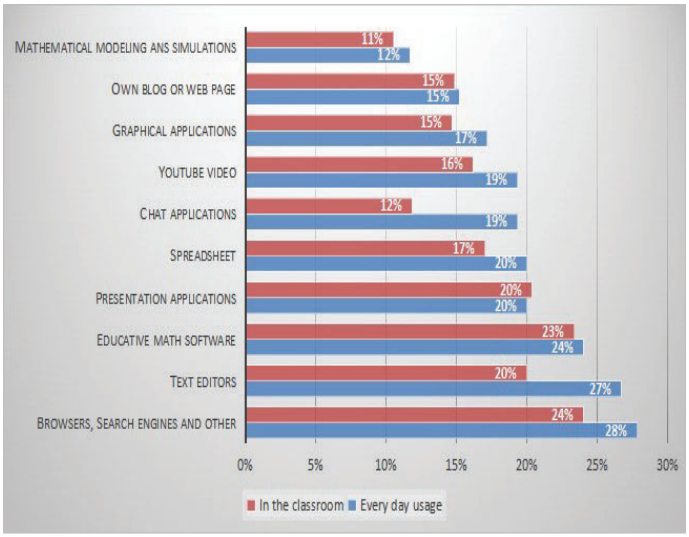


Fig. 1. Usage of different applications in and outside the classroom during the implementation of ICT.

From the results presented in Fig. 1, it can be noticed that the teachers in their everyday usage mostly use the browsers, search engines and other Internet activities. Thus, the everyday usage of the computer is unthinkable without Internet connection, because teachers almost every day use the ICT tools that allow them opportunity to search data, use the social networks etc. The text editors are the second mostly used applications by the teachers in everyday usage, because the teachers in Macedonia in the past ten years have enormous additional administrative work. In the third place are the educative Mathematics applications used by the Mathematics teachers outside the classroom, in order to prepare themselves for the classes.

Browsers, search engines and other Internet activities are also common used ICT tool in the classroom. The second mostly used ICT tool in the classroom are the educative Mathematics applications, while the third mostly used ICT tool are the applications for presentation.

The fourth and last category of the questionnaire aimed to collect general data in order to identify the problems occurred during implementation of ICT in the Mathematics classes. Fig. 2 presents the results from the last category of the questionnaire followed by a brief discussion.

One of the main problems, identified by the teachers, is the hardware fault handling. The computer equipment was installed in 2007, and now after more than one-decade intensive usage, the computer equipment is malfunctioning. Most schools do not have a technical person for computer equipment maintain as an employee with full time job, and that why the teachers are left individually to handle above described problems.

As the second important problem, the teachers identified the shortage of time for creative work and efficient application of ICT in the classes, because of overloading with enormous administrative work in the past ten years.

The third important problem identified by the teachers is that the students do not have access to the installed technology from their homes. This problem covers the problems with the economic status of the students and with the lack of a common educational management system that can provide online learning, at least. Although the "National program for development of education" has predicted design and implementation of Learning Management System, which has to provide online learning, connection between the students and teachers, connection between all teachers in the country, following the students' achievements and sharing of learning materials, this system was not yet created. Thus, although the "National program for development of informational society" aimed to provide access to computer equipment by Internet connection for everyone, there are still many students (mostly students from low-income families) that don not have a suitable access to ICTs.

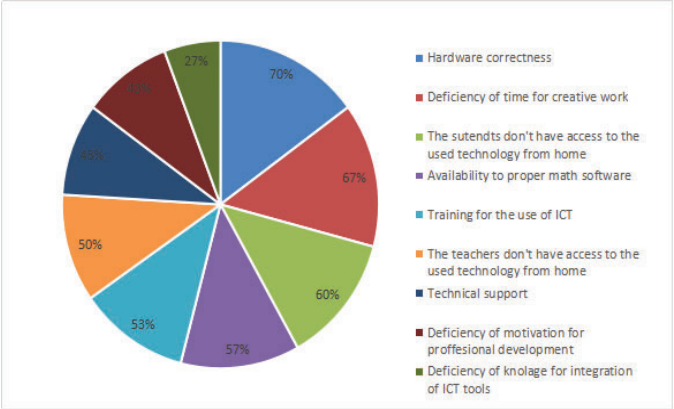


Fig. 2. Arisen problems during the implementation of ICT in education.

V. EDUCATIONAL PORTAL

As a proposal to overcome the most problems occurred during the implementation of ICT in classes, an educational portal was created. This educational portal aims to provide the necessary electronic teaching and learning materials, to propose an innovative manner for realization of teaching and learning process with different ICT tools, also to provide a place for exchanging teaching and learning materials and to enable a collaboration between students and teachers. In other words, the aim of this educational portal is to grow up and to serve as Learning Management System for primary schools in Macedonia.

The first version of the educational portal presents a frame of electronic contents for teaching and learning Mathematics, Biology, Physics, Chemistry and other subjects in primary school as shown in Fig. 3. All the electronic contents presented in this educational portal implement different ICT

tools in order to initiate development of the skills needed in the 21st century.

Fig. 3. The home page of the educational portal.

VI. Conclusion

From the presented results, it can be concluded that Mathematics teachers from the Prilep region municipalities, generally follow the suggestions from the Bureau for education development and at least once per week implement ICT in their classes. In those classes they mostly implement the applications that allow them to browse or search information through the Internet, educative Mathematics software and applications for presentation. During implementation of ICT in the Mathematics classes, teachers had identified the hardware problems, the shortage of time for creative work and the problem related with the lack of access to the used technology from the students' home as three major problems. Hence it can be concluded that malfunctioning computer equipment, technical computer equipment maintainers with full time job in the schools, periodical training for usage of new ICT tool, pointing out more electronic contents about the use of ICT tools, as well as design of Learning Management System can improve the implementation of ICT in the classrooms. Some of the mentioned problems can be solved by using the educational portal described in the previous section.

References

[1] Burkhardt, Gina, M. Monsour, G. Valdez, C. Gunn, M. Dawson, C. Lemke, E. Coughlin, V. Thadani, and C. Martin. "enGauge 21st century skills: Literacy in the digital age." Retrieved June 2 (2003): 2008.

[2] Noor-Ul-Amin, Syed. "An effective use of ICT for education and learning by drawing on worldwide knowledge, research, and experience: ICT as a change agent for education." Scholarly Journal of Education 2, no. 4 (2013): 38-45.

[3] Fu, Jo Shan. "ICT in education: A critical literature review and its implications." International Journal of Education and Development using Information and Communication Technology 9, no. 1 (2013): 112.

[4] Zivanovic R., "Use of computers and Internet in the education system of the Republic of Macedonia", Metamorphosis, Foundation for Sustainable IT Solutions, 2010.

[5] Ministry of Information Society and Administration, Government of the Republic of Macedonia, "National Strategy: e-Content Development Strategy 2010-2015", 2010.

[6] Ministry of Information Society and Administration, Government of the Republic of Macedonia, "National short-term ICT strategy", 2015.

[7] Ministry of Information Society and Administration, Government of the Republic of Macedonia, "The National Strategy for Development of Information Society", 2015.

An Application for Psychological Research using a Modified Stroop Test: Proof of Concept

Adrijan Božinovski
School of Computer Science and Information Technology
University American College Skopje
Skopje, Macedonia
bozinovski@uacs.edu.mk

Sara Temelkovska
School of Computer Science and Information Technology
University American College Skopje
Skopje, Macedonia
saratemelkovska5@gmail.com

Sanja Manchevska
Medical Faculty, Department of Physiology
Ss. Cyril and Methodius University
Skopje, Macedonia
sanja.mancevska@medf.ukim.edu.mk

***Abstract* - The Stroop test is a neuropsychological test that registers the color-word interference tendency, meaning the impairment of the reading speed or color recognition due to interfering information. In simpler words, the interference tendency means that if the color word and color ink do not match, the time needed to name the color is always longer than when the color word and color ink match. This type of test is applicable to children and adults alike. Our Modified Stroop Test application implements a specific testing concept, based upon three types of stimuli within the Stroop test. The testing time lasts 15 minutes, or 5 minutes for each part of the Stroop test, but for our proof of concept the application lasts 60 seconds, or 20 seconds for each part of the Stroop test. The final interference results are calculated based on the average reaction time and number of correct answers. Our Modified Stroop Test application, which was tested on 30 people, showed the existence of the interference tendency during the split attention task and how its its possible effects on everyday life. New variations of the Stroop Test can be added in the future to this application in order to improve the interference calculations. The goal of this application is to be a starting point for future research, with a hope to produce interesting scientific results.**

Keywords—Stroop test, colors, reaction time, interference, neuropsychological assessment, stimuli, number of correct answers

I. INTRODUCTION

The Stroop effect in psychology is best known as a demonstration of semantic interference in the reaction time of a task and is often used to assess automatic processing versus conscious visual control in humans [1]. When a color name is written in different color ink, it takes longer to name the color and is more error prone than when the color name matches the color ink. This Stroop effect is named after its inventor, Ridley Stroop, who was the first person that wrote a paper about this issue. His paper was published in 1935 in England [2].

The stimuli in the Stroop test can be divided in three groups: neutral, congruent, and incongruent. In the neutral stimuli, only the text or the color is displayed. In the congruent stimuli, the color name (text) and the color ink are referring to the same color (e.g., the word pink is written in pink ink). And in the incongruent stimuli the color ink and the color word differ.

The most important finding from these experiments is the aforementioned semantic interference, which states that naming the color ink in the neutral stimuli is easier than in the incongruent stimuli. In addition, there are conditions that are not influenced by the interference, such as the word for the color and the color ink naming speed.

II. PROBLEM SOLVING

A. Theoretical Approach

In the Stroop test, the impairment of the reading speed or color recognition happens due to interfering information. To validate this assumption, various types of Stroop test versions can be used. The version developed for the purposes of this paper is applicable to children and adults alike.

Our Modified Stroop Test application implements a specific testing concept, based upon three types of stimuli. The first type belongs to the group of neutral stimuli, the second type belongs to the group of incongruent stimuli and the third type is a combination of neutral and incongruent stimuli. All of these stimuli are tested in separate parts of the Stroop test, which are connected to each other, and should be executed in the aforementioned order. The main goal is to press the correct button as quickly as possible. In our proof of concept, the testing time is 60 seconds in total, or 20 seconds for each part of the Stroop test.

The final results regarding the interference are calculated based on the reaction times and number of correct answers, using many different formulas, according to literature [1]. Here, we show two which are used the most. These two formulas are not the best option for calculation in our design (since they have been created for Stroop tests that were set up differently), but they support this proof of concept. The first formula calculates the interference using equation (1) [1].

$$IG = CW-((W*C) / (W+C)) \quad (1)$$

In this formula, CW represents the number of correct answers from Stroop II (the second part of the Stroop test), W is the number of correct answers from Stroop I (the first part of

the Stroop test) and C is the number of correct answers from Stroop III (the third part of the Stroop test).

The interference is also calculated using equation (2) [3].

$$I = \text{Stroop III}-((\text{Stroop I} + \text{Stroop II}) / 2). \quad (2)$$

In this formula, Stroop I, Stroop II and Stroop III test related variables are the average reaction times from each of the three parts of the test respectively and the total result is represented in milliseconds.

Until today, the validity of the Stroop test is confirmed by lots of comparative examinations (e.g., [1], [2], [3], [4], [5]). This test is important in many ways and shows how the brain processes information and what effect the interference has. In order to show that, our Modified Stroop Test application was tested on 30 people, but the testing time lasted 60 seconds (enough time for proof of concept), or 20 seconds for each part of the test. For the final interference result, both of the formulae were used.

B. Computer Application Description

For the development of this application, the C# language and the Microsoft Visual Studio development environment were used. The application consists of 10 Forms, ordered sequentially and executed as such (Fig. 1). Going forward and backward is only available between Form 1 and the About form (which displays information about the program developers).

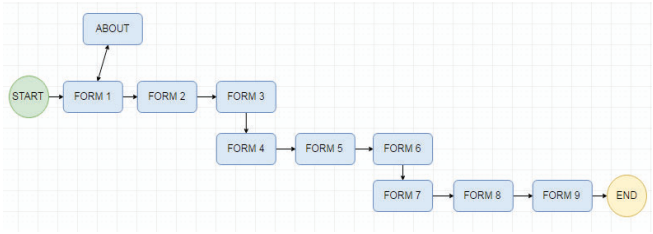


Fig. 1. Flow of the application

When the application is started, the first thing that the user will see is the Form 1, shown on Fig. 2.

Fig.2. Form 1 display

This form welcomes the user to the Modified Stroop Test application and has two buttons. The first buttons have a text “About” and when the user clicks on this button, form About shows details about the developers of the application. The second button, that can be clicked, contains the “Start” text, leads to the next form, Form 2, shown on Fig. 3.

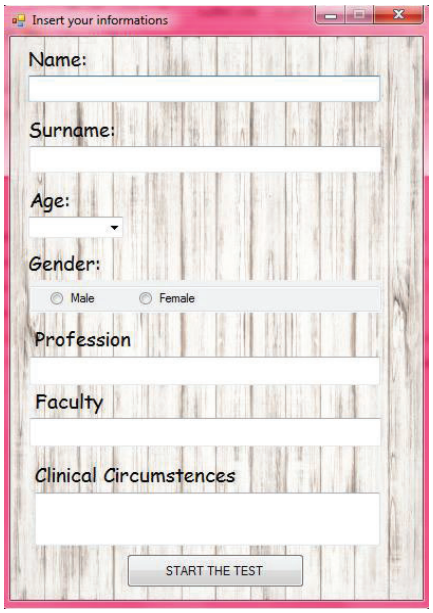


Fig.3. Form 2 display

This Form 2 is used for inserting information about the user (the person that is being tested) such as: name, surname, age, gender, profession, faculty and clinical circumstances. The first five of them are mandatory, but the other two are not. If all of the requested fields are filled, and the user clicks the "Start" button, the Save dialog box shows. The file name contains the name and the surname of the user that were previously inserted and the current date. The user has an option to save or modify the file name. After the file name is confirmed (or a new one entered) and the Save button is clicked, a program message confirms that the new user is inserted successfully. Further on, all the information about the user’s behavior during these tests will be simultaneously recorded in this file, until all of the tests are done.

When the button “Ok” is clicked, Form 3 shows (Fig. 4). In Form 3, instructions for the first part of the Stroop test are shown and the time in seconds, needed to read the instructions fully, is being recorded (although not shown to the user).

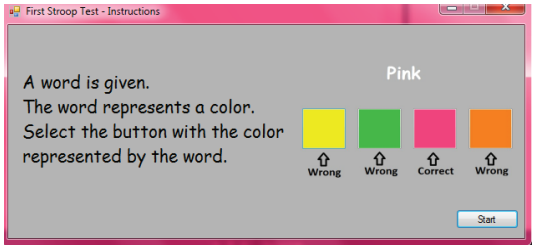


Fig.4. Form 3 displays the instructions for the first part of the Stroop test

After the instructions are read, when the button “Start” is clicked, the time needed for reading the instructions is written into a log file created in the Form 2, and Form 4 is shown, thus starting the first part of the Stroop test (see Fig. 5).

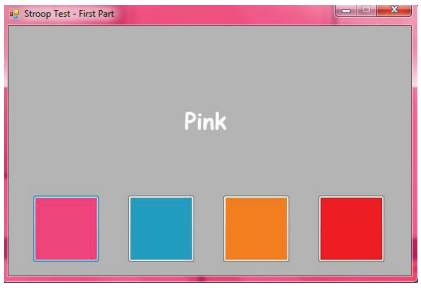


Fig.5. Form 4 display, the first part of the Stroop test

In this part of the test, there is a label containing a name of the color, and there are four buttons painted in different colors. One of those colors must be the same as the meaning of the word (which represents a color), and that is the correct answer. When the user clicks on one of the colors, the form is reloaded, and a new word (representing a color) and color options are given.

In the background, when the user clicks on one of the buttons, the answer is checked and is written in the log file (created in Form 2), together with the reaction time (time needed for answering each question in milliseconds). After the duration of the first part of the Stroop test elapses (20 seconds in the proof of concept presented in this paper), the program control is transferred to Form 5, which shows instructions for the second part of the test (Fig. 6).

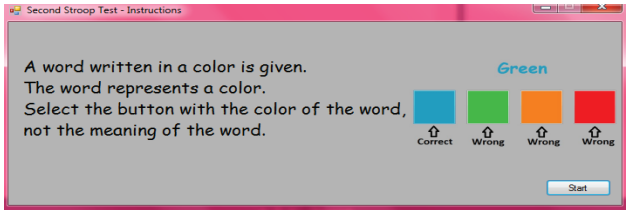


Fig.6. Form 5 display, instructions for the second part of the Stroop test

In Form 5, instructions for the second part of the Stroop test are shown and, again, the time in seconds needed for the user to read the instructions fully is recorded (and again, not shown to the user). After the instructions are read, when the button "Start" is clicked, the time needed for reading the instructions is written into the previously created log file and Form 6, i.e., the second part of the Stroop test, starts, as shown on Fig. 7.

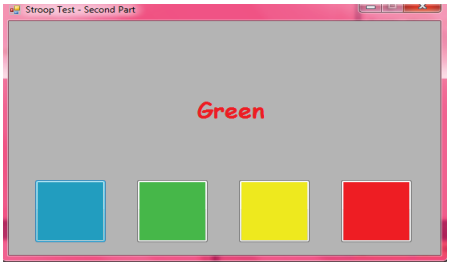


Fig.7. Form 6 display, the second part of the Stroop test

In this part of the test, there is a label containing a word representing a color, written in a different color ink, and there are four buttons in different colors. One of the colors that are given must be the same as the color ink of the word, and that is the correct answer, not the color that is the same as the meaning of the word. To make the test more difficult, a button with the color matching the word is also added, but that is not the correct answer. When the user clicks on one of the colors, the form is reloaded, and a new word representing a color, written in a different color than its meaning, and color options are given. When the user clicks on one of the buttons, the answer is checked and, together with the reaction time (time needed for answering each question in milliseconds) are written in the log file. The second part of the Stroop test again lasts for 20 seconds in this proof of concept, and, after that time elapses, Form 7 loads (Fig. 8).

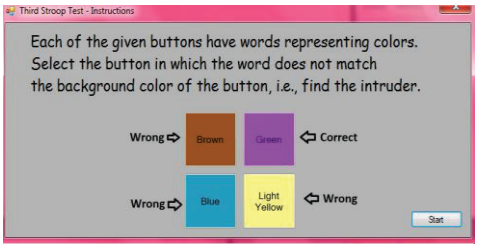


Fig.8. Form 7 display, instructions for the third part of the Stroop test

In the Form 7, instructions for the third part of the Stroop test are given and the time in seconds needed for the user to read the instructions fully is again recorded (again, not shown to the user). After the instructions are read, when the button "Start" is clicked, the time needed for reading the instructions is written into the log file and Form 8, i.e., the third part of the Stroop test, starts, as shown on Fig. 9.

In this part, there are sixteen buttons shown, all having different colors. Each of the buttons contains a text that represents the name of the button's color. Only one of the buttons' text does not match with its background color, and clicking that button is the correct answer. When the user clicks on one of the buttons, the form is reloaded, and new button colors with texts representing colors are given.

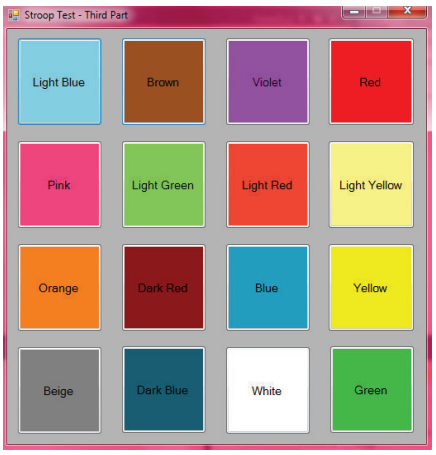


Fig.9. Form 8, the third part of the Stroop test

When the user clicks on one of the buttons, the answer is checked and written in the log file, together with the reaction. The third part of the Stroop test again lasts 20 seconds in this proof of concept, and, after that time passes, Form 9 loads (Fig. 10).

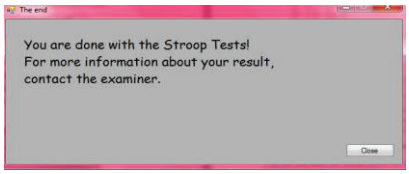


Fig.10. Form 9, End of the Stroop test

Form 9 is the last form of the application, and there is a text saying that the Stroop tests have finished. When the button "Close" is clicked, Form 9 closes and the application is terminated. All details about the user and the user's results are stored in the log file.

III. Results

Our application was tested on 30 people and has shown to work properly by giving valid results. Table I lists the details about the subjects such as their age, gender, total average reaction time, and total number of correct answers, aggregated from all parts of the Stroop test. The last two columns represent the interference results based on average reaction time (using (2)) and number of correct answers (using (1)). The last two rows show the average (mean) and standard deviation values from all of the columns.

To verify the scientific merit of the application, the t-test was performed on the groups of 18 female and 12 male subjects in all of the categories shown in Table I. The male users have demonstrated an insignificantly slower average reaction time (p=0.295) and an insignificantly more pronounced interference effect measured by the reaction time (p=0.217). Thus, further testing on more subjects is needed, in order to find out whether the gender of the subjects plays a role in the interference in the reaction time, as tested by our Modified Stroop Test application. It is estimated that a minimum of 100 subjects would need to be tested, so that these effects would be better studied.

TABLE I: RESULTS FROM THE MODIFIED STROOP TEST PROOF OF CONCEPT TESTING

User	Age	Gender	Stroop I		Stroop II		Stroop III		Total Average Reaction Time (in ms)	Total Correct Answers	Interference result	
			Average Reaction Time (in ms)	Correct Answers	Average Reaction Time (in ms)	Correct Answers	Average Reaction Time (in ms)	Correct Answers			based on average reaction time (in ms)	based on number of correct answers
User 1	21	Female	1912	8	2214	6	3429	1	2518	15	1366	6
User 2	28	Male	1348	5	1663	4	3767	3	2259	12	2262	3
User 3	32	Female	1210	9	1530	6	4383	3	2374	18	3013	4
User 4	53	Male	1636	7	3062	5	5156	1	3284	13	8207	5
User 5	52	Female	1418	15	1268	15	3907	2	2197	32	2564	14
User 6	28	Female	1759	9	1693	11	8831	1	4094	21	7105	11
User 7	21	Male	1205	16	1077	18	6370	2	2884	36	5229	17
User 8	53	Female	1414	13	1508	12	3570	3	2164	28	2109	10
User 9	24	Female	1508	9	1660	10	3529	3	2232	22	1945	8
User 10	57	Male	1428	12	1081	15	7591	1	3366	28	6337	15
User 11	42	Female	1212	13	1235	14	5964	2	2803	29	4741	13
User 12	22	Female	1085	18	1301	15	3806	1	2064	34	2613	15
User 13	25	Male	1181	16	919	20	2598	4	1566	40	1548	17
User 14	15	Female	1035	17	1234	15	4168	3	2145	36	3034	14
User 15	29	Male	1242	15	925	19	2519	4	1562	38	1436	16
User 16	24	Female	1249	15	944	16	9922	1	4038	32	8826	16
User 17	33	Male	2370	8	2180	3	9857	2	4802	13	7582	2
User 18	22	Female	999	10	931	9	3231	1	1720	20	2266	9
User 19	24	Female	786	11	1303	12	2755	0	1614	23	1711	12
User 20	26	Male	1784	9	2125	8	4987	1	2965	18	3033	8
User 21	37	Male	1388	6	2106	5	1656	2	1716	14	1900	5
User 22	45	Female	1959	9	1183	15	4551	2	2564	26	2980	14
User 23	43	Female	1815	11	1494	7	2269	1	1859	19	615	7
User 24	21	Male	1745	9	2384	7	8319	0	4149	16	6255	7
User 25	60	Female	2571	5	2562	3	4506	3	3213	11	1940	2
User 26	18	Female	1020	16	1895	8	4150	1	2355	25	2693	8
User 27	37	Male	1722	10	1330	13	2751	3	1934	26	1225	11
User 28	28	Female	2398	6	1615	8	4781	1	2931	15	2775	8
User 29	39	Female	1278	13	1290	13	1849	1	1849	27	1695	13
User 30	18	Male	2114	8	2814	7	6857	1	3928	16	4393	7
Average (mean)	32.57	/	1526.37	10.93	1517.53	10.67	4734.30	1.80	2638.30	23.43	3446.60	9.90
Standard deviation	12.80	/	446.87	3.77	579.93	4.92	2283.46	1.10	881.06	8.52	2272.98	4.60

IV. Conclusion

The Modified Stroop Test application presented in this paper demonstrates how the interference tendency can be tested for and measured. Our application can be used in different settings such as clinics, in other Stroop Test researches or in studies connected with neurodegenerative diseases. With further research and development of the application, we hope to produce interesting scientific results.

References

[1] Scarpina, F. and Tagini, S., 2017. The Stroop Color and Word Test. Frontiers in Psychology, 8(557).

[2] Stroop, R., 1935. Studies Of Inteferece In Serial Verbal Reactions. [Online] Available at: http://psychclassics.yorku.ca/Stroop/ [Accessed 20 February 2018].

[3] Elst, W. V. d., Boxtel, M. V., Breukelen, G. V. and Jolles, J., 2006. The Stroop Color-Word Test. Assessment, 13(1), pp. 62-79.

[4] Manchevska, S., Pluncevic Gligoroska, J., Bozhinovska, L. and Tecce , J., 2012. Attention and learning in medical students with different levels of anxiety and depression. Physioacta, 6(2), pp. 53-62.

[5] Schuhfried, G., 2011. In: Vienna Test System. Mödling: Schuhfried, p. 80

Implementation of Alert-Management System Using Centralized Log Server

Goran Petkovski, Evgenija Stevanoska, Boro Jakimovski, Goran Velinov
Faculty of Computer Science and Engineering
University Sts Cyril and Methodius
Skopje, Macedonia
{goran.petkovski, evgenija.stevanoska}@students.finki.ukim.mk, {boro.jakimovski, goran.velinov}@finki.ukim.mk

***Abstract*—Modern computing systems generates huge quantity of log data segmented in different categories. This includes high security messages, hardware diagnostics or critical system tasks. Analyzing and extraction of important message data requires respectively knowledge on the information contained in log messages and describing formal methods for efficient parsing, filtering and querying. This paper provides implementation of centralized log-server, which receives log information from two environments: server infrastructure and database system. Use case scenarios that are observed include inspecting SQL commands from database system and analyze security logs from productional server. Logstash is used as software tool for preprocessing of ingested parallel data streams and forwarding encapsulated data messages to other data sources. Results can be used in various scenarios such as detection and prevention of software crash incidents, security vulnerabilities and optimization of services and processes. Furthermore, this system is improved with alert-management system and task schedule based on the importance of discovered computer incident.**

***Keywords*—Alert system, Database systems, Formal methods, Logstash, Server infrastructure**

I. Introduction

Increased use of Internet requires proper monitoring and alerting of security policy violation, faulty hardware or system problems. In today's modern operating system every aspect of the computer system is tracked using log files. These are composed by log entries, including: users login activity, network statistical information, device diagnostics message and other service diagnostics.

Developing infrastructure for real-time analyze of log files that comes from different machines, requires huge storage and computational challenge. Several popular open-source software tools provides efficient filtering of enormous incoming data collected from different system platforms. Fluentd, is open source data collection tool, that use unified logging layer for structuring log messages using JSON format. This allows to combine all facets of processing log data using one logical layer in preprocessing stage[1].

We use Logstash, as main log analyzer. Combined with Elasticsearch and Kibana in one pack called (ELK stack), these tools are powerful analyzer of anomalies in various platforms (database systems, operating systems, virtual machine platform, network infrastructure) or even in monitoring bank transactions as described in this user-case story[2].

In our use-case scenarios, we use log data from two operating system platforms. First, we analyze database and system logs from Windows Server 2016 that is used as e-grade system for students in the faculty. Secondly, we analyze Linux server which stores login information about user login activity on productional servers.

Paper is structured as follows: In Section II we describe similar implementation using log-management tools and their advantages related to this subject; in Section III we describe infrastructure and process of interchange between different components in our solution; in Section IV we demonstrate processing of SQL queries in Microsoft SQL Server and usage of Logstash filter; Section V provides analyze of security audit messages in Linux server and filtering discovered attributes; Section VI demonstrate alert system for detecting unusual user activity; and Section VII concludes our work and presents ideas for future work.

II. Related Work

As part of improving monitoring, visualizing and reporting computing activities in their Large Hadron Collider (LHC) experiments, CERN started to evolve towards data analytics technologies including Apache Hadoop and Apache Spark for scalable processing framework, Docker for isolated environment with processing jobs and Collectd with Elasticsearch for collecting data metrics from various data sources inside their data center as described in [3]. With the new integrated system, they handle approximately 500GB data per day, thus being able to process complete critical system information.

Another implementation of ELK stack is used at Istituto Nazionale di Fisica Nucleare (INFN) in Torino. Their computing center offers IaaS services, that is managed through OpenNebula cloud controller. Communication with Logstash is provided with MySQL plugin that stores data metrics about their virtual environment. As they show in their use-case[4], this system is flexible to use with another environment and can be populated with different categories of metrics.

III. Log Management Infrastructure

Main component for building this infrastructure is using the open-source software stack Logstash, Elasticsearch and Kibana (ELK), which function as complete log management

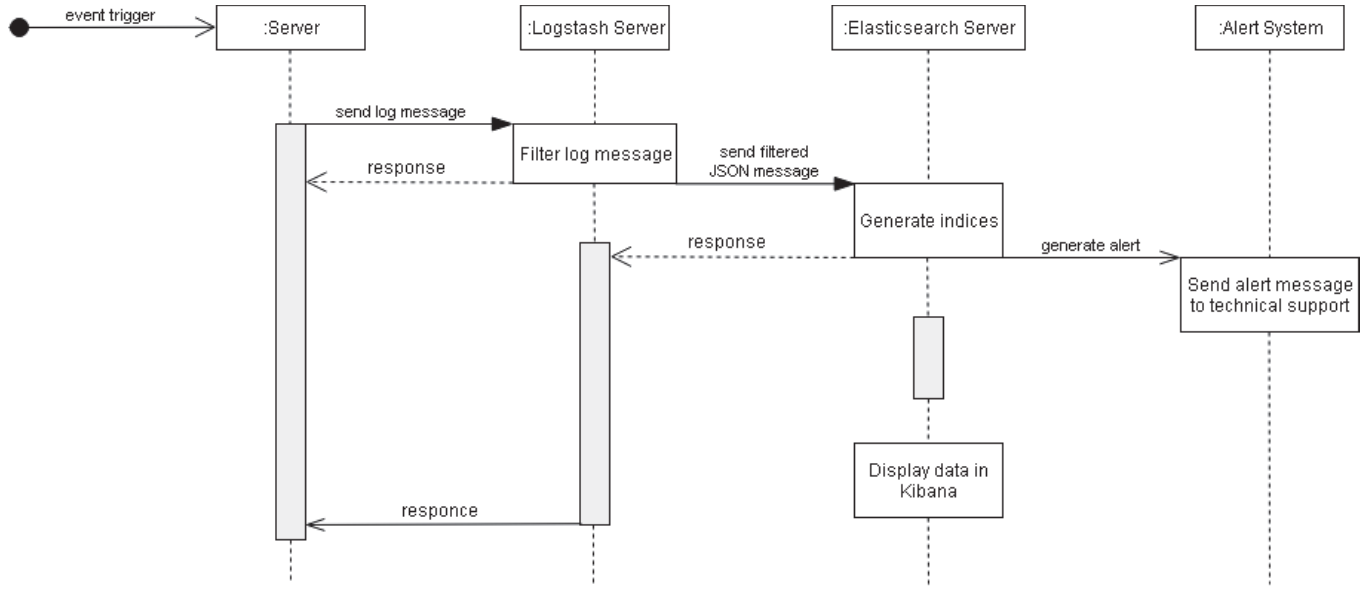


Fig. 1. Sequence diagram of sending log messages.

solution. As shown in Fig. 1, the process of analyzing log data consists of several steps.

First phase is generation of log-data based on some computer event or service handle. In background, service called filebeat is waiting for changes on inspected log-files at productional server machines. After it detect changes, it sends TCP request for establishing connection with the Logstash server. This message is encapsulated in JSON format with several additional meta data fields such as timestamp, source and hostname.

After sending the message using filebeat, Logstash server applies filters, mutation or aggregation, generating meta-data fields into new encapsulated JSON message. Once this process finish, this message is sent through TCP message to Elasticsearch server. If the server is not reachable, Logstash stop the service and can't proceed other incoming log-messages. This anomaly is exceeded with temporary buffer which stores undelivered messages until the Elasticsearch server is available.

Elasticsearch server according to the source of the log message, generate indices and store the metadata field into own database structure. Along with that, it sends alert message to the alerting system for notifying user if some anomaly or security intrusion is detected in the message content. In this phase, response message acknowledge is sent back to Logstash server, for confirmation of completed transaction. This process continues with displaying monitoring data in Kibana dashboard, user reports or tables based on query results.

A. Logstash

Central component for collecting log messages is Logstash. It is open source log-management utility and crucial part of the ELK stack[5]. It consists of three stages during its operation:

- Input phase: Gather logs and events from different platforms, databases and applications. Most common inputs are: file, beats, syslog, http, tcp, udp, stdin but also can ingest data from different data source.
- Filter phase: Support extremely powerful filter plugins that enables to analyze, manipulate, measure or create other events. Along with formal methods, it's efficient way to extract data patterns contained in common log message.
- Output phase: Logstash can push data to various locations, services and platforms. It can output to raw file, CSV, JSON, XML, Elasticsearch and Amazon S3, or convert them into message with RabbitMQ.

B. Elasticsearch

Elasticsearch server is responsible for efficient storing and indexing of log data messages. Main engine is based on Apache Lucene and provides distributed, multitenant full-text search with HTTP web interface provided by Kibana, described in [6]. Also it supports creation of visualizations and dashboards based on some regular expression or context message.

IV. Analyze of Microsoft SQL server queries

This operation consists on detection of CRUD operation in SQL database by non-privileged users.

A. Creating event logs in Windows Event Viewer

4SQL Server Management Studio provide built-in audit specifications inside database schema. Those consists of: database audit specification (logging on user activity in database, such as CRUD operations,triggers or events) and server audit specification (logging on users login activity to database server), as described in [7]. In our case we create

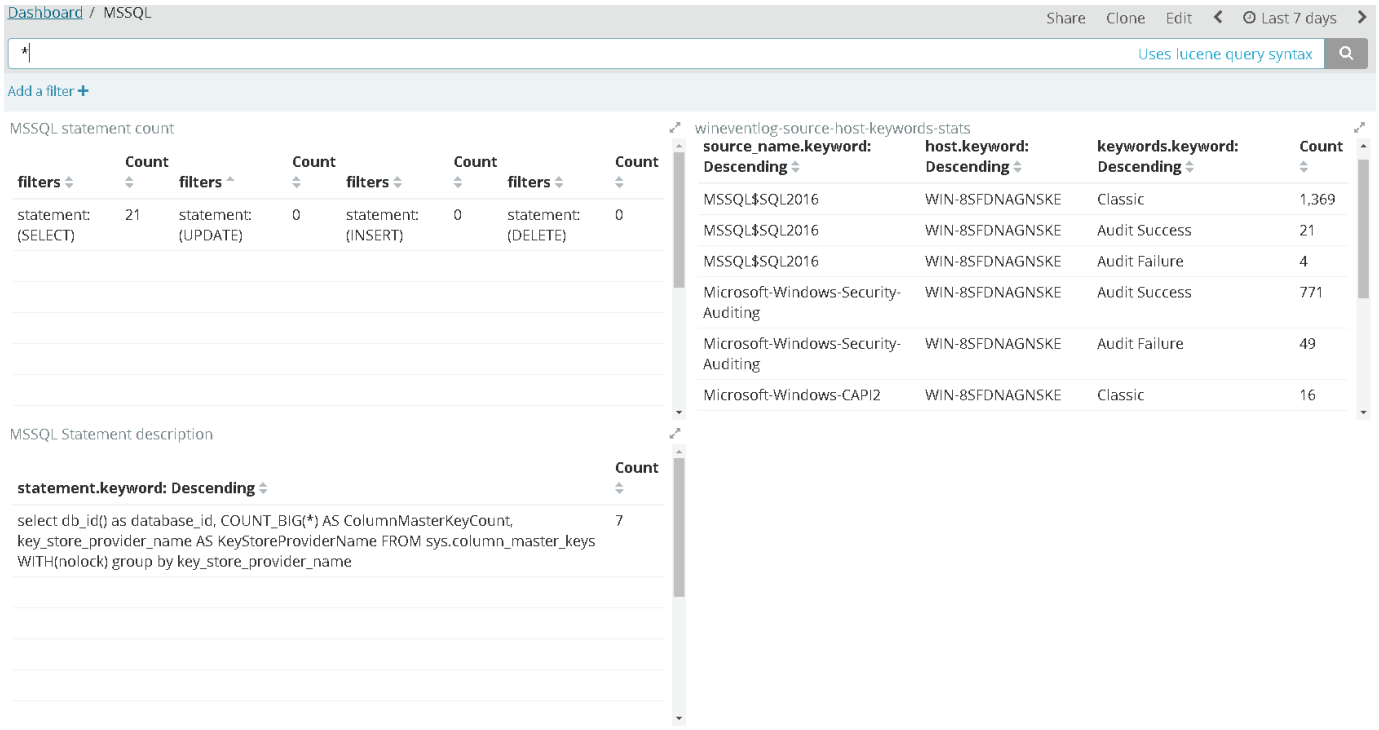


Fig. 2. Dashboard of Microsoft SQL events.

database audit specification that records logs when users execute SELECT command on any tables in the e-grade database schema. Additional requirement is it must be non-privileged, non-applicative or default database user. In Event Viewer, the message contains important fields such as:

- session_id: Identification number associated with open process.
- transaction_id: Identification number associated with SQL statement.
- database_principal_name: Database where the statement is executed.
- statement: Message containing SQL commands.

B. Analyze logs with Logstash filter

In order to perform analyze, Logstash filter must contain all required field that are used in the original message. Below is shown filter with grok options:

```
grok {
match => { "message" => "(?m)Audit event:
audit_schema_version:%{NUMBER:
    audit_schema}
event_time:%{TIMESTAMP_ISO8601:event_time
    }
session_id:%{NUMBER:session_id}
transaction_id:%{NUMBER:transaction_id}
database_name:%{GREEDYDATA:database_name}
statement:%{GREEDYDATA:statement}"
}
}
```

Important notice is the sign of ***(?m)*** which indicate persistence of fields. This means that some fields are optional, they can be omitted in the original event message and Logstash ignores them when it comes to the filtering phase.

C. Use Case scenario of Microsoft SQL log data

Tables with results are shown on Fig. 2. We use three visualization tables, first table display the number of commands in every SQL statement that is executed on the SQL Server, second table shows the type of event that was triggered along with keywords added by Event Viewer. And the third table shows details about SQL statement along with count meter.

V. Analyze Security Logs in Linux Server

Another issue presented in our solution was to develop security policies, that track every users activity in our production servers. Those includes ssh intrusions from another locations, prevention of undesirable system commands and protection of personal information leakage.

A. Analyze logs from Linux server using Logstash

Linux maintain security messages in different system paths based on the type of its distribution. The syntax of the log-file is same in both cases and includes:

- syslogtimestamp: Timestamp when the log has occurred.
- program: Program used for logging to the Linux server.
- process_id: Identification number of ssh process.
- auth_user: User with user-name that tried to login to the server.

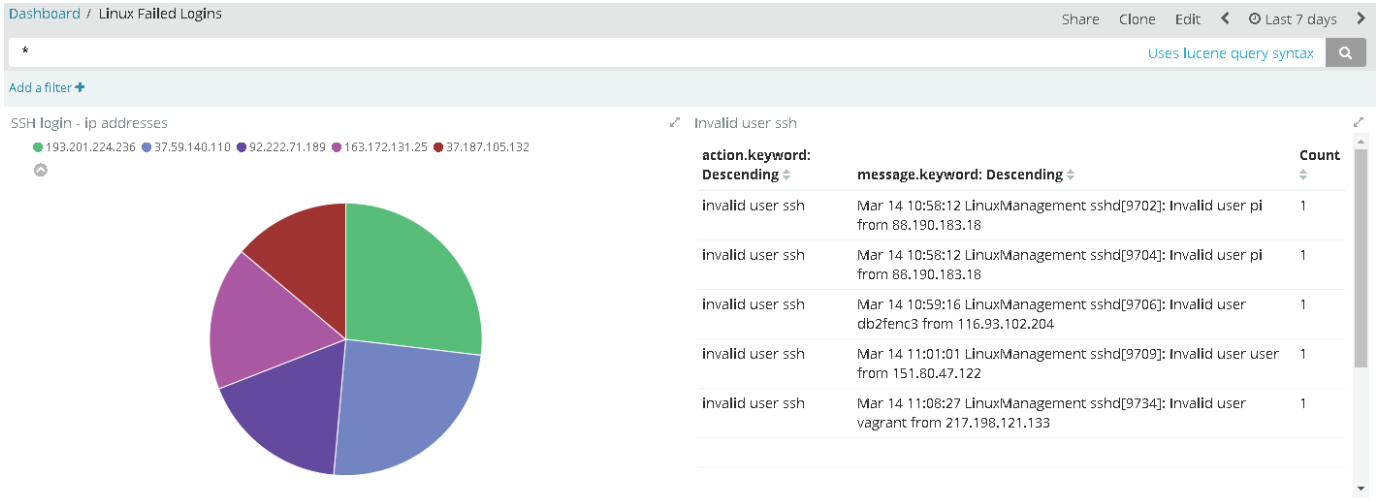


Fig. 3. Display of IP address logged with invalid user.

- ipv4address: IPv4 address from where the user has tried to login.

These fields are mandatory in every scenario, with other optional fields. They describe additional information about user and include: command that user executed on the server and duration of session which user has logged to the server.

Logstash filter for detecting user login through ssh is given in below code:

```
grok {
#unknown user SSH
match => [ "message", "%{SYSLOGTIMESTAMP:
    timestamp}
%{PROG:program}\[{NUMBER:pid}\]: Failed
    password for invalid user from
%{IPV4:ipv4address} port %{NUMBER:port}
%{WORD:protocol}" ]
add_field => { "action" => "failed ssh
    login" }
add_field => { "connection" => "ssh" }
add_tag => "unknown_user"
add_tag => "alert"
}
```

Another common security message is tracking session number opened for logged users on console:

```
grok {
#connect SSH
match => [ "message", "%{SYSLOGTIMESTAMP:
    timestamp}
%{PROG:program}\[{NUMBER:pid}\]:
pam_unix\(sshd:session\): session opened
    for user by %{GREEDYDATA:uid}" ]
add_field => { "action" => "login" }
add_field => { "connection" => "ssh" }
add_tag => "session_opened"
add_tag => "alert"
}
```

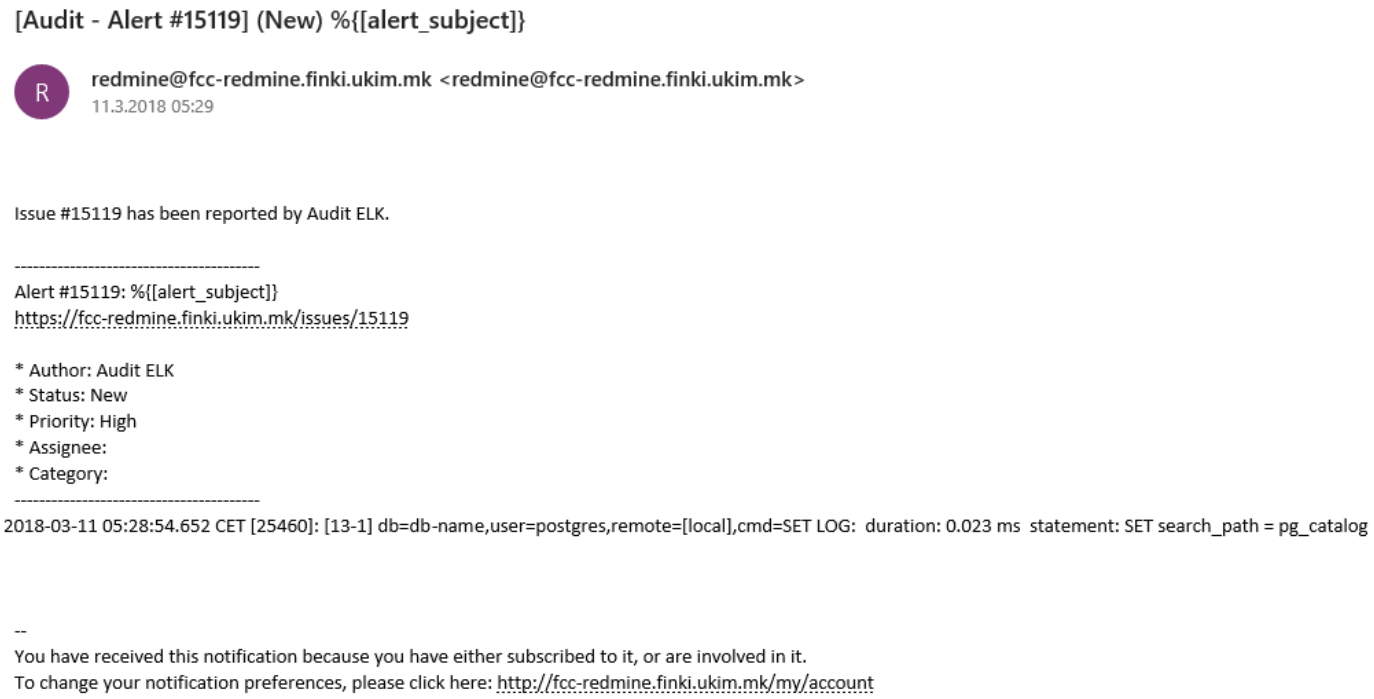
B. Use case scenario in Linux

In Fig. 3 are shown ip addresses from where users has tried to login through ssh protocol. Pie-chart display percent of total number of incorrect login divided by the number of ip addresses which is investigated. This chart can be extended with the hostname of the machine where the event has triggered and timestamp of the event.

C. Metrics with Logstash filter

Another advantage using filters is metrics specifications and its operations for aggregation (count, sum, avg). Using aggregate function requires mapping data field to external variable and perform the needed operations. For example, if we want to calculate total duration of queries that contains **SELECT** keyword in its SQL statement message we use the following code:

```
if [commandTag] == "SELECT"
{
 aggregate
 {
task_id => "%{session_ID}"
code => "map['sql_statement'] = map['
    sql_statement'] + event['statement']"
code => "map['sql_duration'] += event['
    duration']"
map_action => "update"
 }
}
```


[Audit - Alert #15119] (New) %{[alert_subject]}

R redmine@fcc-redmine.finki.ukim.mk <redmine@fcc-redmine.finki.ukim.mk>
11.3.2018 05:29

Issue #15119 has been reported by Audit ELK.

Alert #15119: %{[alert_subject]}
https://fcc-redmine.finki.ukim.mk/issues/15119

* Author: Audit ELK
* Status: New
* Priority: High
* Assignee:
* Category:

2018-03-11 05:28:54.652 CET [25460]: [13-1] db=db-name,user=postgres,remote=[local],cmd=SET LOG: duration: 0.023 ms statement: SET search_path = pg_catalog

--
You have received this notification because you have either subscribed to it, or are involved in it.
To change your notification preferences, please click here: http://fcc-redmine.finki.ukim.mk/my/account

Fig. 4. Alert email message for SQL command.

VI. Alert Management System

Another feature in our system is using built-in plugin for alerting users about security incidents. This plugin sends alert message to user when detects metafield tag with alert keyword, as shown in the example. Fields that are included in the original Logstash message can be sent to users, so they can be informed about the incident.

Our implementation uses Redmine plugin, available as Logstash output plugin that generates new event to Redmine issue tracking system. Configuration is provided in separate file that is parsed after input and filtering stage.

```
output {
  if ("alert" in [tags]) {
    redmine {
      token => <token specified to
          project id>
      url => "https://fcc-redmine.finki.
          ukim.mk"
      project_id => 8
      tracker_id => 4
      status_id => 1
      priority_id => 3
      ssl => true
      subject => "%{[alert_subject]}"
      description => "%{[message]}"
    }
  }
}
```

As described in above code, **subject** and **description** are fields populated with data from filtered Logstash messages. Those can be extended with another meta-data fields according to use-case scenario and user requirements.

Example message that is sent to user is shown in Fig. 4. This alert message shows execution of SQL command on production Linux server.

VII. Conclusion and Future Work

Processing log data in today modern computing is challenging and complex job due to enormous generated information from different systems and platforms. In this paper, we present infrastructure for analyzing two categories of log information coming from database system and productional server.

Main parts of the solution are open source software tools, packed in one ELK stack. Pipeline processing occurs in three phases: input data stream (collect data information from events and logs), filtering stage (apply formal logic and generate filters to analyze and refine crucial information). And in final phase, output stage (send alert notification about security incident or suspicious SQL statements to user and generate visualizations or dashboards on Kibana web interface).

We shown scenarios for analyzing SQL queries based on CRUD operation and inspecting audit logs for detection of user login intrusion. These events are monitored every day with several technical support persons who are responsible for manual intervention.

In future work, we want to extend the domain of usage on different platforms (network switches and routers) for detection of network congestion and automatic reaction in case of detected problems. We also plan to apply this solution in

virtual infrastructure (Vmware Esxi), and monitor the status of all virtual machines.

ACKNOWLEDGMENT

This work was partially financed by the Faculty of Computer Science and Engineering at the Sts. Cyril and Methodius University in Skopje.

REFERENCES

[1] ”What is Fluentd?”, Fluentd.org, 2018. [Online]. Available: https://www.fluentd.org/architecture. [Accessed: 24-Mar-2018]

[2] ”How Wirecard uses the Elastic Stack to monitor transactions”, Elastic Blog, 2018. [Online]. Available: https://www.elastic.co/blog/how-wirecard-uses-the-elastic-stack-to-monitor-transactions-and-analyze-errors. [Accessed: 24- Mar- 2018]

[3] A. Aimar, A. Corman, P. Andrade, S. Belov, J. Fernandez, B. Bear and M. Georgiou et al., ”Unified Monitoring Architecture for IT and Grid Services”, Journal of Physics: Conference Series, vol. 898, p. 092033, 2017.

[4] S. Bagnasco, D. Berzano, A. Guarise, S. Lusso, M. Masera and S. Vallero, ”Monitoring of IaaS and scientific applications on the Cloud using the Elasticsearch ecosystem”, Journal of Physics: Conference Series, vol. 608, p. 012016, 2015.

[5] ”Logstash Introduction”, Elastic.co, 2018. [Online]. Available: https://www.elastic.co/guide/en/logstash/6.2/introduction.html. [Accessed: 24-Mar-2018]

[6] ”Basic Concepts of Elasticsearch”, Elastic.co, 2018. [Online]. Available: https://www.elastic.co/guide/en/elasticsearch/reference/current/_basic_concepts.html. [Accessed: 24- Mar- 2018]

[7] ”Windows Event Log (Windows)”, msdn.microsoft.com, 2018. [Online]. Available: https://msdn.microsoft.com/en-s/library/windows/desktop/aa385780(v=vs.85).aspx. [Accessed: 24-Mar-2018]

External Factors Destabilizing the Operation of Data Centers

Rosen Radkov
Department of Software and Internet Technologies
Technical University of Varna
Varna, Bulgaria
rossen.radkov@tu-varna.bg

***Abstract*—The operation of a Data Center is dependent on multiple external factors: disasters, incidents, emergencies, accidents. They may, directly or indirectly, affect its operation in such a way that its activity or the security of information is impaired in whole or in part. Therefore, it is important to assess the impact of these factors when designing a data center and to make adequate decisions to suppress their impact. An analysis and classification of the external factors have been conducted and a model of the surrounding environment has been created.**

Keywords—data center, availability, integrity, risk assessment

I. INTRODUCTION

Ensuring the integrity and availability of data is one of the most important tasks in the field of information technology. This is determined by the dependence of business processes on the IT Infrastructure (ITIS), which ensures their implementation. Contemporary business environment conditions are such that this dependency is almost 100% and it is required of ITIS to provide an acceptable interruption of business processes, which in most of cases, especially when it comes to large enterprises, is reduced to a few minutes or even seconds. Failure to meet the requirements of continuity of business processes and disaster recovery leads to loss of confidence and reputation in front of clients and partners, and financial losses, often resulting in the cessation of business.

Data Centers (DCs) are the core of the ITIS of each organization and therefore must be designed and built to meet the set requirements. However, the functioning of a DC is also dependent on many external factors: disasters, incidents, emergencies, accidents. They may, directly or indirectly, have an impact on the functioning of the DC in such a way as to impair partially or entirely its activity or the security of the information. Therefore, it is important to assess the impact of these factors and to take adequate decisions to suppress it when designing a DC or assessing its quality parameters.

The purpose of this article is to analyze the factors that exist in the environment and affect the performance of a DC, causing damage, errors, and failures [1] of individual components, subsystems, or systems, thereby disturbing or ceasing to provide the services expected from it. These factors will be called destabilizing factors (DFs).

II. ANALYSIS OF DESTABILIZING FACTORS

A. Types of disasters

Based on an analysis of literary sources [2], [3], a disaster (catastrophe) is defined as a significant disturbance to the normal functioning of society caused by natural phenomena and/or by human activity and resulting in negative consequences for the life or health of the population, private property, the economy and the environment, the prevention, containment and overcoming of which exceeds the capacity of the system servicing the usual activities aimed at the protection of society. It is therefore an event or a series of events that are possible but unlikely and difficult to predict.

The analysis of literary sources [2]–[8] has shown that disasters can be classified into two categories:

- natural: cosmogenic, lithospheric, atmospheric, hydrospheric and biogenic phenomena. The cosmogenic phenomena include solar storms, cosmic rays, the fall of massive cosmic bodies (meteorites, comets, asteroids), the gravitational impact of space objects, the landing of foreign extraterrestrial microorganisms, etc. The lithosphere consists of earthquakes, landslides, volcanic eruptions, erosion, conversion of Earth's magnetic poles, radiogenic zones, deposits of harmful earth elements. Atmospheric phenomena are related to events, such as cyclones, storms, hurricanes, tornadoes, hailstorms, frost and drought, global climate change and landscape fires. Hydrospheric phenomena include tropical cyclones, tsunamis, sea level fluctuations, floods, droughts, and the biogenic ones are: invasion of different types of macro- or microorganisms, which causes damages of technogenic nature; infectious diseases - epidemics, pandemics; damage to crops from diseases and pests [3];
- man-made: technological (cyber attacks, chemical threats, biological bombs, pollution, power outages, equipment failure, etc.), industrial (fires, explosions, etc.), military (wars, coups, etc.), social (civil unrest, strikes, terrorist attacks, etc.).

Natural disasters can be predicted to a certain extent by analyzing statistical data and the changing trends, both of which are collected by relevant specialists (geophysicists,

meteorologists, etc.), whereas the technological ones cannot be predicted, but an assessment of the risk of their occurrence can be made and, depending on it, the location of the DC can be determined.

Determining the DC location should also be tailored to the possibility of disasters. Examples of risk grading depending on the location of the DC in case of natural or man-made disasters can be found in the Uptime Institute DC standard [9], which are presented in Table I and Table II.

TABLE I. SITE LOCATION-NATURAL RISK CATEGORY

Natural Disaster Risk Category	**Scale of Risk**	
Component	*Higher*	*Lower*
Flooding and tsunami	< 100-year floodplain	> 100-year floodplain
Hurricanes, tornadoes and typhoons	High	Medium
Seismic activity	>0.8m/s^2	<0.8m/s^2
Active volcanoes	High	Medium

TABLE II. SITE LOCATION-MAN-MADE RISK CATEGORY

Man-Made Disaster Risk Category	**Scale of Risk**	
Component	*Higher*	*Lower*
Airport/Military airfield	<3 miles from any active runway; inside a 1x1.5-mile runway extension	>3 miles from any active runway; outside a 1x1.5-mile runway extension
Adjacent property exposures	Chemical plant, fireworks factory, etc.	Office building, undeveloped land, etc.
Transportation corridors	< 1 mile	> 1 mile

The threats and consequences of different types of disasters have changed their character over the years. In recent years, the danger of social disasters has increased considerably: terrorist attacks, refugee flows, etc., all of which can seriously affect the accessibility and efficiency of DCs.

B. Statistics on the types of disasters and the losses caused by them

Statistics on the number of crisis events in Bulgaria and the damages caused by them for the period 2010-2016 based on information from the National Statistical Institute are presented in Table III.

The statistics show that the number of events is not determinant of the amount of damage. The statistics do not indicate damages to businesses due to data integrity and accessibility disruptions caused by technological disasters, such as power surges, sudden power outages, fires, etc.

TABLE III. CRISIS EVENTS IN BULGARIA IN THE PERIOD 2010-2016

Indicators	**Number of reported disasters**	**Total amount of damages (thousand BGN)**
Fires	38.49%	11460
Landslides	1.38%	567749
Earthquakes	0.13%	60238
Droughts	0.16%	268
Floods	8.04%	660977
Storms, tornadoes, whirlwinds	2.00%	164864
Hail	0.23%	52288
Snowstorms (snowdrifts)	1.33%	8988
Freezing, frosting	0.97%	465
Accidents	2.00%	1083
Vehicular accidents	44.64%	24719
Contamination (with chemicals, hazardous waste, household waste, etc.)	0.34%	30158
Human epidemics	0.11%	5
Animal epidemics (including birds)	0.05%	84
Calamity	0.02%	431
Other	0.10%	579

According to DataMagic Inc. only 3% of data losses occur due to natural disasters and 97% happen because of employees. But it is a good idea to assess the likelihood of a particular disaster and the impact of this event on business as well as to assess the risk to business because the statistics indicate that the floods, which are only 8.04% of all disasters, cause the biggest damage.

C. Impact of disasters on the DC operation

As a result of the impact of the DF on the operation of DCs, the following events can take place:

- Inaccessibility of the DC premises and therefore impossibility of: servicing its infrastructure, supply of raw materials and consumables for DC and replacement of staff;
- Interruption of main and backup power supplies;
- Interruption or inability to use the communication lines;
- Unavailability of the ITIS components;
- Unavailability of the provided services;
- Inability to fulfill contractual obligations to clients and suppliers.

An assessment of the impact that the different DFs can have on the performance of DCs is done through risk assessment methods. The risk assessment for the same type of DC differs

depending on its territorial location due to the difference in DFs that occur in the respective region.

As a result of the impact of DF on the functioning of the DC, the following can be affected to the greatest possible extent:

- staff life;
- the DC buildings (following an earthquake, flood, fire, hurricane, terrorist attacks, catastrophes, etc.);
- engineering subsystems that provide vital raw materials (electricity, water, air conditioning);
- ITIS components (hardware, system and application software);
- data medias (backups and archive copies);
- documentation in paper or electronic form;
- data.

According to the National Archives and Records Administration in Washington D.C. 93% of companies that lost data for 10 days or more due to disaster filed for bankruptcy within one year. Half of all business without data management for this same time period filed for bankruptcy immediately. This means that if one wants to survive, each company must develop and implement adequate business continuity and disaster recovery plans.

III. Model of the surrounding environment

Based on the analysis of the different DC components and the specifics of their functioning and the external factors influencing their work, a model of the surrounding environment has been created and presented in Table IV, including the main threats to the integrity and availability of the data, the destabilizing factors that can lead to them, attributed to the relevant disaster category.

Using the presented model would help to produce a risk assessment that identifies all risk sources by identifying possible threats, likelihood of their occurrence, and potential consequences. The purpose of this step is to generate an exhaustive list of risks based on events that could affect the functioning of DCs. Full identification is crucial because the risk that is not identified at this stage will not be included in a further analysis. Identification should include risks, whether their source is under the control of the organization or not. It is a fact that not all DFs have the same impact on the functioning of DC. Some DFs depend on the territorial location of the DC and there are cases where a DF will never occur (example: Varna is not endangered by a volcano eruption). Not all DFs can partially or totally damage a DC, so when designing high-reliability DC, those that are potentially possible are selected. The calculation of their impact is done on the basis of risk assessment. Principles and guidelines for risk management and risk assessment are presented in international standards ISO 31000 and ISO 31010 [10], [11].

TABLE IV. Model of The Surrounding Environment

Threat	Disaster type	Destabilizing factors
Disturbance of data integrity	Technological	Power interruption
		Damage of HVAC system
		Using poor quality hardware and software
		Implementing inadequate solutions for IT security
		Cyber attacks
Disturbance of data availability	Natural	Demolition of the DC buildings as a result of geophysical, meteorological, hydrological and climatological activities like: hurricanes, floods, landslides, earthquakes, tsunami, etc.
		Fires near the DC
		Damage to the DC equipment due to cosmogenic and atmospheric phenomena like static electricity
		Damage to the DC engineering equipment due to meteorological phenomena like extreme temperatures
		Demolition of the communication lines used by the DC due geological activities like landslides, earthquakes, etc.
	Technological	Demolition of buildings due to industrial accidents: explosion, fire, collapse, gas leak, poisoning, etc.
		Demolition of buildings due to vehicular accidents
		Damage to engineering facilities inside DC in proximity to important parts of IT infrastructure: water mains, HVAC, etc.
		Interrupted cable lines due to construction and assembly activities
		Damages to equipment due to fire or explosion
		Power interruption
		Use of outdated or non-prospective equipment
		Implementing inadequate IT solutions
		Cyber attacks
	Social	Terrorist attacks
		Mass riots and strikes
		Stampedes
		Military actions
		Other
	Biological	Epidemics: viral disease, bacterial disease, etc.
		Pandemic
		Insect infestation

Many methods can be used for risk assessment and one of the best ones is Failure Mode and Effects Analysis (FMEA). FMEA is a method designed to identify potential damages for a product or process, the possible causes for their occurrence and the effect of their occurrence on the organization. The method allows the classification and identification of significant risks and the identification of the appropriate corrective actions to address the most serious problems. Performing failure analysis and its effects can lead to predicting and preventing problems, reducing costs, shortening of product development times, and delivering safe and highly reliable products and processes.

FMEA has the potential to be a very powerful tool for achieving high product and process reliability and, when performed accurately, it is extremely effective. FMEA is an inductive risk assessment tool that identifies the risk by a Risk Priority Number (RPN) as a product of the following metrics according to the formula:

$$RPN = S \times O \times D \quad (1)$$

where

- S – is a non-dimensional number that stands for severity, i.e. an estimate of how strongly the effects of the failure will affect the system or the user;
- O – denotes the probability of occurrence of a failure mode for a predetermined or stated time period – even though it may be defined as a ranking number rather than the actual probability of occurrence;
- D – means detection, i.e. an estimate of the chance to identify and eliminate the failure before the system or customer is affected. This number is usually in an inverse relationship with the severity or occurrence numbers: the higher the detection number, the less probable the detection is. The lower probability of detection consequently leads to a higher RPN, and a higher priority for a resolution of the failure mode.

Each of the indicators has a maximum value of 10, therefore the maximum possible risk is 10 x 10 x 10 = 1000. The method does not specify certain RPN values, which are to be considered as limits and against which to be determined whether the risk is low, high. But according to good practices, the following division is considered:

• low risk (acceptable risk) - <125, the risk is acceptable and the controls applied are good, i.e. it does the functioning is not endangered;

• medium risk (unacceptable risk) - 125 ÷ 300, a risk that can be controlled and action should therefore be taken to reduce RPN <125;

• high risk (unjustified risk) - > 300 and S> 7, a risk that is likely to cause major damage and immediate action must be taken. Typically, this risk is outsourced because it is not wise to be managed by the organization.

The EN 60812 standard [12] uses tables to determine the values of the three components S, O and D.

IV. Conclusions

The analysis of the impact of the surrounding environment on the operation of the DC shows the need to assess the risk of occurrence of each of its DFs. On the basis of this assessment, adequate decisions can be made in regard to the composition of the DC and the organization of its management in order for the organization that uses the DC to be protected from violation of the integrity and accessibility of its data and any negative consequences that would stem from the lack of such protection.

Applying the proposed environmental model when assessing the risk allows for the inclusion of all potential threats to DC operation.

References

[1] A. Avizienis, J. Laprie, and B. Randell, “Fundamental Concepts of Dependability,” 2014.

[2] МС, Закон за защита при бедствия. 2017, p. 50.

[3] Р. Илиев, “Природни бедствия - същност, проявление и класификация,” in Четвърта научна конференция с международно участие “Географски науки и образование”, гр. Шумен, 2015, p. 6.

[4] МС, Национална програма за защита при бедствия 2014-2018. 2013, p. 73.

[5] DRI, “10 Predictions for 2018 : DRI ’ s Third Annual Predictions Report,” 2018.

[6] EU, “Зелена книга относно застраховането срещу природни и причинени от човека бедствия,” 2013.

[7] ICDO, “International Civil Defence Organisation” [Online]. Available: http://www.icdo.org/en/disasters/.

[8] Restore your Economy, “Disaster Preparedness • Economic Recovery • Resilience.” [Online]. Available: http://restoreyoureconomy.org/disaster-overview/types-of-disasters/.

[9] Uptime Institute, Data Center Site Infrastructure Tier Standard : Operational Sustainability. 2014.

[10] ISO, ISO 31000, vol. 2009. p. 34.

[11] International Organization for Standardization, “ISO/IEC 31010:2009 Risk management - Risk assessment techniques,” vol. 31010, p. 92, 2009.

[12] CENELEC, EN 60812 "Analysis techniques for system reliability - Procedure for failure mode and effect analysis (FMEA)", 2006.

Analysis and Evaluation of Data Center Quality Indicators

Rosen Radkov
Department of Software and Internet Technologies
Technical University of Varna
Varna, Bulgaria
rossen.radkov@tu-varna.bg

***Abstract*—Data Centers are an important part of an organization's IT infrastructure. In order to ensure the normal operation of the business processes in the organization, the data center needs to provide services with the expected quality. The article proposes an approach for analyzing and evaluating the values of the data center quality indicators and assessing their compliance with the project values. The approach involves installing means to monitor the IT infrastructure's components and apply a method of analyzing its reliability.**

Keywords—data center, availability, disaster recovery, business continuity

I. Introduction

We can hardly imagine the modern world without the presence of information technology. It is all around us. We use it when we are at work when we are relaxing, and even when having fun. With its help we can keep track of our rhythm of life throughout the day. Running any of these activities in a way that will please us depends on our justified expectations of the quality of the services we use. This is not possible if the high-quality IT infrastructure (ITIS) that is needed to ensure these services is not provided. Failure in the services we use when relaxing or having fun is inconvenience, disappointment, and in some cases a loss of money, but when it comes to services necessary to ensure business processes, it is related to loss of large amounts of money, trust, reputation, and even human life.

Each ITIS changes during its operation due to a change in the values of various indicators: number of hours worked by each of its components, amount of processed and stored data, number of serviced users, changes in business process requirements, destabilizing environmental factors, and others. The following issues arise for the organizations for which ITIS are built:

- How can they be sure that the ITIS built has the qualities outlined by the project?
- How can uninterrupted high-quality support of ITIS services be ensured?

This article presents the author's approach to finding answers to the questions asked.

II. Components of IT Infrastructure

Each ITIS is unique in its composition and purpose, however, its components can be grouped according to the tasks performed.

A. Detailed Description of ITIS

According to research conducted on literary sources [1]–[4], a typical ITIS structure for an organization includes DCs, a local network at each of the organization's sites, and Internet connections, presented on Fig.1.

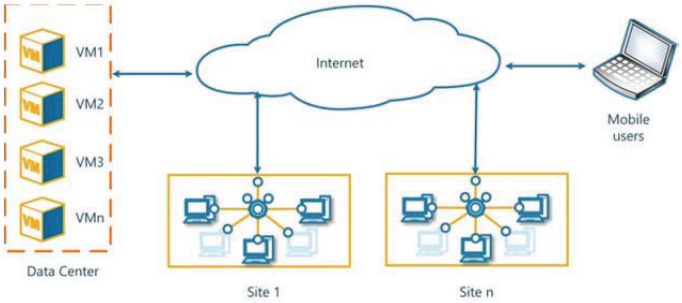


Fig. 1. Typical corporate ITIS

DC is the core of this ITIS and consists of the following components:

- Specialized premises;
- Structured cabling system;
- Communication lines;
- Access control system;
- HVAC;
- Fire detection and fire suppression systems;
- Power lines;
- Server and communication cabinets;
- Network infrastructure;
- Servers and storages;
- Workstations;
- IT staff.

DCs can be categorized according to whether they serve a private domain (corporate center) or a public domain (ISP's DC, Internet DC, collocated DC), and whether they use own, rented DC or collocated proprietary equipment.

A typical IT infrastructure of DC is shown on Fig. 2 [3].

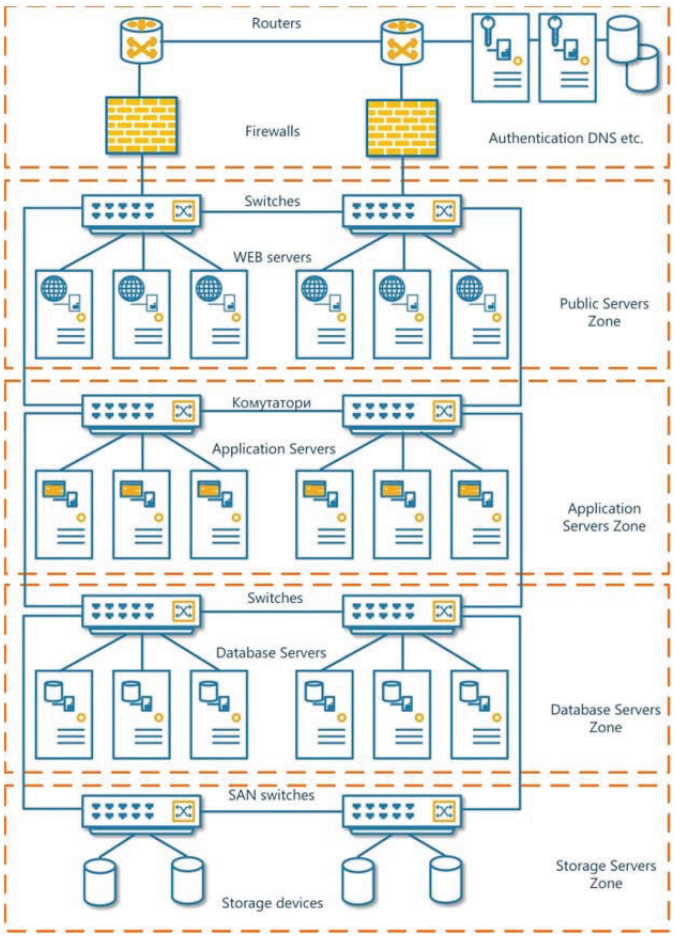


Fig. 2. Typical IT infrastructure of a DC

B. The Most Important Indicators of DC Quality

The quality of a DC can be assessed by different criteria by measuring indicators related to service availability, performance, energy efficiency, TCO, ROI, etc.

According to the analyses conducted [5]–[7] from the users' point of view, the most important indicator for assessing the quality of DC is availability. It is usually measured in percent or in number 9's. The availability is defined as a quantitative assessment of the possibility for users to access the services provided by DCs at any point in time when they are needed. The higher its value, the higher the quality of the DC. But higher quality is achieved at the expense of a more complicated IT solution and larger capital and operating costs.

Availability in turn is dependent on reliability, maintainability RTO, RPO, and others. Some of them are technical in nature, others are dependent on process management at the DCs. Another indicator that has a direct relation to customer satisfaction is the performance of the DC. It evaluates the time for processing the information and whether it corresponds to the user's requirements. This indicator is dependent on the utilization of the equipment and the capabilities to expand the capacity of the DC in response to the increased number of business processes, users, or the volume of data processed.

C. Approach for Indicator Evaluation

The proposed approach uses availability and performance indicators to assess the quality of an existing DC. Their values are determined on the basis of experimental data.

Analysis of the ITIS components and the DC structure shows that availability is a combination of the presence of communications (A_L) and accessibility of the components (A_C) of the DC. User connectivity to DC differs depending on the size of the organization and the number of its locations. For organizations that have their own DC located at the same site as the users, the connection to it is via a LAN, and most of the services it provides do not depend on other organizations. In the case of an organization whose users are located at many different sites, the connection to a DC is dependent on the quality of the services provided by the telecommunication operators supplying the Internet or the communication lines. Therefore, in order to have an adequate assessment of the availability of the DC, it is necessary to monitor the connections to it and the presence of its components.

In order to get answers to the above questions, it is suggested to install monitoring systems that take into account the following indicators.:

- The availability and quality of communications from DC to each location;
- Availability of DC components;
- Utilization of the DC components (CPU, RAM, Storage, network connections etc.);
- I/O connections to storage system.

Systems for monitrong use protocols like SNMP and ICMP to collect data from devices.

Since in one corporate ITIS not all locations are of equal importance, the organization should determine which ones to take into account when determining the aggregate accessibility factor.

Usually, monitoring systems use quantitative values, which are measured in time, number or percentages.

The availability of a separate component $A(t)$ is calculated using the formula:

$$A(t) = \frac{Uptime}{Uptime + Downtime} * 100\ [\%] \qquad (1)$$

where *Uptime* is a time during which a component is operational, and *Downtime* is the time when it isn't operational.

The indicator is usually calculated on an annual basis. Having in mind that, in order for the whole ITIS to work, it is necessary for all the components (DC and communication lines) to be in normal operating mode, then in a reliable sense a reliable circuit with sequentially connected components (series system) is obtained [8]. Therefore, the overall accessibility index K_{av} is calculated using the following formula:

$$K_{av}=\prod_{i=1}^{n} A_i(t) \quad (2)$$

where $A_i(t)$ is the availability of an ITIS component (DC and network connections) $i=1..n$.

Utilization of some of the elements of ITIS is reported as percentage with a maximum value of 100%. For quantification of other indicators, units are used, for example: input/output system load is measured in number of I/O operations. In order to calculate the utilization K_u, the number of units reported (operations, connections, etc.) is compared with the maximum number of units defined by the respective equipment manufacturer using the formula:

$$K_u=\frac{C_i}{C_{max}} * 100\ [\%] \quad (3)$$

where C_i is the value of the indicator and C_{max} is the maximum value defined by the equipment manufacturer. Table I gives the author's classification of the load of a component or system.

TABLE I. UTILIZATION RATING

K_u [%]	Rating
0-29	Low
30-59	Medium
60-84	High
85-100	Critical

III. APPLICATION OF THE APPROACH

In order to demonstrate the approach, a corporate ITIS, whose topology is presented in Fig. 3 is selected. The requirements that are set at the design stage are as follows: 99.9% availability and utilization rating - middle. It consists of eight nodes each one of which contains users of the system. Each node is connected to the Internet with one or two ISPs, as shown in Fig. 3. Connection between offices is built using IPSec VPN tunnels established and maintained via firewall devices at each node.

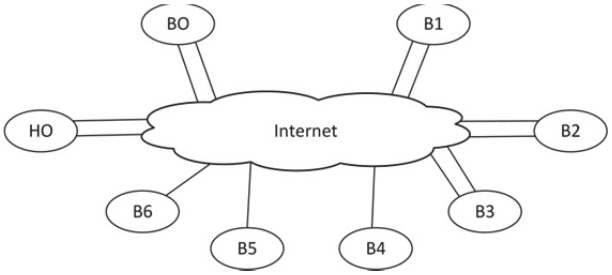


Fig. 3. Typical IT infrastructure of a DC

The components of DC are shown in Fig. 4. DC consists of two clusters: primary (Bourgas) and secondary (Varna). The primary one is located in the central node (HO) and it has two hosts (BS1 and BS2) and one storage server (SAN), and the secondary in the backup node (BO) has one host. Both clusters work in an active-passive mode.

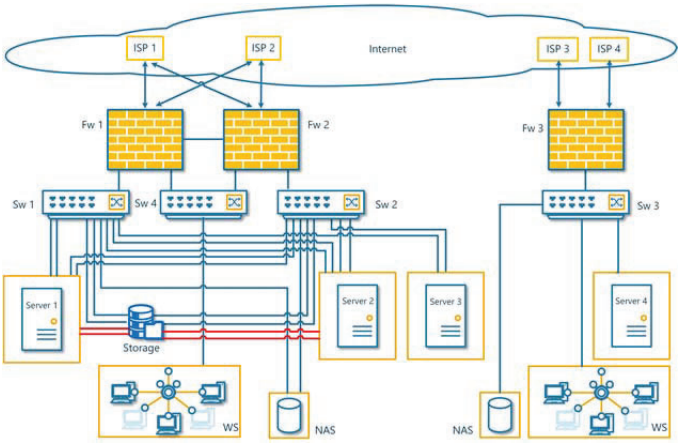


Fig. 4. Typical IT infrastructure of a DC

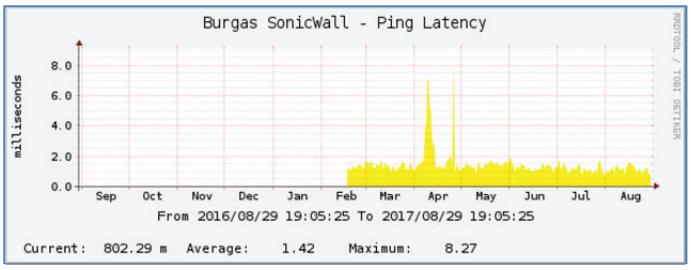
The critical business processes of the company require the availability of servers with the following roles: terminal server, database server, domain controller and DNS, and mail server. They are realized through four virtual machines. In the Bourgas cluster they are: RDS1, DBS1, DC1 and MS1 and in the Varna cluster these are: RDS2, DBS2, DC2 and MS2. Replication of data is done via applications.

The analysis of the selected quality indicators for DCs is based on the experimental data obtained through the installed monitoring systems. Four monitoring systems have been installed: to measure the availability of communications - CACTI, Observium for availability and utilization reporting, and storage system performance - STOR2RRD and LPAR2RRD. The control panel of one of the uninterruptible power supplies (UPS1) of the DC has been reported to have been active for 1 year 57 days 1 hour and 41 minutes. Therefore, the power supply system has worked well and there has been no interruption in the operation of ITIS due to problems with it, although the detailed overview of the UPS1 logs shows a number of registered events related to short-term power outages.

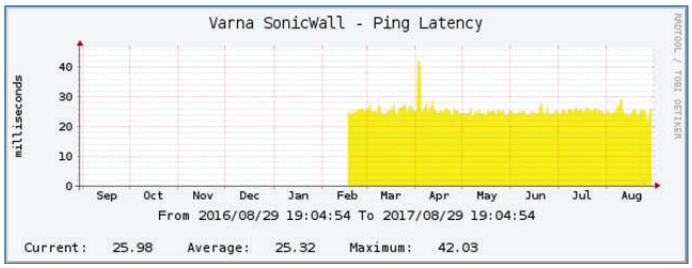
A. Network Connections Availability and Utilization

In Fig. 5a-h the graphs illustrate the presence of communication over the VPN tunnels to the firewalls (Fw) in each of the nodes for the time period 15.02.2017-29.08.2017.

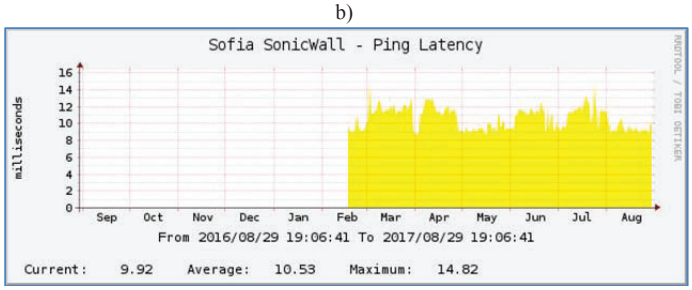
The graphs are based on tests performed by the CACTI monitoring system via ping, at a 5-minute interval. The presented results show that VPN tunnels were available throughout the entire timeframe.



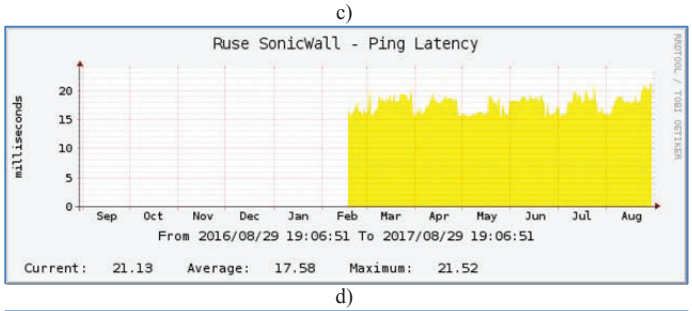
a)



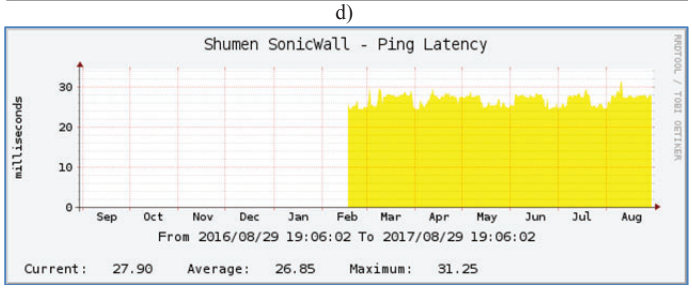
b)



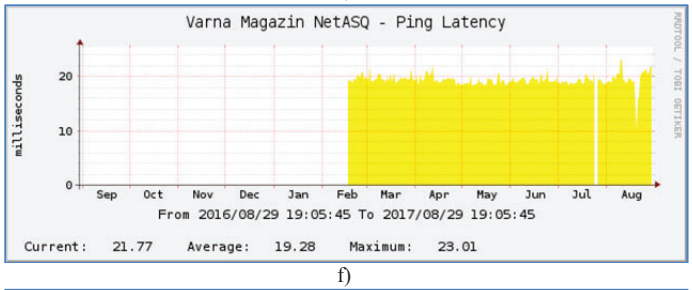
c)



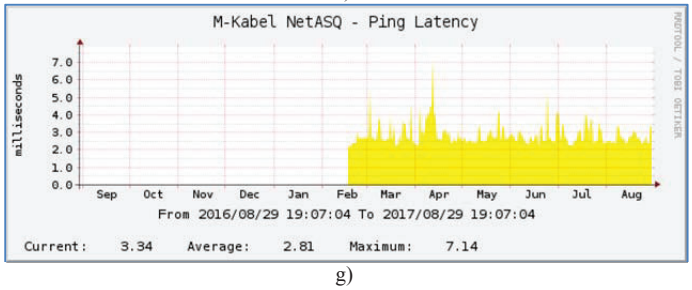
d)



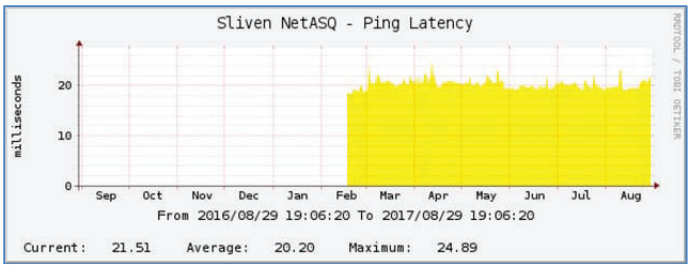
e)



f)



g)



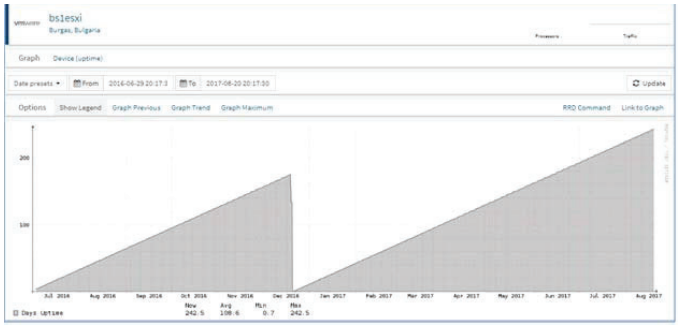
h)

Fig. 5. Availability of VPN tunnels: a - Fw (HO), b – Fw (BO), c – Fw (B1), d – Fw (B2), e - Fw (B3), f – Fw (B4), g - Fw (B5), h – Fw (B6)

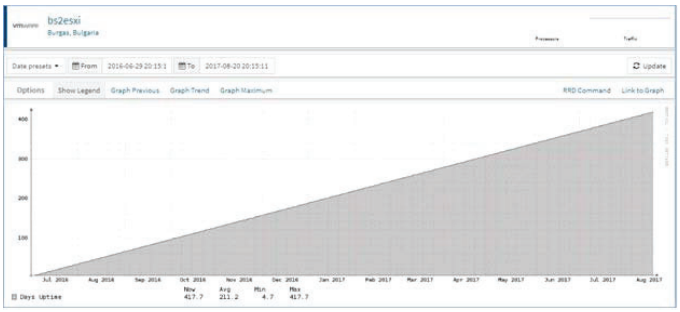
An interruption for several hours was observed only in the B4 node, between 18:40 on 23.07.2017 and 10:00 on 24.07.2017 due to a damaged firewall power supply. The average values of the connection latency are as follows: 1.42 ms for HO, 2.81 ms for B6, 10.53 ms for B1, 17.58 ms for B2, 19.28 ms for B4, 20.20 ms for B5, 25.32 ms for BO, and 26.85 ms for B3. The impression is that when the connection between the nodes takes place through the network of the same Internet provider, the latency is much lower, approaching the values typical of the local network. An example of this is the relationship between the nodes HO and B6, which is carried out through the MAN network of one supplier and between the HO and B1, which are connected to the same Internet provider. The other three knots have similar latency values of around 20 ms. Understandable are the values of the latency of the B3 node, which is connected through the supplier with the worst parameters, and that of the BO, in which the higher latency value is related to the large service traffic exchanged between the two nodes.

B. Availability and Utilization of the Servers

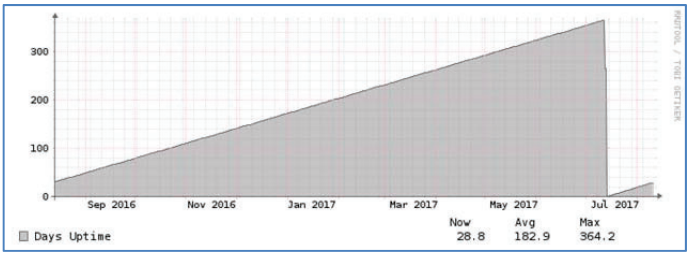
The analysis of achieved ITIS availability was conducted using experimental data collected through SNMP protocol and monitoring system Observium. Results on the status of all devices that depend on the ITIS work have been obtained. Figures 6a-d illustrate graphs generated by the Observium system, presenting the availability of the core cluster components. The graphs in Fig.7-a give statistics about number of users connected to server RDS1 and Fig.7-b and Fig.7-c give information about the CPU utilizations of the RDS1 - the highest load-carrying server. The statistics show that the average load for each of the four cores is between 13% and 15%. Fig. 7-c shows the change in processor load over a 24-hour period. At the same time, Fig. 7-d TCP connections is significant.



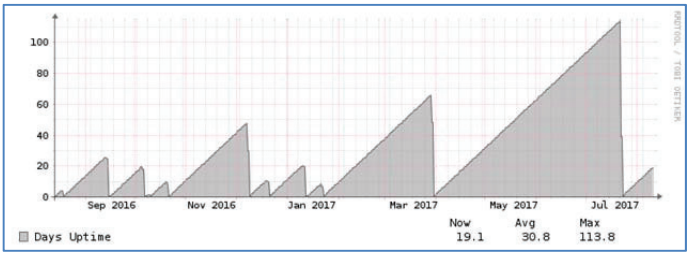
a)



b)

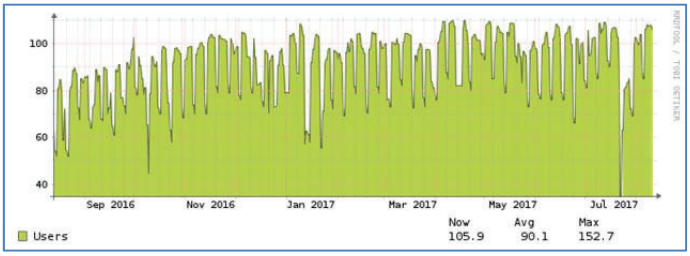


c)

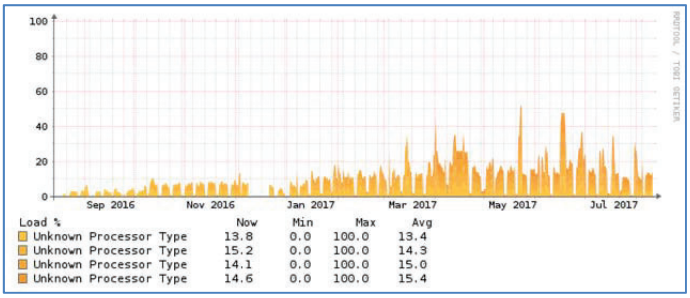


d)

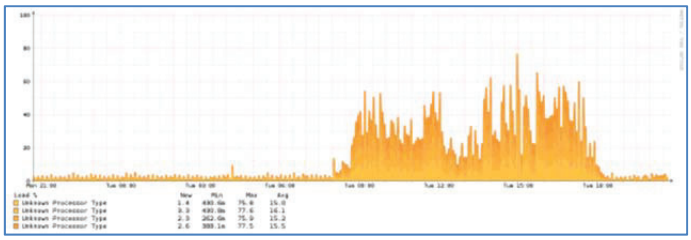
Fig. 6. Availability of the most important components of the Bourgas Cluster: a - hypervisor BS1ESX1 on host1; b - hypervisor BS2ESXi on host2; c - database server DBS1; d - Terminal server RDS1



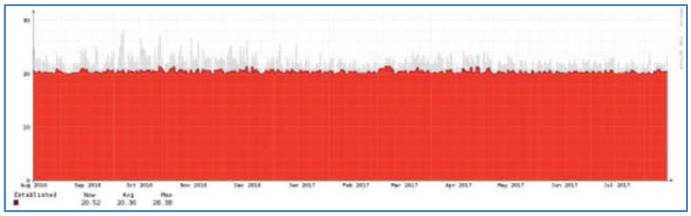
a)



b)



c)



d)

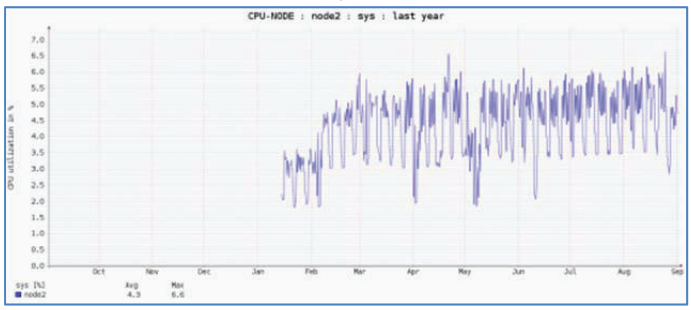
Fig. 7. Statistics about a - number of users on RDS1; b - CPU utilization of RDS1; c - CPU utilization on RDS1 for a period of 24 hours; d-TCP connections to RDS1

Therefore, the server configurations shown are selected correctly and without difficulty can take 4-5 times the load.

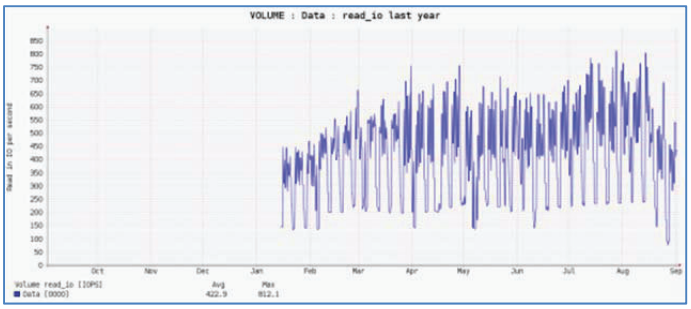
C. *Performance of the SAN System*

An important element of the DC, which is highly dependent on its productivity, is the storage system. Figures 8a-f show graphs taken from the STOR2RRD system showing load statistics of the processors of the two control modules as well as graphs for the I/O operations for both the Data and Quorum storage system volumes.

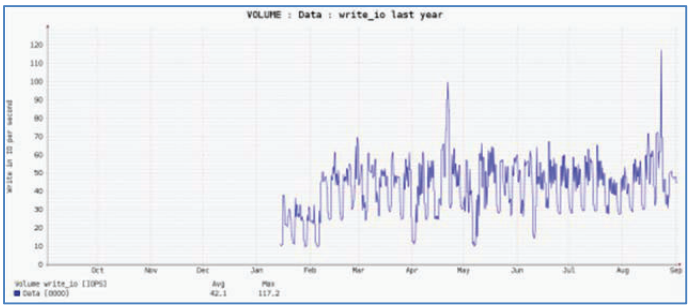
a)



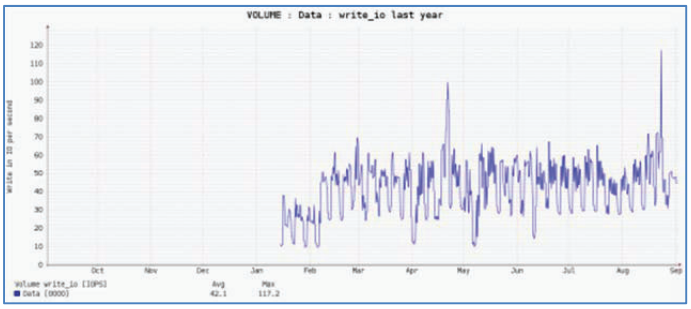
b)



c)



d)



e)

f)

Fig. 8. SAN utilization: a - CPU of node1, b - CPU of node2, c - I/O read to volume Data, d - I/O write to volume Data, e - I/O read to volume Quorum, f - I/O write to volume Quorum

The analysis of the graphs in Fig.8-a and Fig.8-b and of the studies conducted by Watson and Storage News categorically indicate that the storage system is far from reaching the limit. According to the above-mentioned sources, the total maximum number of I/O operations is about 80 000. It is seen that for the monitored period of time, the second storage module (node2) was active on the storage server.

D. Results Assessment

The availability indicators calculated by the monitoring systems, according to the company's working time, and the calculated total accessibility indicator of the system are presented in Table II. The system-reported values of the performance characteristics of ITIS are presented in Table III.

TABLE II. K_{AV}

Component	$A_i(t)$ [%]
DC (RDS,DBS,DC,MS)	100%
Line HO-BO	100%
Line HO-B1	100%
Line HO-B2	100%
Line HO-B3	100%
Line HO-B4	100%
Line HO-B5	100%
Line HO-B6	99.977%
K_{av} of ITIS	**99.977%**

TABLE III. UTILIZATION RATING BY COMPONENTS

Component	Utilization rating
RDS CPU	Low
RDS RAM	High
SAN CPU node 1	Low
SAN CPU node 2	Low
SAN I/O write	Low
SAN I/O read	Low

IV. CONCLUSIONS

From the observations and analysis of the work of the specific ITIS it can be seen that the achieved accessibility of the system for the monitored period is 99.977% and the load of the individual components is less than 16%. It has been concluded that during the monitored period of operation of the system it has worked according to preliminary expectations, there are possibilities for servicing of new business processes and it allows service and planned preventive maintenance without interrupting the business processes.

Applying the proposed approach to analyzing the functioning of ITIS allows for real results about for the operation of the system. From their analysis it can be deduced whether the quality indicators of the implemented ITIS correspond to the values set in the project, as well as whether the quality of services is as required.

REFERENCES

[1] Reichle & De-Massari AG, R & M Data Center. 2011.

[2] P. Turner, J. H. Seader, V. Renaud, and K. G. Brill, “Tier Classifications Define Site Infrastructure Performance,” 2008.

[3] R. Radkov, “A design approach for high reliable Data Center,” Technical University of Varna, Bulgaria, 2017.

[4] Cisco Inc., “Data Center Technology Design Guide,” 2014.

[5] ReliaSoft Corporation, “Availability and the Different Ways to Calculate It.” [Online]. Available: http://www.weibull.com/hotwire/issue79/relbasics79.htm.

[6] E. Vargas, “High availability fundamentals,” Sun Blueprints Ser., no. November, 2000.

[7] E. Dubrova, “Fault-tolerant design,” 2013, pp. 1–185.

[8] P. Antonov and R. Radkov, “Analysis of the Data Centers dependability characteristics,” Computer Sciences and Technologies, vol. IX, no. 1, pp. 43–49, 2011.

Improving the Efficiency of Business Processes in Catering Industry

Suad Saliu[1], Sasko Ristov, *Member, IEEE* [1,2], and Marjan Gusev *Senior Member, IEEE*[1]

[1]'Ss. Cyril and Methodius' University, Faculty of Computer Science and Engineering, Skopje
[2]University of Innsbruck, Distribute and Parallel Systems Group

***Abstract*—The Catering industry is often associated with inefficient business processes. Self-service mechanism (kiosks) is being introduced by big enterprises (McDonalds, Wendy's) in this industry to improve these issues and improve the sales process. The purpose of this study is to investigate and offer a solution for companies, which still hesitate to implement self-service in their work, focusing on Small and Medium Enterprises. Businesses from Macedonia and around the world reported the main barriers for implementing technology (self-service mechanisms) in their work. The hesitation comes from the high costs of implementation of the current solutions in the market and high maintenance costs. In this paper, we propose a solution that will improve the efficiency of business processes in catering industry by applying ICT. It also improves the business processes by using Artificial Intelligence. We introduce a Software as a Service solution which does not require any implementation on client side and has low maintenance costs. It also makes the process of ordering and booking services more efficient by reducing the waiting time of customers, mitigates the risks of human mistakes, reduces operating costs, and improves the overall data processing.**

***Keywords*—Software as a Service, SME, hospitality, catering, scalability, digitalization, efficiency, effectiveness, business processes, self-service**

I. Introduction

Due to the dynamics of the developing world, companies need to adapt to changes in real time and continuously improve themselves. It is necessary to continuously measure the efficiency and effectiveness of the performance of the work tasks and the business processes as a whole. One approach to make this happen is digitalization of the work tasks.

For example, if we look at catering sector and their services, there are several challenges to improve the efficiency and effectiveness of their business processes. Difficulties arise when they use the services and at the same time space for improvement of various business processes [1].

Firstly, there is a problem in organizing the work, respectively the lack of accurate information and analyzes of the data they have in connection with the performance of their business processes as well as the preferences of their clients (users of their services) [2]. Processing the data of caterers will lead to knowledge of how to redefine and restructure their business processes, which one of them are unnecessary, which ones are too long and which consume resources. They will know what their users from different categories prefer (younger, older, women, men, etc.), when there are bigger requirements, what happens with the holidays and various other circumstances. Very importantly, they now have a measure of accuracy for the employees, they have answers from users how satisfied they are from the offer, and much more.

Artificial Intelligence (AI) is also introduced here for the first time in this manner. With the use of current AI techniques and the data produced by the use of this solution, it will suggest caterers certain best practices and advices how to behave in order to maximize the profit, how to improve the customer experience and their business processes. Difficulties were also noted in the core business process that is the process of orders [3]. The long waiting time is an issue that should be considered, starting from waiting for taking the service offer to the final order and the inconvenience by calling several times the waiters, than violating the privacy in deciding, forgetting the orders, over-spending on time by employees while customers are being overlooked, calculation errors, etc.

For these processes there is a possibility for improvement, like cutting the non-productive spending of time and resources, reducing the risk of additional costs for errors, inability to have full control and insight on the work and processes, inadequate human resources, dissatisfied customers etc. Adding technology in the catering sector could do the job. Digital ordering in the catering sector is growing rapidly, with biggest brands like McDonald's installing kiosks and extending their plans for mobile order and pay, Starbucks testing a voice ordering app and Wendy's installing kiosks at 1,000 of its locations [4].

In this paper, we will present a model that has SaaS on the top of the infrastructure to be used from the businesses and the mobile application for clients. This model gives the business that operates in hospitality sector the opportunity to redefine and restructure business processes, to improve the efficiency and effectiveness of employees, the quality of managers who manage business processes and users who use their services. Reengineering of business processes with ICT assets can dramatically improve the performance of companies. The essence of their use is to save time and to have accurate real-time information about the work in the company, and by processing these data companies achieve even greater goals. With the use of ICTs, many human errors and cyclical repetitions of some tasks can be eliminated and financial and human resources are saved. ICT assets can provide a competitive advantage if they are used rationally. The monthly cost of getting these services is minimal for the businesses, with no need for infrastructure

and staff training.

The rest of the paper is organized as follows. In Section II, the current efforts in improving the industry are presented along with examples. Section III describes the model used and its main implementation challenges. Advantages and disadvantages of the model are discussed in Section IV. Finally, Section V represents the conclusion of this paper and planned extension of our work.

II. Related Work

There are various attempts to introduce catering technology, but most of them are specifically made for certain caterers like the McDonald's, Starbucks or Wendy'??s solutions mentions above. Other minor example is GetWaiter which integrate an architectural concept that partially is proposed by us.

GetWaiter concept is similar to our model, since businesses use their SaaS and clients have mobile application to use the services [5]. The main downside is their limited options. With GetWaiter as a client you have only four options: a) to call the waiter, b) to ask for more drinks based on your previous order, c) to cancel the order and 4) to ask for bill. With these limited options, businesses do not have much data to process and reengineer their business processes.

Regarding the small amount of services that our competitors are offering, we can say that we don't have any limitation on that or in other words almost every service that the businesses are offering to their clients it can be reflected with the solution presented here. That is because of a simple reason, we have a dynamic form from which the businesses itself can define the service that they are offering along with the behaviors. And this gives the opportunity of creating of almost infinite number of services using our solution.

Far more serious work has been done by QikServe. QikServe are constantly getting funds, till now they have won approximately 4 million dollars as investment [6]. But they differ in the model. It is made of a client application and Oracle POS system. The mobile or kiosk application, has features like seeing the menu, ordering and paying from the app. A downside of the mobile application is that you have to pay to process the order. However, Federal Reserve Board research declares that only 46 percent of US citizens have used mobile payments[7]. Their advantage is that they capture customer data in order to power more targeted campaigns for the businesses. But they have a big downside especially for SMEs. The caterers that want to implement their solution have to change their current infrastructure and way of work. They also have to contact QikServe directly, negotiate and make a deal with them directly and let them install the infrastructure that is needed. And this of course costs a lot of money and time, which can't afford SMEs.

We have exceeded all of these problems mentioned above. In our model, there is not an obligation to pay order, you can place it using the mobile application and pay cash instead. The model presented here includes all of the above listed features, from seeing the menu, ordering the dinner and the option to pay (very important that it is not obligatory) and also suggestions for restaurants, hotels and clubs located nearby with pictures, videos and captions depending on what businesses have entered and data processing. The difference is that to implement our solution, the businesses only need 2 minutes of time to access the SaaS. And this means that they don't need to change their infrastructure at all, we use their current implementations. They only have to pay a minimum cost monthly. This also means that the interested parties, doesn't need to contact us directly in order to negotiate and make a deal. This process is automatized.

It is worth mentioning that neither of the above mentioned solutions doesn't use AI to help the caterers. There are some efforts made in introducing AI especially in hotels (like the Marriot hotel-chain), but they are merely focused on improving the customer experience and not the overall work and business processes.

III. Business Model And Architecture

Even though the costs for building a decent hardware infrastructure for a company are decreasing, an SME still can't afford enterprise software solutions, (which uses those resources), because they are expensive. Even a more negative fact is that usually SMEs use only a small percent of the capability of an enterprise software solution. A research from the National Restaurant Association shows the main barriers to adding technology in restaurants. 63% of the interviewers have mentioned that the cost of implementation is the main barrier, 50% of them mentioned lack of infrastructure, 49% service and repair, 49% per transaction/usage costs, 48% customer acceptance and 44% staff training [8]. That's why most SMEs do not want to invest in large infrastructure and big software solution that has to be maintained and upgraded with new features by a development team and train the staff to use the new software [9].

Our model enables monitoring and improving the efficiency of business processes of all three types of business processes in the hospitality sector; management, core and support business processes [10] [11]. The target group that will use the pro, cess management system is all companies, but with main focus on SMEs and startup companies, so the best approach is to develop the system as SaaS.

A. Architecture

The SaaS model must be built on top of infrastructure that allows high-concurrency, high volume of data, and responds to the challenges of the system performance, security, elasticity and scalability. In other words, the solution will be built on top of a platform that possesses the SaaS features. The system will be used by multiple companies, so the model of the system must accommodate all of them. The solution to this problem is a single-instance multiuser model. That model allows configuration of the user interface and feature as well on company of user level. The system lives only as one instance for all the companies. If anything is changed or upgraded in the application, it will be available to all users. The relevant data to each company must be isolated

and secured, because the users are concurrent to each other. Sharing the same infrastructure and the improved maintenance gives benefits regarding the costs to the SMEs [9].

We propose a model that consists of three main components that will communicate between each other, a server, mobile application and web application. The server is responsible for serving the mobile and web application. It is in charge for communicating with databases, creating, reading from it, writing and updating it and securing them from any attacks. The server also has the web server that serves the caterers that will access the web application from their web browser. It is hosted in cloud. That means that it will have the key benefits of cloud, like flexibility and scalability, accessing extra resources as and when required, cost-effectiveness, reliability. They are more reliable than traditional servers, due to the number of available servers, if there are problems with some, the resource will be shifted so that clients are unaffected.

The mobile application is of crucial importance for the system. It facilitates and improves the customer experience. Clients have to download only one application (iOS or Android), which can be used in every caterer that has implemented the system. They can see the services that the caterer is offering, order or book them depending on the service, paying if needed etc. Caterers also can make various targeted campaigns that will be executed in the mobile platform using a web application. They can place image or video ads that will be shown to certain customers depending on the filters added by them.

The model includes a web application also. The web application is intended for satisfying the caterers' need. From the web application they can set up the configuration, set up their profile, creating their services and other configuration information. The caterer is ready to receive orders or bookings from clients that have the mobile application. The orders pop up as a web browser notification in real time. This is very important part because it means that the caterers do not have to change their current way of working with their current systems if they don't want to. So they can add the feature of receiving orders for services from clients without having to change the current way of functioning. The AI assistant is placed on the web application. It can be used to improve their current business processes and deliver a superior service to the customers.

B. Features

Having technology and especially a self-service mechanism in the hospitality sector is proven to have lots of benefits. We will start with the first and maybe the most important one, that introducing ICT and self-service mechanisms significantly makes a reduction in costs and undoubtedly increases the sales.

When self-service is in role in the hospitality sector with terminals on the rooms or tables the stats shows 30% reduction in cost and 10-15% increase in sales. Even better results are shown when the customers are left to their own devices spend up to 75% more[12]. More than half of the world's population now uses a smartphone [13]. With that will grow the consumer expectations for mobile self-service in every industry. They have adopted an intuitive digital mindset that is only going to grow over time.

In order to improve efficiency, caterers have two options. The first one is to reduce the inputs, i.e. costs that they use in the process of production of services, and the second option is to increase outputs, i.e. sales, which refers to revenues and occupancy rate of rooms and tables. The ideal solution for increasing revenues is raise prices while offering guests some added value with regard to the caterers main competitors at the same time [14]. As we mentioned that introducing self-service mechanism improves both inputs (reduces costs) and increases sales, we conclude also that introducing self-service improves the efficiency of the caterers.

Mindberry reported that 77% of people use their mobile phones in restaurants [15]. It would be better to use them for looking what kind of services the host is offering, rather than doing something else. Another statistic shows that 66 percent of diners prefer restaurants that have a rewards scheme (eMarketer) and 80% said it is important to see a menu before they dine at a restaurant (Everything Mobile) [16]. This being said means that the businesses will have reduction in cost and significant increase in sales, much more satisfied customers and taking advantage of contemporary addictions.

A very important feature is the possibility of processing the data. Even though nowadays everyone is collecting data, processing them is very difficult for SMEs. Information technology has changed the way companies operate and is affecting the entire process by which companies create their products. Furthermore, it is reshaping the product itself: the entire package of physical goods, services, and information companies provide to create value for their buyers [17]. It is the data processing that makes the difference, mainly because of the following features [18]:

- *Better results and increased productivity*. Data can be processed in different forms to obtain the required information, without data it will be impossible to take a good decision. For example: the software gives the exact number of man or woman clients, or it shows how frequented their business is by young, more mature or old people. Having this kind of knowledge, they can make much more targeted campaigns to improve the stats where needed, like making discounts for women's day. This information can help them make other decisions as well, like what kind of music they should play.
- *Gain in speed, accuracy and reliability*. It insures that data that is collected from the work done is collected pretty fast and accurately. When a data is collected and figured through computers, there are no or negligible chance of errors. Accurate input means accurate output.
- *Reduced cost*. The collected data is a valuable asset for every need and having it stored provides easy access to when required. This eliminates the need repeat the process of collecting them. Moreover it is much moreeasy and convenient to manipulate with data that are stored in

digital form. Sending or transferring the data is also much easier. This directly helps in cost reduction.

Having the possibility of getting the data processed by such software is indeed a great advantage for the whole sector, making improvements by economical, ecological and social aspects.

A step further is also made by using an AI feature with current techniques to mark important notes for the businesses itself. It mainly uses technique known as rule-based expert system with clear rules and outcomes and by the data already processed as an input. The expert system will handle these kind of problems, for example, if the caterer is mostly visited by young people, it will suggests what kind of music to play, or if their business is visited far more often by men than women, it will suggest to work on this direction, by leading maybe a women's campaigns or advertise their business to women users. Another example would be the system will analyze the average time of service from receiving the service to purchasing the service every day (day or night) / week (weekdays or weekend) / year (summer or winter), and Where is that average time over a certain critical limit, so it will notify and suggest the business to intervene. This features automates calculations and logical processes.

Such an example of use is the following: if the system realizes that caterer is not visited enough by women, than it suggests to do a campaign with special offers for them in order to attract them, or it will suggest to do advertisements on the mobile application which will be shown mostly to women.

Another example is if by some part of the day or week, the caterer is visited and their services are used in some significant number by young people than the system suggests a playlist of music songs to play. If some part of day, week or year is very crowd and if that leads to increase significantly the average time from the time of order to the time of serving the service or food than this needs action in order to improve this time. It may suggest to increase the number of stuff in that particular part of the year, month, week or day. These are just few examples to have a feeling of what can the system do. A lot more rules and outcomes can be defined.

This model presented here is directly changing one of core business processes in the hospitality industry. That is the process of ordering and serving meals and beverages and booking other services that are part of this industry. Fig. 1 presents the process of ordering on restaurants / coffee bars / night clubs, it has quite some difficulties and a lot of time is wasted for placing the order. For ordering your meal, you have to pass few steps and a lot of time is spend only for purchasing it. A worst case scenario that quite often shows up in every restaurant / coffee bar / night club which doesn't offers self-service is as follows: Some of the clients calls the waiter by raising their hands and they wait him to come or others choose just to wait for the waiter until he comes on his own. The client than asks for the menu if he hasn't bring them with himself. Here we add another time for waiting until the waiter brings up the menu. Leaving up the menu, the clients are able now to take a look of the services that the host is

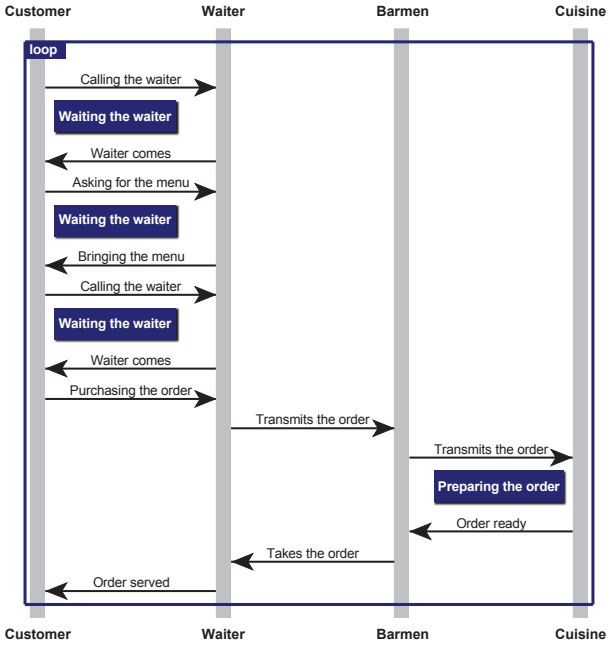


Fig. 1. The process of serving meals and beverages in restaurants without using technology.

offering. After they make a decision there is again a repeating procedure, calling and waiting for the waiter to purchase the order. Finally after this long procedure you can place your order. Introducing self-service reduces the wasting time for making the order, and by that it makes room for making more orders, i.e. increasing the sales. Another advantage is that it mitigates the risks of human mistakes. As shown on Fig. 2, the model presented here simplifies this process. It skips the steps of calling and waiting the waiters multiple times. Immediately after getting in the caterer the clients can see the offers on their mobile and purchase their order. By simplifying the ordering process the clients have plenty more time to purchase more orders making more sales for caterers.

Almost the same problem is occurring in hotels that offers some kinds of services like all kind of sport courts, or Jacuzzis, massages or room food service. The current procedure that comes in play in hotels, shown in Fig. 3 is as follows: for reserving a service the client has to make a call or walk down to the reception and ask them for availability of the requested service. Than they check the availability and show that to the client. We have to add here also a time that is wasted during the check procedure. And if the client decides to book the service, the booking again needs to go through the receptionist, including the waiting time to do the process and sending him acknowledgement that the booking is done successfully. It is worth mentioning that a receptionist can serve only one client at the same time, which is not the case with digitalized processes. This business process can be significantly improved if hotels implement the SaaS presented here. It saves time,

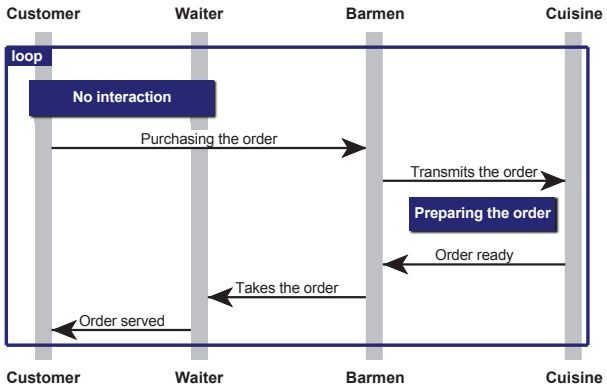


Fig. 2. Improved process of serving meals and beverages in restaurants using an SaaS solution defined in our model.

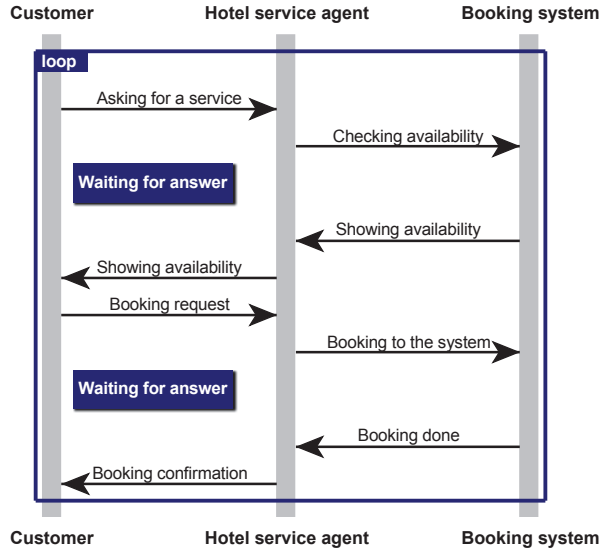


Fig. 3. The process of booking services without using our SaaS.

resources, mitigates human errors in bookings (that happens a lot) and delivers superior service. As shown in Fig. 4 the clients can book a service or make an order by the comfort of their room, simply by opening the mobile application, viewing the availability and other features and making a decision in any time.

IV. Discussion

This section discusses the advantages and disadvantages of the proposed system.

A. Benefits

There are several important advantages why a SaaS e-Caterer solution has to be adopted by the caterers emerged. One of the most important reasons is that the cost for using the service is very low. Along with that, the risk of failure

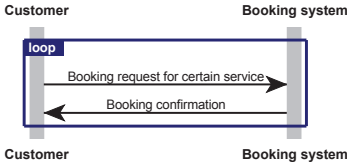


Fig. 4. The process of booking services using the SaaS proposed here.

of the software is very low, as it can easily be replaced or removed.

The vendor of the SaaS solution has the responsibility of the infrastructure, fixes and upgrades of the software. Most of them guarantee high percentage availability of their cloud services. The data resides in a secure environment and is backed up regularly. The solution is accessible by any device that has internet access 24 hours a day. The SaaS solutions are usually updated with new features more frequently than the standalone solutions. As software for caterers, SaaS solution is more stable solution. The ease that SaaS offers to the users is not achievable from the large software suites.

The system offers data mining and data processing. Nowadays these are very useful information, which can be used in improving the business processes. Specialized campaigns can be built to suit the business budget and goals. The added possibility of having advices to improve the business processes of the providers by the AI implemented in the system is the first of a kind. It also means that the companies are far more informed, better process monitoring, better and more efficient running of business processes. Accurate input means accurate output.

As logic suggests and as mentioned above, researches have shown that with technology added to the caterer, it leads to more orders from users during their stay. That means more workloads and more profit. More workload can lead to increase the number of staff also. It has been shown that a significant reduction in costs is done by using this model.

B. Challenges

Apart of many advantages, this model has several disadvantages that should be handled. One of the biggest challenges is to handle the possibility of false ordering because of static QR Codes in every table. One of solutions could be to make the payment obligatory during the process of ordering. And that is very elegant solution, but as the system should be used by wide range of random people, we could face another problem. That is, as mentioned in Section II, the low percentage of using the mobile payment feature by now. This could dramatically decrease the usage of the mobile application.

Another approach is to let the payment be optional and add some security layers to the mobile application like location or NFC tags (which would increase the cost of implementation). The waiters should make sure that the order was actually made by customers sitting in the table before processing it.

Internet access is required all the working time. This could be one of the downsides of the system, especially for some parts of the world. Also the employers with no technological experience and education could be a barrier for implementing this solution.

C. Risks

In order features to be fully useful it is crucial that the software is used by both side, clients and caterers. The risk here is if caterers implement this solution and clients by some reason don't. There will be no data and a lot of the features for caterers will not be useful. The risk of other case when caterers don't implement this solution is quite obvious.

V. Conclusion and future work

The number of cloud SaaS providers is increasing rapidly, together with variety of multi-tenant cloud services and applications. SMEs should consider the offers seriously, especially e-Business SaaS solutions. This paper helps the SMEs to better manage the risks and benefits of implementing a SaaS e-Business solution. They can better solve the problems and challenges. In this paper, we have analyzed problems in realization of business processes in hospitality sector and propose a new model which overcomes them and improves the efficiency of work.

Our model introduces AI improves business processes of the provider and customer experience. Efficiency of business processes are also significantly improved.

Future work needs to be done on marking notes for businesses, exploring the possibilities of other AI techniques, like supervised and unsupervised learning, so that it will not depend on humans rules and outcomes.

References

[1] W. Jin-zhao and W. Jing, "Issues, challenges, and trends, that facing hospitality industry," *Management science and engineering*, vol. 3, no. 4, p. 53, 2009.

[2] D. Buhalis and H. Main, "Information technology in peripheral small and medium hospitality enterprises: strategic analysis and critical factors," *International Journal of contemporary hospitality management*, vol. 10, no. 5, pp. 198–202, 1998.

[3] R. J. Kwortnik Jr, "Clarifying fuzzy hospitality-management problems with depth interviews and qualitative analysis," *Cornell Hotel and Restaurant Administration Quarterly*, vol. 44, no. 2, pp. 117–129, 2003.

[4] M. Young, "The technical writers handbook. mill valley, ca: University science, 1989," *Checklist before starting the analysis*, vol. 2.

[5] G. Ltd, "Get waiter software." https://www.getwaiter.com.

[6] E. I. Capital, "Qikserve," 2012.

[7] A. Brown, S. Dodini, A. Gonzalez, E. Merry, and L. Thomas, "Consumers and mobile financial services 2015," *Board Governors Federal Reserve Syst., Washington, DC, USA, Tech. Rep*, 2015.

[8] N. R. Association *et al.*, *Mapping the restaurant technology landscape*. National Restaurant Association, 2016.

[9] K. Kolic, S. Ristov, and M. Gusev, "A model of saas e-business solution," in *Telecommunications Forum Telfor (TELFOR), 2014 22nd*, pp. 1146–1149, IEEE, 2014.

[10] B. Krstic, E. Kahrovic, and T. Stanisic, "Business process management in hotel industry: A proposed framework for operating processes 4," *Ekonomika*, vol. 61, no. 4, p. 21, 2015.

[11] M. Drljača, "Methodology of business process development in a hotel," in *18th Biennial International Congress Tourism & Hospitality Industry 2006*, 2006.

[12] Qikserve, "So, Mobile!," January 2013.

[13] S. Kemp, "Digital in 2017: Global overview," *We are social*, 2017.

[14] K. Poldrugovac, M. Tekavcic, and S. Jankovic, "Efficiency in the hotel industry: an empirical examination of the most influential factors," *Economic research-Ekonomska istraživanja*, vol. 29, no. 1, pp. 583–597, 2016.

[15] M. C. GmbH, "Smartphone users in Germany, Austria and Switzerland," 2015.

[16] Google, Nielsen company, "Mobile search moments, understanding how mobile drives conversions." URL: https://www.thinkwithgoogle.com/consumer-insights/creating-moments-that-matter/, March 2013.

[17] M. E. Porter, V. E. Millar, *et al.*, "How information gives you competitive advantage," 1985.

[18] J. Alex, "Importance of data processing." URL: https://planningtank.com/computer-applications/importance-of-data-processing, August 2017.

Digitalization of Banking services as a Driving Force towards Profitability – IT Perspective of the Macedonian Banking Sector

Aleksandra Karadja
School of Business Economics and Management
University American College Skopje
Skopje, Republic of Macedonia
aleksandra.karadza@uacs.edu.mk

Veno Pachovski
School of Computer Science and Information Technology
University American College Skopje
Skopje, Republic of Macedonia
pachovski@uacs.edu.mk

Marjan Bojadjiev
School of Business Economics and Management
University American College Skopje
Skopje, Republic of Macedonia
provost@uacs.edu.mk

Marija Andonova
School of Business Economics and Management
University American College Skopje
Skopje, Republic of Macedonia
nacova@uacs.edu.mk

***Abstract* — In the last decade, with the increasing advancement of information technology, banks are introducing innovation in their products and services, by continuously upgrading their technologies, systems and processes. Their investment in the IT segment is aimed for providing innovative products and services at the market, coping with the technologically improving external environment, abiding by the latest rules, policies, and regulations, increasing their customer base and achieving higher profits. The purpose of this paper is to measure the effect of one of the aspects of digitalization of banking services, i.e. investment in information technology, on the profitability of banks which constitute the Macedonian banking sector – represented by ROE and ROA. Furthermore, a comparative analysis is performed between the performance of banks constituting the small, medium and large banking groups, in terms of their investment in IT and its impact on the profitability performance. This paper could be beneficial both scientifically, in terms of initiating research in the area of digitalization of banking services in the country, the Balkans and in Europe, as well as for providing recommendations to the banks' management in terms of improving the constituted strategies and becoming more digitalized.**

Keywords — digitalization, banking, ROA, ROE

I. INTRODUCTION

In the 21st century, the globalization has exerted a tremendous influence on the banking business operations, since traditional branches of commercial banks have been transformed into places which implement the latest trends in digitalization of banking products and services. With the development of e-banking, e-commerce and the technologically advanced environment, banking institutions are gradually introducing alternative channels. A higher portion of their products and services are accessible online, as well as through mobile phones, tablets, Automated Teller Machines (hereinafter: ATMs) and other electronic devices used by their customers. Therefore, these financial institutions are offering flexibility in terms of usage of banking services 24 hours a day, 7 days per week. The implications of digitalization are various, and among others, they include achieving higher customer satisfaction, higher profitability figures and possessing higher share of the financial services market. Despite the benefits of the digitalized era, banks operating under the principle of *traditional banking* are still unaware of the benefits of implementing an intensive digitalization of their products and services, and transforming their branches into 'banks of the future'. Therefore, banks' annual and financial reports rarely include data regarding profits achieved by offering digitalized products and services in the portfolio.

The main aim of this paper is to analyze whether and to what extent one of the main factors of digitalization, i.e. investment in information technology (hereinafter: IT) affects the profitability figures of small-sized, medium-sized and large-sized banks in Macedonia throughout the period 2012 to 2015. In order to investigate whether and to what extent the IT investment influences the profitability indicators, i.e. Return on Equity (ROE) and Return on Assets (ROA), multiple regression models will be computed for each of the banking groups.

II. INVESTMENT IN INFORMATION TECHNOLOGY AND BANK PROFITABILITY

A. Positive Relationship between Investment in Information Technology and Bank Profitability

The literature suggests that there is scientific evidence indicating a positive relationship between IT investment as one of the aspects of digitalization and bank profitability. According to Ho & Mallick (2006), there are two positive effects stemming from the relationship between IT and banks' performance. Firstly, through the introduction of IT, banks are experiencing a reduction in their operational costs, an example of which would be using online banking for conducting

transactions. Secondly, the implementation of IT provides the possibility for doing transactions within the existing bank network, as usage of ATMs on dispersed locations in a country [1]. The study of Chowdhury (2003) analyzed the impact of IT on the profitability performance of a sample composed of 327 banks in Asian and Pacific Basin countries (Australia, Hong Kong, Japan, Malaysia, New Zealand, South Korea, Taiwan and Thailand). The results from the conducted profitability-based analysis indicate that the IT capital has a positive and significant effect on the profitability variables of the bank observations being analyzed – ROA and ROE [2]. Furthermore, by implementing the Stochastic Frontier Approach (hereinafter: SFA), Romdhane (2013) conducted an empirical analysis of the IT effect on the performance of a bank sample of 15 Tunisian banks during the period 1998 to 2009. The results generated designate that the IT variable positively and significantly affects the cost efficiency of banks [3].

When comparing the impact of increasing IT investments on the profitability of high IT level banks, in their research study, Leckson-Leckey, Osei and Harvey (2011) found the following. Banks which made higher investments in IT experienced higher profitability figures in comparison with low IT level banks, which proves for a positive and statistically significant relationship between IT and ROA of high IT level banks. Furthermore, when estimating the effect of IT expenditures on the profitability of all banks represented by ROE, the same results were produced. The relationship between IT investments and ROE of high IT level banks was positive, but not statistically significant. These findings imply that IT investments exert a greater influence on the profitability of high IT level banks, compared to the effect on profitability of low IT level banks [4]. By analyzing the impact of IT on the bank performance of selected 11 banks in Nigeria, during the period 2001 to 2011, the following results have been generated (Muhammad, Gatawa and Kebbi, 2011). The authors have conducted a regression analysis with ROE as a dependent variable and Net Profit, ATM and e-banking services in Nigeria as independent variables. Their findings suggest that there is a strong, positive and statistically significant relationship established between IT and ROE [5].

B. Negative Relationship between Investment in Information Technology and Bank Profitability

Despite the scientific evidence showing a positive relationship between IT investment and bank profitability, there are numerous research studies indicating that there is a negative association between these two variables.

A mid-nineties research conducted in the United States of America (hereinafter: USA) by Prasad and Harker (1997) suggests that banks' investment in IT does not influence their profitability expressed in terms of ROA and ROE, and that it also does not represent a barrier for entry in the retail banking industry [6]. These findings have been confirmed in the study by Markus and Soh, 1993, cited in Beccalli, 2007, who argue that there is no significant relationship established between investment in IT in small-sized banks and their profitability performance. Regarding the performance of banks which belonged to the large banking group, the authors suggest that these institutions experienced negative returns as a result of their permanent spending on IT operations [7]. Furthermore, Mallick and Ho (2008) were also examining whether investment in IT influences the financial performance of a panel set composed of 68 banks in the USA during the period of analysis ranging from 1986 to 2005. Their research findings indicate that during the observed period, banks which report higher figures for IT also show lower profitability, which implies that IT has a negative effect on the profitability of banks in the USA [8].

Farouk and Dandago (2015) were analyzing the relationship between investment in IT and banks' profitability in Nigeria, as measured by ROE, ROA, Net Profit Margin (hereinafter: NPM) and Earnings per Share (hereinafter: EPS). Their research suggests that there is an existence of a significant relationship between the IT performance and ROA. They further elaborate that the banks' investment in IT does not have a positive contribution to the profitability performance, since a 1% increase brought for a decrease in this financial ratio of 1.587%. The same results were also generated for the impact of IT investment on the financial performance of Nigerian banks, measured by ROE, NPM and EPS [9]. In addition, the study of Beccalli (2007) examined the relationship between IT investment (hardware, software and other IT services) and financial profitability measures of 737 commercial banks in the European Union located in France, Germany, Italy, Spain and in the United Kingdom (hereinafter: UK), during the period ranging from 1993 to 2000. The empirical results of the study have indicated that on short-term, the correlation between investment in IT and ROA ratios is both – negative and statistically significant. When calculating the correlation between IT investment and business performance (ROA and ROE) for each country which was part of the sample, he learned that there is a negative and statistically significant relationship for four of the countries, apart from Germany [10]. Further research (Günsel and Tükel, 2011) analyzed the relationship between IT capability and bank profitability of a research sample composed of 15 banking institutions in Turkey. The relationship between these two variables had been tested by using several methods, among which were the correlation and regression analyses. The regression model analyzing the relationship between IT capability and ROE, suggests there is no statistically significant relationship between the two variables. The same results were generated by the regression analysis conducted between IT capability and ROA [11].

III. DATA SAMPLE AND DESCRIPTIVE STATISTICS

A. Quantitatve Data Analysis

The bank-specific data which we use in this paper were obtained from the annual and financial reports of Macedonian small-sized, middle-sized and large-sized banks throughout the period 2010 to 2015. These figures have been analyzed by using Excel spreadsheets and the statistical software Statistical Package for the Social Sciences (hereinafter: SPSS). Multiple regression analyses were performed for the purpose of

indicating whether, and to what extent the aspects of digitalization (IT and Software and Marketing and Administrative expenses) influence the profitability indicators (ROA and ROE) of Macedonian small-sized, middle-sized and large-sized banks.

Due to existence of a small banking sample in Macedonia, two multiple linear regression analyses will be performed. The first one will be used to determine the nature of the relationship between ROE as a dependent variable and the investment in IT and Software and the Marketing and Administrative expenses of the Macedonian banking sector as independent variables. In addition, the second multiple linear regression analysis will be used to determine the nature of the relationship between ROA as a dependent variable and the figures representing the factors of digitalization, used as independent variables.

B. Descriptive Statistics of the Sample

TABLE I. AVERAGE VALUES OF SMALL, MEDIUM AND LARGE BANKING GROUPS IN MACEDONIA DURING THE PERIOD 2010 TO 2015

Values	Banking groups		
	Small banking group	***Medium banking group***	***Large banking group***
ROE	-0.0620	0.0235	0.0872
ROA	-0.0108	0.0024	0.0161
IT and Software expenses (in MKD thousand)	40.708	22.021	39.077
Marketing and Administrative expenses (in MKD thousand)	10.222	15.406	62.620

[a] Developed by the author, based on calculations obtained from publicly available information from annual reports of Macedonian small-sized, middle-sized and large-sized banks for 2012, 2013, 2014 and 2015

Table I. shows a comparison of average values for the Macedonian banking sector used in the paper. Regarding the ROE figures, the Table indicates that the management of the large banking group was the most successful in paying higher dividends to its shareholders and supporting future growth. Its ROE of 0.0872 suggests that per 1 Macedonian Denar (hereinafter: MKD) investment in equity, banking groups' shareholders managed to earn an average return of 8.72%, applicative for the period from 2012 to 2015. In addition, the medium banking group reported ROE at the value of 0.0235, implying that during the four fiscal years, by investing MKD 1 in average equity, medium-sized banks' shareholders earned an average return of 2.35%. Lowest profitability expressed as ROE was achieved by banks classified as small-sized, due to their negative average return on the equity invested on part of the shareholders, amounting to -0.0620. These figures show that throughout the period 2012 to 2015, a MKD 1 investment in equity, has yielded a negative average return of 6.2%. Therefore, on the basis of these average values, it was concluded that the shareholders of large Macedonian banks managed to earn the highest return on the equity they had invested (Karadja, 2017) [12].

The average ROA of the large banking group was the highest in comparison with the figures of the small-sized and middle-sized banks, at the value of 0.0161. Therefore, by investing 1 MKD in their total assets, during the four-year period of analysis, banking institutions in Macedonia classified as large-sized banks, managed to make an average return of 1.61%. On the other hand, by investing MKD 1 in total assets, the middle banking group reported an average return of 0.24%, whereas the small banking group achieved a negative return on its investment at the amount of – 1.8%.

Astonishing results are shown for the average IT and Software expenses for the small banking group, which were the highest in comparison with the medium-sized and large-sized banking institutions in Macedonia. The average IT investment incurred by the group composed of small-sized banks has amounted to 40.708 thousand MKD. These figures suggest that on average, these banks invested a significant portion of their budget on IT infrastructure, software and systems as well as necessary licenses for conducting their operations. When analyzing the IT figures of the large banking group, it can be concluded that during the period from 2012 to 2015, large-sized banks in Macedonia made an average investment of nearly MKD 40.000 thousand. On the other hand, the IT and Software expenses for the middle banking group were the lowest, in comparison with their competitors, implying that medium-sized banks in the country did not spend a lot of their funds on a digital transformation of their services.

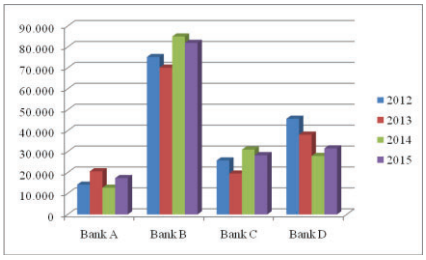


Fig. 1. IT and Software expenses in the group of large-sized banks in Macedonia during the period 2012-2015

On the basis of Figure 1, one can conclude the following. In comparison with its competitors, Bank B incurred the highest expenses. This financial institution's IT investment experienced a minor decline from 2012 to 2013, followed by a tremendous increase in 2014 and a minor diminishment in 2015. Furthermore, in comparison with the other large-sized banks, Bank D incurred the second-highest IT and Software investment in 2012 and 2013, whereas in the two consecutive years, it experienced a decline in this type of expense on its Income Statement. On the other hand, it is evident from Figure 1 that Bank C reported the highest IT and Software costs in 2014, which was preceded by several fluctuations. In comparison with its competitors, Bank A reported the lowest expenditures in the banking group during the period 2013 to 2015. In the year of 2013, the IT and Software expenses of this banking institutions were higher in comparison with those of Bank C.

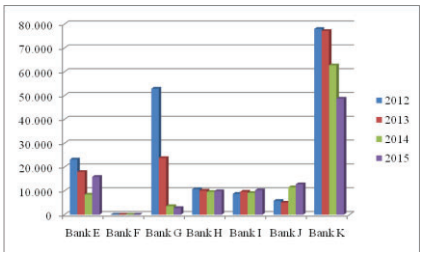


Fig. 2. IT and Software expenses in the group of middle-sized banks in Macedonia during the period 2012-2015

On the basis of Figure 2, what can be concluded is that Bank K reported the highest IT and Software expenses in the group of middle-sized banks during the period 2012 to 2015. Bank G incurred the second-highest IT investment expenditures on its Income Statement in the first two fiscal years, in comparison with its competitors. The lowest figures were reported in the last fiscal year. When analyzing the performance of Bank E, one can advocate that in comparison with its medium-sized bank competitors, this financial institution reported a decline in this type of operating expenses from 2012 to 2013. Moreover, these figures decreased by a tremendous amount in 2014, followed by an increase in IT and Software investment in 2015.

On the other hand, throughout 2012 to 2013, Bank J reported the lowest IT and Software figures within the group. In the following fiscal years, it experienced an increasing trend in terms of its investment. Bank I experienced various fluctuations in its operating expenses throughout the years, starting from an increase in IT and Software expenses from 2012 to 2013, followed by a decline in 2014 and an increase in 2015. Furthermore, Bank H experienced a downward trend in its IT and Software expenditures, followed by a slight improvement in the last fiscal year. Bank F's figures cannot be compared to the other middle-sized banks, since they have not been disclosed in the annual or financial reports.

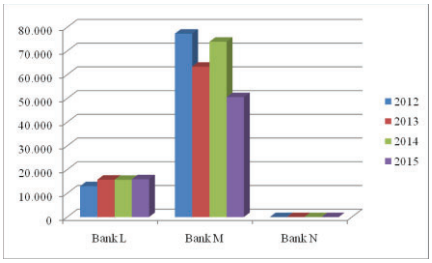


Fig. 3. IT and Software expenses in the group of small-sized banks in Macedonia during the period 2012-2015

When analyzing the performance of small-sized banks in Macedonia in terms of their IT and Software expenses during the period 2012-2015, on the basis of Figure 3, we can conclude that Bank M experienced fluctuations in its figures. They decreased by a small extent from 2012 to 2013, followed by a moderate increase in 2014 and a tremendous decline in 2015. On the other hand, Bank L did not experience various fluctuations in its IT and Software expenses during the four-year period of analysis. It incurred the highest IT and Software costs in 2013, which declined by a small extent in 2014. In the last fiscal year, the Bank managed to increase its investment by a considerable degree. The data of Bank N cannot be analyzed and compared to the competitors, due to their unavailability in the annual reports.

IV. METHODOLOGY OF THE RESEARCH

The Macedonian banking sector is comprised of fifteen banks, i.e. one state owned and fourteen privately owned banks, classified in three groups on the basis of their assets size: large-sized, medium-sized and small-sized banks. Data used in this paper have been obtained from banks' Balance Sheet and Income Statements, as publicly disclosed information. The data which have been obtained include the ROE, ROA, investment in Marketing (Marketing and Administrative expenses) and investment in IT (IT and Software expenses) of 14 bank observations in Macedonia during the period 2012-2015.

V. DATA ANALYSIS AND FINDINGS

A. Multiple Linear Regression Model with ROE as a Dependent Variable

Since a dependent variable is a continuous variable, first, we apply a multiple regression model with two independent variables. We would like to examine the relationship between investments in IT and ROE as well as the relationship between Marketing and Administrative expenses and ROE of the Macedonian banking sector. The linear regression model is defined with the following equation:

$$\Pi_i = \beta_0 + \sum \alpha B_j + \varepsilon_i \quad (1)$$

where Π_i is the dependent variable represented by an observation of ROE; the independent variable includes the intercept β_0; the independent variables B_j; the α represents the coefficients, and ε_i is the error.

Hence, we can derive the following regression model:

$$ROE = \beta_0 + \beta_1 \ln IT + \beta_2 \ln Marketing + \varepsilon_i \quad (2)$$

where:
ROE = ROE of the Macedonian banking sector;
ln *IT* = natural logarithm of investment in IT and Software
ln *Marketing* = natural logarithm of Marketing and Administrative expenses;
β_0, β_1, β_2 = independent variable coefficients, which define the amount of change in y per every change in predictor variable;
ε_i = random error in ROE, assumed to be well-behaved.

In Table II., we find that the adjusted R square (R^2) of our model amounts to 0.144, whereas the R square equals to 0.181. Therefore, the coefficient of multiple determination (R^2 = 0.181) indicates that 18.1% of the variance in the ROE of the Macedonian banking sector is explained by the variation in the investment in IT and the variation in the marketing and administrative expenses. The reason for the small amount of R square can be attributed to the unavailability of data in some of Macedonian banks' annual reports for the fiscal period from 2012 to 2015, as well as the small number of banks constituting the sample size.

TABLE II. MODEL SUMMARY

Model	*R*	*R Square*	*Adjusted R Square*	*Std. Error of the Estimate*
1	,426[a]	,181	,144	,1032021

[a] Predictors: (Constant), lnmarketing, lnit

Furthermore, Table III. shows the results of the overall F-test in order to determine whether there is a significant relationship between the dependent variable, i.e. ROE and the entire set of independent variables, represented by natural logarithm of IT investment and natural logarithm of

investment in marketing. One can draw a conclusion that the regression model is significant, since there is a significant relationship established between the independent variable in our model: natural logarithm of investment in IT; natural logarithm of Marketing and Administrative expenses and the ROE of the Macedonian banking sector as a whole.

TABLE III. ANOVA[A]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	,104	2	,052	4,870	,012[b]
	Residual	,469	44	,011		
	Total	,572	46			

a. Dependent Variable: ROE
b. Predictors: (Constant), lnmarketing, lnit

Since the p-value of 0.012 is lower than 0.05, we can conclude that all variables are jointly significant.

Table IV. depicts the regression coefficients, the intercept, the significance of all coefficients as well as the intercept in the model.

We find that our regression analysis estimates the linear regression function to be the following:

$$ROE = -0.267 - 0.021 \text{ x ln it} + 0.052 \text{ x ln marketing} \quad (3)$$

In our first multiple linear regression model, depicted in Table IV., the coefficient which is significant is Marketing and Administrative expenses, represented by *lnmarketing*, which is shown from its p-value, which is smaller than the significance level of p equal to 0.05. Moreover, the coefficient lnit does not represent a statistically significant indicator of profitability of the Macedonian banking sector, since the value of 0.229 is higher than the p-value of 0.005.

TABLE IV. COEFFICIENTS

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-,267)	,215		-1,242)	,221
	lnit	-,021)	,017	-,170)	-1,221)	,229
	lnmarketing	,052	,017	,426	3,059	,004

a. Dependent Variable: ROE

B. Multiple Linear Regression Model with ROA as a Dependent Variable

Since we would also like to assess the relationship between IT and Software and ROA as well as the relationship between Marketing and Administrative expenses and ROA of the Macedonian banking sector, we apply the second multiple linear regression. The second linear regression model is defined with the following equation:

$$\Pi_i = \beta_0 + \sum \alpha B_j + \varepsilon_i \quad (3)$$

where Π_i is the dependent variable represented by an observation of ROA; the independent variable includes the intercept β_0; the independent variables B_j; the α represents the coefficients and ε_i is the error.

Hence, we can derive the following regression model:

$$ROA = \beta_0 + \beta_1 \ln IT + \beta_2 \ln Marketing + \varepsilon_i \quad (4)$$

where:

ROA = ROA of the Macedonian banking sector;

ln *IT* = natural logarithm of investment in IT and Software

ln *Marketing* = natural logarithm of Marketing and Administrative expenses;

$\beta_0, \beta_1, \beta_2$ = independent variable coefficients, which define the amount of change in y per every change in predictor variable;

ε_i = random error in ROA, assumed to be well-behaved.

TABLE V. MODEL SUMMARY

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,352[a]	,124	,084	,0193131

a. Predictors: (Constant), lnmarketing, lnit

TABLE VI. ANOVA[A]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	,002	2	,001	3,111	,054[b]
	Residual	,016	44	,000		
	Total	,019	46			

a. Dependent Variable: ROA
b. Predictors: (Constant), lnmarketing, lnit

TABLE VII. COEFFICIENTS

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-,042)	,040		-1,036)	,306
	lnit	-,003)	,003	-,134)	-,927)	,359
	lnmarketing	,008	,003	,354	2,455	,018

a. Dependent Variable: ROA

On the basis of the conducted multiple linear regression model with ROA as an independent variable, we can conclude that the variables are not jointly significant. Therefore, a further elaboration of this regression model is not be provided.

VI. CONCLUSIONS AND RECOMMENDATIONS

A. Limitations of the Study

Due to lack of publicly available data, it is impossible to produce a comprehensive overview of the digitalization of the banking sector in Macedonia. Furthermore, some of the banks have not published their data in their annual reports. Those who have published the information have published only the figures for total investment in externally developed software, and not providing information per segment of digitalization (ATMs, SMS banking, Phone banking, etc.) and statistics per customer segment (retail and business).

B. Conclusions

This paper examined the digitalization of banking services and its impact on profitability indicators of the Macedonian banking sector, during the period 2012 to 2015. On the basis of the results of the multiple linear regression models, one can conclude that as an independent variable, the IT and Software

expenses are negatively related to the profitability of the Macedonian banking sector, represented by two independent variables – ROE and ROA.

C. Recommendations

The main goal of this paper is to initiate research in the area of digitalization of banking services in Macedonia, in the countries of the Balkans and in Europe, in general. Analyzing the digitalization of banking services and their impact on the profitability performance of banking institutions, contributes for a better formulation of strategies on part of the top management, which brings for profit maximization, better market positioning and attracting new customer base in the Macedonian banking sector. Therefore, further research on this topic, not just in Macedonia, but also in the Balkan countries and in Europe, will help the management of banking institutions to better structure their budget aimed for investments in IT infrastructure. Moreover, they will have the ability to improve the evaluation of the bank's financial performance based on the adoption and implementation of new and innovative banking services and their usage rate in part of the customer base. By spreading the research in the neighboring countries, banking institutions in Macedonia will be able to compare their performance and locate the areas of digitalization which need to be a subject of improvement, for the purpose of better transitioning from 'traditional banks' to 'banks of the future'.

D. Abbreviations and Acronyms

ATM	Automated Teller Machines
IT	Information technology
ROE	Return on Equity
ROA	Return on Assets
SFA	Stochastic Frontier Analysis
USA	United States of America
NPM	Net Profit Margin
EPS	Earnings per Share
UK	United Kingdom
SPSS	Statistical Package for the Social Sciences
MKD	Macedonian Denar

REFERENCES

[1] S. J. Ho, and S. K. Mallick, "The impact of information tehnology on the banking industry: Theory and empirics", 2006.

[2] A. Chowdhury, "Information technology and productivity payoff in the banking industry: evidence from the emerging market" in Journal of International Development, vol. 15, issue 6, 2003, pp. 69 –708.

[3] S. B. Romdhane, "Impact of information technology on the performance of Tunisian banks: a Stochastic Frontier Analysis with panel data", in Asian Academy of Management Journal of Accounting and Finance (AAMJAF), vol. 9, no. 2, 2013, pp. 95–125.

[4] G. T. Y. Leckson-Leckey, K. A. Osey, and S. K. Harvey, "Investments in information technology (IT) and bank business performance in Ghana", in International Journal of Economics and Finance, vol. 3, no. 2, 2011, pp. 133–142.

[5] A. Muhammad, N. M. Gatawa, and H. S. B. Kebbi, "Impact of information and communication technology on bank performance: A study of selected commercial banks in Nigeria (2001 – 2011)", in European Scientific Journal, vol. 9, no. 7, 2013, pp. 213–238.

[6] B. Prasad, and T. P. Harker, "Examining the contribution of information technology towards productivity and profitability in U.S. retail banking", Wharton School Center for Financial Institutions, University of Pennsylvania, 1997.

[7] E. Becalli, "Does IT investment improve bank performance? Evidence from Europe", in Journal of Banking and Finance, vol. 31, no. 7, 2007, pp. 2205–2230.

[8] S. K. Mallick, and S. J. Ho, "On network competition and the Solow paradox: Evidence from US banks", in Manchester School Supplement, 2008, pp. 37–57.

[9] B. K. U. Farouk, and K. I. Dandago, "Impact of investment in information technology on financial performance of Nigerian banks: is there a productivity paradox?", in Journal of Internet Banking and Commerce, vol. 20, no. 1, 2015, pp. 1–22.

[10] E. Becalli, "Does IT investment improve bank performance? Evidence from Europe", in Journal of Banking and Finance, vol. 31, no. 7, 2007, pp. 2205–2230.

[11] A. Günsel and A. Tükel, "Does information technology capability improve bank performance? Evidence from Turkey", in International Journal of eBusiness and eGovernment studies, vol. 3, no. 1, 2013, pp. 41–49.

[12] A. Karadja, "Digitalization of banking services as a drivng force towards profitability – Comparative analysis of the Macedonian banking sector", MSc Thesis, October 2017.

[13] Alpha Bank AD Skopje, Annual reports for 2012, 2013, 2014, 2015 and 2016.

[14] Centralna Kooperativna Banka AD Skopje, Annual reports for 2012, 2013, 2014, 2015 and 2016.

[15] Eurostandard Banka AD Skopje, Annual reports for 2012, 2013, 2014, 2015 and 2016.

[16] HalkBank AD Skopje, Annual reports for 2011, 2012, 2013, 2014, 2015 and 2017.

[17] Kapital Banka AD Skopje, Annual reports for 2013, 2014, 2015 and 2016.

[18] Komercijalna Banka AD Skopje, Annual reports for 2011, 2012, 2013, 2014 and 2015.

[19] NLB Banka AD Skopje, Annual reports for 2011, 2012, 2013, 2014 and 2015.

[20] Ohridska Banka AD Ohrid, Annual reports for 2012, 2013, 2014, 2015 and 2016.

[21] ProCredit Bank Macedonia, Annual reports for 2011, 2012, 2013, 2014 and 2015.

[22] Sparkasse Bank Makedonija AD Skopje, Annual reports for 2011, 2012, 2013, 2014 and 2015.

[23] Stopanska Banka AD Skopje, Annual reports for 2012, 2013 and 2015.

[24] Stopanska Banka AD Skopje, Financial statements for 2012, 2013, 2014 and 2015.

[25] Stopanska Banka AD Bitola, Annual reports for 2012, 2013, 2014 and 2015.

[26] TTK Banka AD Skopje, Annual reports for 2012, 2013, 2014, 2015 and 2016.

[27] UNI Banka AD Skopje, Financial statements for 2012, 2013, 2014, 2015 and 2016.

facetC: Highly customizable and CRUD based facet browser for Semantic Web

Nasi Jofce
Faculty of Computer Science and Engineering
Skopje, Macedonia
jofce.nasi@gmail.com

Riste Stojanov
Faculty of Computer Science and Engineering
Skopje, Macedonia
riste.stojanov@finki.ukim.mk

Dimitar Trajanov
Faculty of Computer Science and Engineering
Skopje, Macedonia
dimitar.trajanov@finki.ukim.mk

***Abstract*—As the Linked Open Data (LOD) Cloud is continuously expanding with new interlinked datasets, its general public accessibility becomes a challenging task. The facet browsers are trying to bring closer the Linked Open Data to the regular users, enabling full text querying of the datasets and concepts traversal using hyperlink navigation. However, their generic user interface is not suitable for presentation and interaction with different concepts.**

In this paper we present the *facetC*, a customizable facet browser solution that visualizes Linked Data resources in a highly customizable manner. The power of *facetC* lies in its pluggable templating mechanism, where a user can configure and select a custom user interface for each resource type. Besides that, it provides a predefined, but customizable, CRUD UX[1], which can make it easier for the general public to contribute to the LOD data gathering. *FacetC* is visioned to serve as a marketplace, where the innovative developers can provide custom templates for the LOD resources that will be more suitable for interaction with the resources.

***Index Terms*—Facet browser, Dataset, LODC**

I. Introduction

Consuming Linked Open Data Cloud from the general public needs to be achieved through a special tool which actively interacts with different dataset endpoints, and provides a rich, easy to use interface.

A vast range of facet browsers have been developed for simplifying the LOD consummation process. These browsers provide rich functionality for exploring and visualizing interlinked datasets, and are often faced by challenges referring to offline data access, dataset size, variety of dataset consumers and tasks needed to be handled. However, all these facet browsers provide a generic user interface for resource presentation and hyperlink navigation that is not suitable for all kinds of concepts. And all these challenges are handled in different ways by different browsers, each providing its set of unique features. [2]

On the other hand, modification of the datasets and resources is a challenge of its own, and most of the facet browsers don't provide a solution.

Providing a rich user experience is of an essential importance when creating highly customizable facet browser that both solves the above mentioned challenge and provides the necessary means to customize the user interface that will actually interact with the datasets in a CRUD fashion. The facet browser we are presenting in this paper, called *facetC*, provides a predefined CRUD user experience, with the ability for further customization intended to be used by the general public. As it provides capability for different presentation per resource type, *facetC* can be used as a marketplace for developers trying to publish their components for different LOD resources.

[1]CRUD stands for Create, Retrieve, Update, Delete, while UX for User eXperience

II. Related Work

A facet browser known as */facet* was proposed addressing the ability to select and navigate through facets of resources of any type and to make selections based on properties of other semantically related types based on modern Web standards. It also offers capability for refining the default provided facet browser configuration. [7]

Facet browsers based on graphical query builders leverage the graphical representation of the resources to build dynamic queries that can be executed on numerous SPARQL endpoints, relieving the user from excessive SPARQL knowledge. The user is only prompted to express the type of data he wants to retrieve, and the browser itself takes care of query building and the endpoints.

SPARQL enthusiasts on the other hand can choose to build their own queries utilizing the functionality provided by numerous faceted browsers based on auto-complete query builder, in order to ease out the entire query building process. Several auto-complete browsers are have been proposed, and we can mention the browsers that ease out the process of remembering the resource URI. [1]

Other efforts have been made in providing rich facet browsers that leverage the capabilities of HTML5 for providing a better user experience while querying multiple datasets. A proposed browser offers the abilities to search and filter through data, store data in the browser's local storage for faster retrieval and overcoming the endpoint downtime difficulties, as well as providing a wide range of data export formats, such as JSON, RDF/XML, CSV, Turtle. [8]

It is of an essential importance mentioning Swoogle as a widely used tool for discovering and indexing data residing

TABLE I
COMPARING TOOLS FOR NAVIGATING SEMANTIC DATA

	/facet	Swoogle	Sgvizler	WYSIWYQ	facetC
Multiple Sources	x	x	x	x	x
Resource Navigation	x	x	x	x	x
Resource Modification					x
Resource Visualization			x	x	
Customizable Templates				x	x
Customizable SPARQL Visualization			x	x	x

in distributed datasets based on Semantic Web technologies. It also provides a vast range of useful functionalities for data analysis while querying different sources.[4]

Another facet browser that allows users to specify SPARQL queries directly into HTML elements, and visualize the results with interactive charts, treemaps and graphs, is Sgvizler - a JavaScript based wrapper for visualizing SPARQL results. [10]

A recently proposed novel approach has taken the faceted browsers to the next level. Based on the widely used WYSIWYG (What You See is What You Get) paradigm, a more advanced form of facet browser, known as WYSIWYQ(What You See is What You Query) allows users to interactively visualize, enrich or re-purpose a given SPARQL query as a browsing UI. It aims to provide a two-way binding between SPARQL queries and RDF-based faceted browsers. [9]

Table I gives a comparison highlighting the customizability features between the above mentioned tools for exploring semantic data and our proposed facetC.

III. FACETC BROWSER PLATFORM

The *facetC* browser platform consists of two main parts: the end user journey and the UX provider. The CRUD functionality is the main feature of the end user journey, which offers representation of the data contained in the dataset by utilizing the FOAF vocabulary [6], as well as interactivity with the data. The UX provider handles data display layout definition and SPARQL query definition and execution.

A. End user journey

In order to model the entire CRUD functionality in an easy-to-use interface, the *facetC* platform defines four main presentation contexts. The default context shown to the user when accessing the application contains a list of the data types included in the dataset.

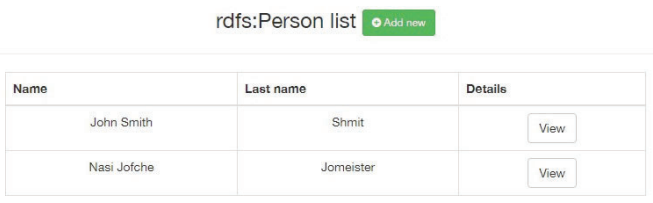


Fig. 1. List context for displaying a list of resources of the selected data type in a predefined layout

The user has the option to select one of the types, and navigate to the list-context, which loads the resources of the selected data type and displays the results in a defined layout, which can be customized from the UX provider.

The other properties of the selected resource can be shown in the details context, in a UX provider defined layout also. The template used in this layout may or may not display all the triples associated with the selected resource.

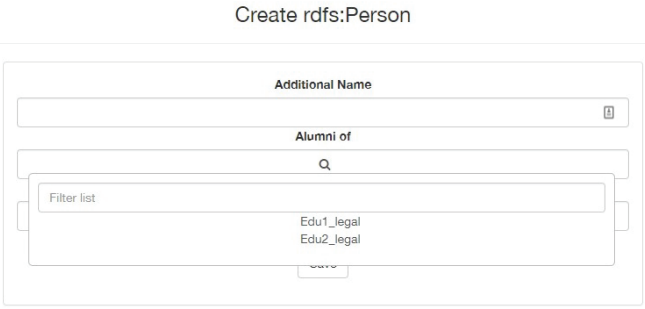


Fig. 2. Create/Edit context for defining/editing the resource's details

Besides showing the details of the resource, the user needs to create or edit a resource of the selected type. The create/edit context makes this functionality available, by providing a dynamically built form layout, which includes several sub-contexts regarding the type of data, whether it would be a text or date literal, or even a resource. In case the required data type is a resource, the select sub-context provides a select picker with search capability for resources of the specified data type. The results displayed in the select box are based on a UX provider defined template also. UX provider defined layouts for the above mentioned contexts are displayed in Figure 1 and Figure 2.

B. UX provider

This is the core of the entire application, as it provides functionality for defining layouts and retrieved data binding for the above specified contexts.

This can be described as a process containing more than one step, where the user can define specific data types based on preferred vocabulary. Options for updating and deleting data types are also available.

The next step is defining a configuration for each data type property that is going to be used in the layout. The configuration includes defining labels for the properties associated with a specific data type, as well as the object's type, whether it would be literal or another resource. As an object may have a

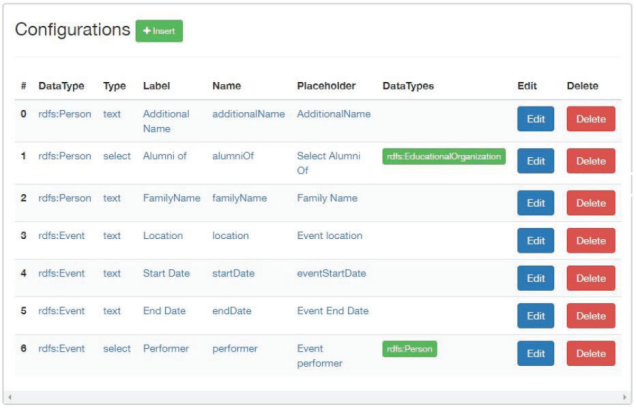


Fig. 3. Configuration handling

number of types, the user is prompted to pick the appropriate data types contained in the database. Figure 3 shows the user interface for the configuration specification functionality. The edit and delete functionality are available for this part as well.

Template definition can be considered the most important functionality of the UX provider, since it offers the means for defining layouts to be accessed by the public, as well as the bindings for the results obtained by executing specified queries for specific display contexts. When defining templates, the user can select the appropriate data type and the display context.

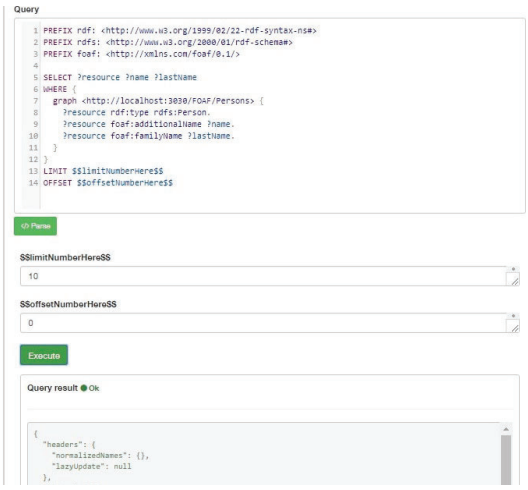


Fig. 4. Creating, binding and executing a SPARQL query

A SPARQL query must be defined for retrieving the most appropriate data for the selected display context, and this query may contain variables that can be resolved in runtime, depending on the display context. Such variables may include resource id, offset and limit numbers. The query can be executed directly in the page, by first binding the variables with input values and executing the parsed query. SPARQL response will be displayed in the specified container, showing whether the query parsing succeeded or incorrect data had been entered in the binding stage. Figure 4 shows defining, parsing and executing the query.

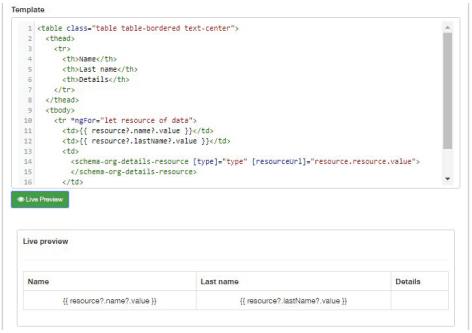


Fig. 5. Defining the template layout

The results obtained by the SPARQL response can be represented in a HTML layout, which can be defined in the appropriate section, along with the live preview option. The response data bindings must be in compliance with Angular interpolation style, in order for the values to be rendered correctly. Additional bootstrap classes, or even custom CSS styles can be added for further layout customization and improved user experience. Figure 5. represents this step.

The application offers custom defined wrapper components for the buttons that can be used in any of the layouts, annotated with schema-org-details-resource. These components require input based on retrieved data from the SPARQL query and can be placed anywhere in the layout.

IV. The Architecture

FacetC offers a wide range of functionalities, which are powered by the underlying layered architecture configured for code and component reusability.

To begin with, the data layer consists of two sub layers, one for a relational database, and one for SPARQL endpoints. The relational database is used to store information about the data types used in the dataset, the specified configurations, as well as the template data - the queries and layout templates.

This data is retrieved and stored by a template handler, represented by a service, which gets invoked by the configuration handler. This service is responsible for retrieving the appropriate configuration and template from the database, and pass the query obtained from the template along with the user specified parameters to the query builder. This service performs the required bindings and executes the SPARQL query.

The obtained result is then transfered to the end user journey components, along with the appropriate layout templates, and that data is rendered in the browser by leveraging the power of Angular's Just In Time compiling. That result can also be redirected to the UX provider panel, whenever the user decides to execute a query in the template definition section. This architecture is visualized in Figure 6.

V. Discussion and Future Work

FacetC is a facet browser platform that supports specific UX for diverse resource types and contexts with performance op-

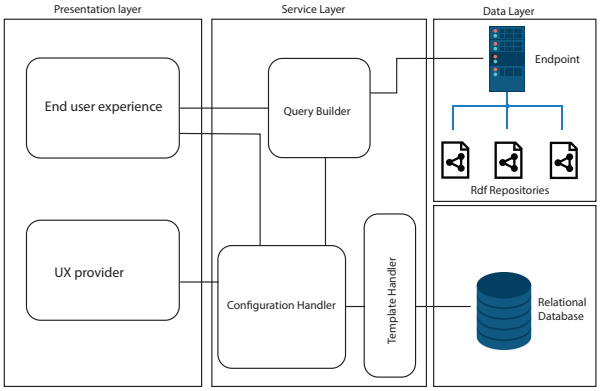


Fig. 6. Application architecture describing the main building blocks

timization through specific SPARQL queries for the specified template, and provides a marketplace with customizable and extensible UX for the common templates. We are now facing new challenges into making the platform more user friendly and upgrading it with other complex features.

An interesting feature that is proven to be challenging as well adheres to resource grouping and classification into a meaningful hierarchy. Dealing with pre-configured classification attributes is a straightforward task, but the continuously updated resources and collections present a tough challenge. Implementing this feature in our application by utilizing the recently proposed Reconfigurable Faceted Thesauri [5] can take it's own priority as well.

We can classify the guided natural language transformation into SPARQL as a generally useful feature for this particular browser platform. Several systems have been proposed regarding this topic, like SPARKLIS, a Semantic Web tool that helps users explore and query SPARQL endpoints by guiding them in the interactive building of questions and answers, from simple ones to complex ones [3] and other platforms that utilize query processing approach based on associated semantic context inference [11].

VI. Conclusion

The general public requires a special tool for interacting with Linked Open Data Cloud that provides easy to use interface with CRUD capabilities and customizable interface at hand. Our proposed application overcomes these challenges and provides the appropriate user experience for interacting with interlinked datasets, while visualizing results in a highly customizable fashion.

References

[1] A. Andreevski, R. Stojanov, M. Jovanovik, and D. Trajanov. Semantic web integration with sparql autocomplete. In *The 12th International Conference on Informatics and Information Technologies*, pages 1–4, 2015.

[2] N. Bikakis and T. Sellis. Exploration and visualization in the web of big linked data: A survey of the state of the art. *arXiv preprint arXiv:1601.08059*, 2016.

[3] S. Ferré. Sparklis: an expressive query builder for sparql endpoints with guidance in natural language. *Semantic Web*, 8(3):405–418, 2017.

[4] T. Finin, L. Ding, R. Pan, A. Joshi, P. Kolari, A. Java, and Y. Peng. Swoogle: Searching for knowledge on the semantic web. In *Proceedings of the National Conference on Artificial Intelligence*, volume 20, page 1682. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, 2005.

[5] J. Gayoso-Cabada, D. Rodríguez-Cerezo, and J.-L. Sierra. Browsing digital collections with reconfigurable faceted thesauri. In *Complexity in Information Systems Development*, pages 69–86. Springer, 2017.

[6] R. V. Guha, D. Brickley, and S. Macbeth. Schema. org: evolution of structured data on the web. *Communications of the ACM*, 59(2):44–51, 2016.

[7] M. Hildebrand, J. van Ossenbruggen, and L. Hardman. /facet: A browser for heterogeneous semantic web repositories. In *International Semantic Web Conference*, pages 272–285. Springer, 2006.

[8] M. Janevska, M. Jovanovik, and D. Trajanov. Html5 based facet browser for sparql endpoints. In *Proceedings of the 11th Conference for Informatics and Information Technology*, pages 149–154.

[9] A. Khalili and A. Merono-Penuela. Wysiwyq–what you see is what you query, 2017.

[10] M. G. Skjæveland. Sgvizler: A javascript wrapper for easy visualization of sparql result sets. In *Extended Semantic Web Conference*, pages 361–365. Springer, 2012.

[11] Y. Yao, J. Yi, Y. Liu, X. Zhao, and C. Sun. Query processing based on associated semantic context inference. In *Information Science and Control Engineering (ICISCE), 2015 2nd International Conference on*, pages 395–399. IEEE, 2015.

Aggregator of repositories and archives with the implementation of the protocol OAI-PMH

Bojan Despodov
Faculty of Computer Science
Goce Delcev University
Stip, Macedonia
despodovbojan@gmail.com

Zoran Despodov
Faculty of Natural and Technical Sciences
Goce Delcev University
Stip, Macedonia
zoran.despodov@ugd.edu.mk

Todor Cekerovski
Faculty of Electrical Engineering
Goce Delcev University
Stip, Macedonia
todor.cekerovski@gmail.com

Abstract— **The Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) is a protocol established for harvesting (or collecting) metadata descriptions of records in an archive so that services can be built using metadata from a number of repositories. Publications from Macedonian universities, magazines and libraries are available through their repositories. These repositories have implemented support for the OAI PMH metadata harvesting protocol and allow harvesting of their records by service providers which using the OAI-PMH protocol collect the records from multiple repositories and archives and enable search of records from multiple different repositories and archives through a single interface.**

In this paper is elaborated how with our implementation of the OAI-PMH protocol as a service provider to collect metadata such as title, author, topic, description, publisher, date, type, etc. from each record in these repositories and archives and later use the collected data to develop a Web application. The developed Web application provides advanced search and analysis of the records, as well as links to the repositories and archives websites from which these records are collected. To achieve these goals algorithms in the PHP programming language were developed by the authors of this paper that collect and store the metadata of records from these archives in a separate MySQL database, and enable search by appropriate attributes of records. This paper first discusses the basic characteristics of the data harvesting protocol OAI-PMH followed by a description of concepts and technologies in the process of data extraction. Then the application of these concepts in the PHP programming language is shown for parsing and storing metadata from every record in Macedonian repositories and archives.

Keywords—OAI-PMH, metadata harvesting, service provider, parsing, PHP, MySQ

I. INTRODUCTION

The number of publications that are allowed to be freely available on the Internet by institutions that have joined the Open Archives Initiative such as libraries, research institutions, scientific and cultural archives, is constantly growing. Libraries are trying to find a faster and better way to provide access to resources they own or resources available elsewhere. The different content of the various repositories can be described through metadata. Metadata is data about data. Metadata can be used for almost any content such as texts, pictures, movies, web pages and software. In July 1999, Paul Ginsparg, Rick Luce and Herbert Van de Sompel sent a Call for Participation in a meeting to explore the collaboration between scientific electronic archives. The meeting, held in October 1999 in Santa Fe, led to the establishment of an Open Archives Initiative (OAI). The Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) has found a widespread adoption for metadata exchange. It is used to collect metadata descriptions so that services can be developed using metadata from multiple archives [2].

Commercial search engines have also begun to use OAI-PMH to collect more resources. Google uses OAI-PMH to collect information from the National Library of the Australian Digital Repository. In 2004 Yahoo retrieved content from OAlster (University of Michigan) by collecting metadata with OAI-PMH. In this paper, techniques for collecting data were applied to automatically retrieve metadata and store the information on a separate MySQL database. The stored data were then used to develop a Web application that can be used as a tool for analyzing and searching downloaded records from various Macedonian repositories and archives.

II. OAI-PMH PROTOCOL

OAI-PMH provides an application independent framework for interoperability based on metadata collection. OAI-PMH is based on client-server architecture and uses XML over HTTP. A full picture of what OAI-PMH represents is shown in Fig. 1. With simple terms, the data that should be available by entities named Data Providers are available through OAI-PMH servers viewed as Repositories. The data can be collected via OAI-PMH clients, named as Harvesters, which are managed by entities called Service Providers, these entities are called “providers”, since it is assumed that they will again use the data that is collected to provide some type of service [3].

A. Characteristics of OAI-PMH

OAI-PMH does not support search, because its purpose is to transfer defined metadata sets between repositories. The services provided through the data are outside the scope of the protocol.

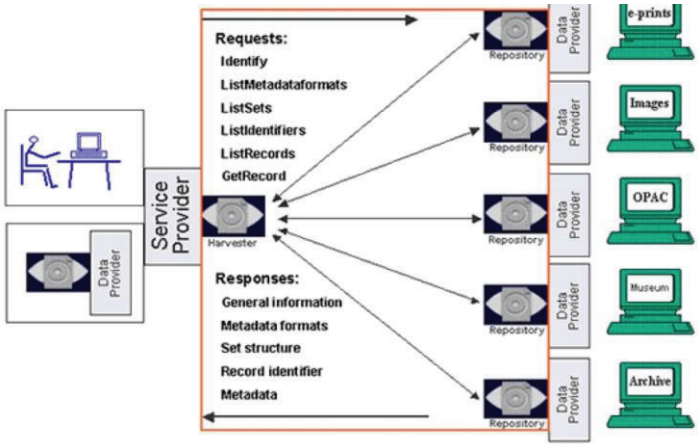


Figure 1. OAI-PMH: review and structure of the model.

This also means that it is scalable because all the data that is needed can be stored in a central repository and queries or processing can be made locally. There are two types of collection criteria that can be combined in the OAI-PMH request: sets and date stamps. The data available in the repository can be grouped into sets. The repository at least one set, but data providers may decide to publish several sets (one record may belong to more than one set). When collecting data with the selection of a set, all records of that set are given to the answer. Data collection using a date bookmark selection can be used to request records that belong to the set that are created, deleted, or modified in a given time period. Requests in OAI-PMH are submitted using HTTP GET or POST method. Responses are always XML documents encoded in UTF-8 in accordance with the XML schema, and have a description in the response to the request. Dublin Core is a mandatory format, but the repository can provide records in other metadata formats. An important feature of OAI-PMH is its simplicity: the requests are simply one GET or POST request, the answers are given in XML and the protocol is asynchronous (the only information that the client must store is resumptionToken for large sets of records).

B. Basic concepts of OAI-PMH

1) Types of participants: As previously mentioned, the two types of participants were defined: **Data providers** that represent entities that support OAI-PMH as a tool for metadata exposures, and **Service Providers**, who use the collected metadata to build value-added services. Some systems can act as both types.

- **Data Providers** refer to entities that have metadata and are willing to share with others
- **Service Providers** are entities that collect data from Data Providers to provide high-level services to users. Typical services are: search, browsing, and so on. An interesting scenario is when Service Providers can process data and use them to provide services like Data Providers (such as collecting all data from a group of libraries and re-providing
- them as regrouped in new sets by mixing records from more than one source).

2) OAI-PMH Software:

- **Harvester** is a client application that sends OAI-PMH requests. It is managed by Service Provider as a tool for collecting metadata from the repositories. It is the mechanism that Service Providers have for obtaining data from Data Providers.
- **Repository** is a server available in a network capable of processing OAI-PMH requests. It is managed by a Data Provider to expose metadata in order to be collected. To explain what a repository is, there are three concepts that must also be described. The first concept is **Resource** – which is what metadata relates to, the second is **Item** which is an integral part of the repository and the thirds concept is **Record** which is metadata of an item in a specific metadata format encoded in XML.
- **Examples of deployment** – it is possible to see a network of systems that support OAI-PMH as in Fig. 2. There are multiple Data Providers and Service Providers. Each Service Provider collects (harvests) data from multiple Data Providers [3].

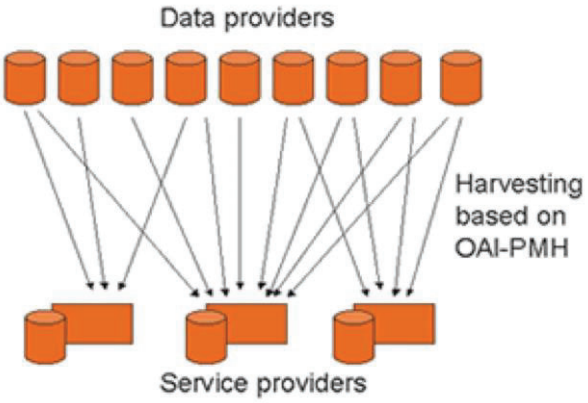


Figure 2. More service providers.

A more complex example is shown in Fig. 3, a network of systems that support OAI-PMH. There are multiple Data Providers for each Service Provider, and also an aggregator that collects metadata from multiple Data Providers, which can also provide data to Service Providers [3].

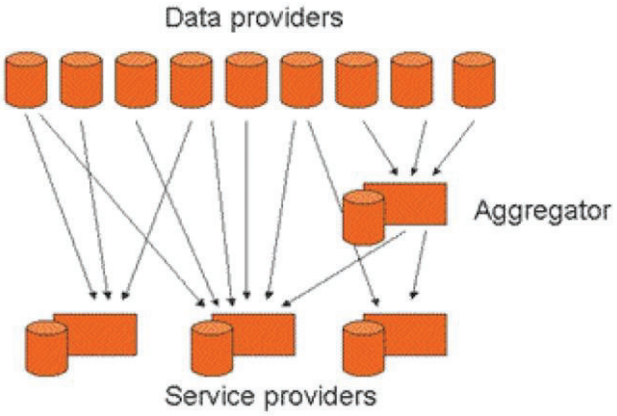


Figure 3. Aggregators.

3) Metadata formats: OAI-PMH uses metadata of resources rather than resources themselves. Bellow the resources are the items, the most abstract entity in OAI-PMH. The item has a unique identifier for all metadata of resources. Below item are records. The records have unique identifiers and contain metadata in a specific format in XML: MARC, Dublin Core, Qualified Dublin Core, MODS, METS, TEL-AP, etc.

4) Identifiers: An identifier in the OAI-PMH must be unique so it can unequivocally identify an item within a repository. The unique identifier is mapped into an item, and all possible entries available from one item share the same unique identifier. Unique identifiers play two roles in the protocol: in the answer for the verbs ListIdentifiers and ListRecords to identify each record, and request a specific record in the GetRecord request.

5) Sets in repository are an OAI-PMH mechanism to enable the collection of sub-collections whose semantics are defined outside the protocol. The sets are defined by conventions defined between Data Providers and Services, or only from the Data Providers. They may be more logical aggregations such as: Magazines, Institutional Repositories, and so on.

6) Deleted entries: deleted entry means that the record was once in the repository but is no longer available. There are three types of support for deleted entries: no support, persistent and transient.

C. OAI-PMH Requests and Resumption Token

As previously described, all requests are HTTP GET or POST requests, and any responses, any errors or regular responses are given in XML. The following section will describe the six types of requests available in OAI-PMH. In OAI-PMH there are six types of requests (known as verbs) that can be added to the OAI based URLs along with other arguments to access the contents of the repository. Although OAI-PMH is intended for machine to machine communication, it returns the results as an XML that can be displayed by all known Web browsers [4]. For this reason, the following examples are shown as direct links.

1) Identify: The server responds with information about the repository. Some of this information is required, namely: its name, primary URL, version of OAI-PMH protocol, the earliest date bookmark, the type of support for deleted entries, the administrator's email and granularity of time. The basic OAI-PMH URL address of the repository at the University "Goce Delcev" is http://eprints.ugd.edu.mk/cgi/oai2. Adding the Identify request to the basic OAI-PMH URL address of this repository will result in:

http://eprints.ugd.edu.mk/cgi/oai2?verb=Identify.

The results of this request are shown on Fig. 4.

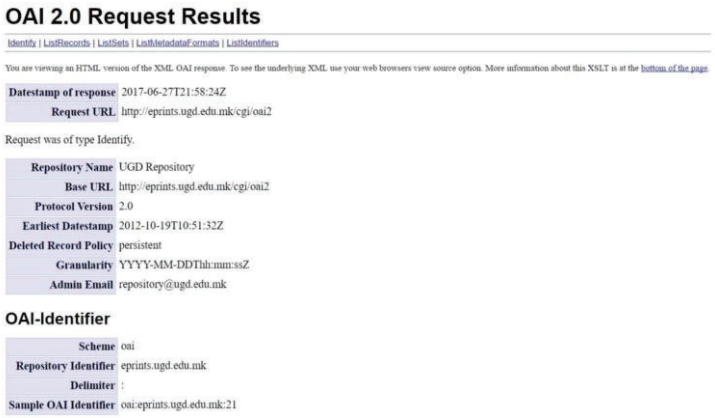


Figure 4. Results from OAI-PMH Identify request to the UGD Repository.

2) ListMetadataFormats: Lists the available metadata formats by the repository. By adding the ListMetadataFormats request to the basic OAI-PMH URL of the UGD Repository will result in

http://eprints.ugd.edu.mk/cgi/oai2?verb=ListMetadataFormats.
Part of the results of this request are shown on Fig. 5.

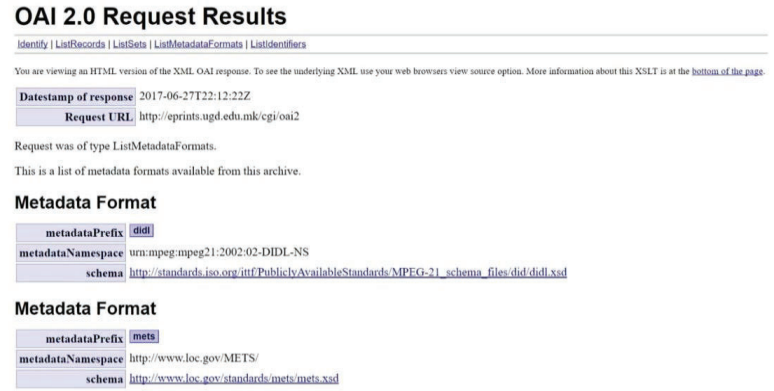


Figure 5. Results from OAI-PMH ListMetadataRecords request to the UGD Repository.

3) ListSets: This set is used to retrieve the available sets of the repository. As an optional argument, it is possible to use Resumption Token for flow control, which is described later in the paper. Adding ListSets request to the basic OAI-PMH URL address of the repository at University "Goce Delcev" will result in http://eprints.ugd.edu.mk/cgi/oai2?verb=ListSets.

Part of the results of this request are shown on Fig. 6.

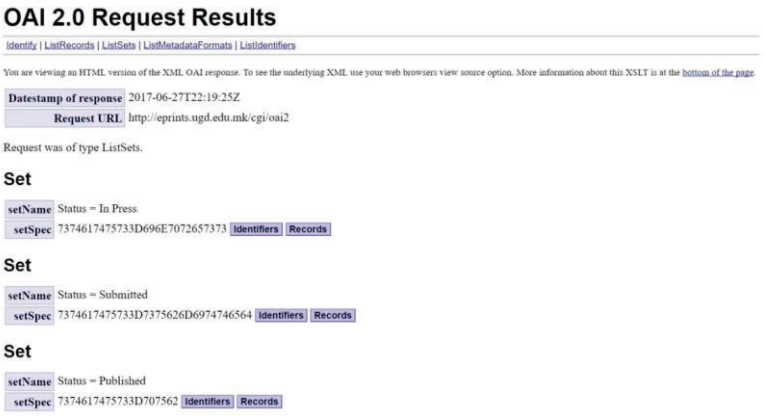


Figure 6. Results from OAI-PMH ListSets request to The UGD Repository.

4) ListIdentifiers and ListRecords: These requests are very similar, and both return the list of records identifiers. The difference between them is that the ListRecords response also includes records in the metadata format that is followed as a parameter. It is possible to limit the list that is returned using arguments from and until, which limits the response to the records updated after a given time and before the given time accordingly. It is also possible to limit the records to a given set using the setSpec parameter. If the repository stores information about deleted entries, the response also includes the entries that were deleted, displaying in the status of the attribute of the title the value "deleted". As with the ListSets verb, it is possible to use Resumption Token for flow control which is later described. Adding ListIdentifiers request to the basic OAI-PMH URL address of the repository at "Goce Delcev" University will result in: http://eprints.ugd.edu.mk/cgi/oai2?verb=ListIdentifiers.

Part of the results of this request are shown on Fig. 7.

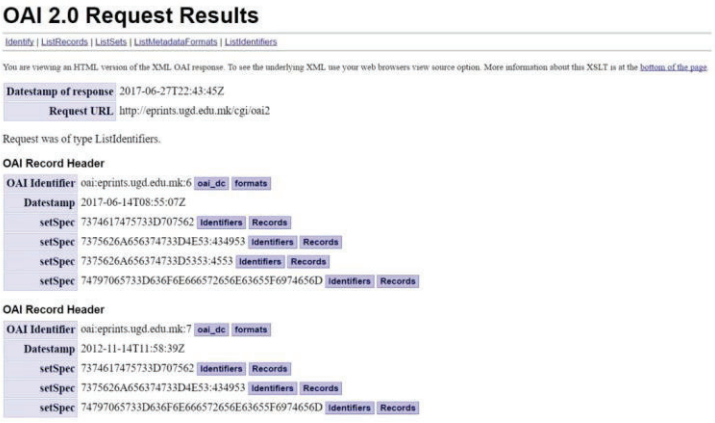


Figure 7. Results from OAI-PMH ListIdentifiers request to the UGD Repository.

Adding ListRecords request to the basic OAI-PMH URL address of the repository at "Goce Delcev" University will result in: http://eprints.ugd.edu.mk/cgi/oai2?verb=ListRecords. Part of the results of this request are shown on Fig. 8.

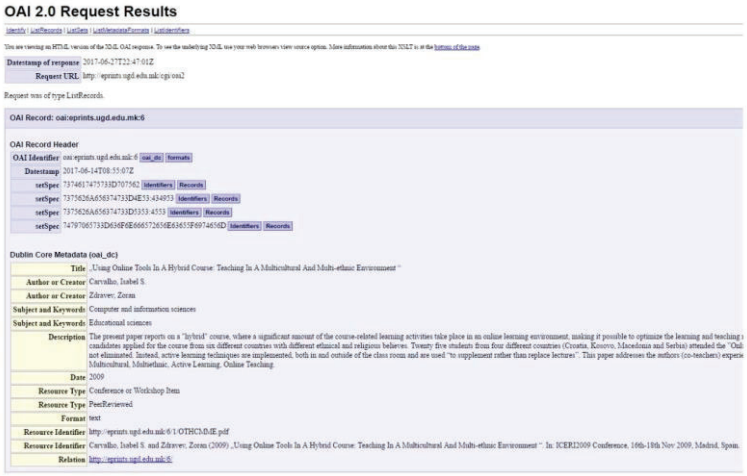


Figure 8. Results from OAI-PMH ListRecords request to the UGD Repository.

5) GetRecord: Gets a record from the repository by specifying the identifier and the metadata format in which the record should be returned (metadataPrefix parameter). If the repository stores information about deleted entries, it is also possible to monitor if the record was deleted, displaying in the status of the attribute of the title value "deleted". Adding the GetRecord request to the basic OAI-PMH URL address of the repository at "Goce Delcev" University will result in: http://eprints.ugd.edu.mk/cgi/oai2?verb=GetRecord&metadataPrefix=oai_dc&identifier=oai:eprints.ugd.edu.mk:6. The results of this request are shown on Fig. 9.

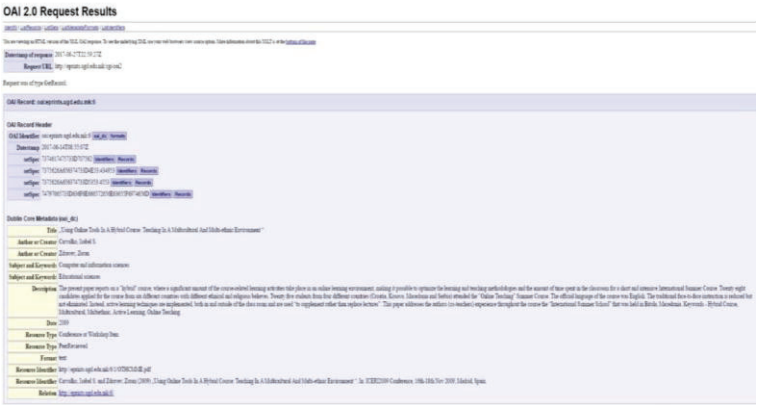


Figure 9. Results from OAI-PMH GetRecord request to the UGD Repository.

6) Resumption Token: Verbs ListSets, ListIdentifiers, and ListRecords return a list of discrete entities. When these lists are too long to be sent in one response, it may be practical to split the list and take a series of requests and answers. Mechanism in OAI-PMH who deals with this is the use of resumption token [1]. When a request can not be answered in one response, a list is sent in the response together with the resumption token, which can be used in the next request to obtain the next partition from the list. There are two optional attributes in the following: expiration date, the overall size of the required list and the cursor

(the number of items returned. On the list that is returned, starting with 0).

III. Used Technologies

The Web application "Aggregator of repositories and archives" represents implementation of the OAI-PMH protocol as a Service Provider. The Web application using the OAI-PMH protocol collects the records from multiple Macedonian repositories and archives and enables search of records from multiple different repositories and archives through a single interface. The application provides advanced search and analysis of the records, as well as links to the repositories and archives websites from which these records are collected.

This application was built on two core technologies. The first technology is the PHP programming language which is the most used server-side scripting language for development of web applications. PHP has been widely ported and can be deployed on most web servers on almost every operating system and platform. The second technology is MySQL open-source relational database management system which is used for storing of the collected records from the repositories. Technology for web server that was used is Apache. Our application represents full implementation of OAI-PMH Service Provider. The supported metadata format is Dublin Core and with little modifications of the code all other metadata formats could easily be added. This application can send OAI-PMH requests of all six types and deals with all types of XML responses from repositories. Additionally, Resumption Token mechanism is fully implemented. The architecture of the application is shown on Fig. 10.

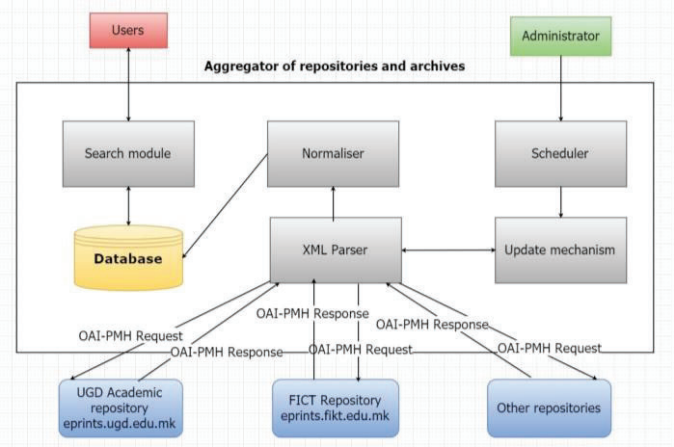


Figure 10. Architecture of the web application.

The sections with grey on the picture are large number of php scripts. Scheduler scripts are responsible for collecting all the records from given repository. Update mechanism scripts are responsible for collecting new records that are added to the repositories after last harvesting from the application. XML Parser scripts send OAI-PMH requests to repositories and extract(parse) data from responses from repositories which are XML documents encoded in UTF-8 in accordance with the XML schema. Normaliser scripts prepare the extracted data from XML responses and inserts them into a local MySQL database. Search module scripts allow searching of harvested records from repositories based on some tags such as: title, author, topic, description, publisher, date, type, etc.

IV. The Application

The functionality and features of the created web application will be explained in this section. The main interface of the web application is shown on Fig. 11.

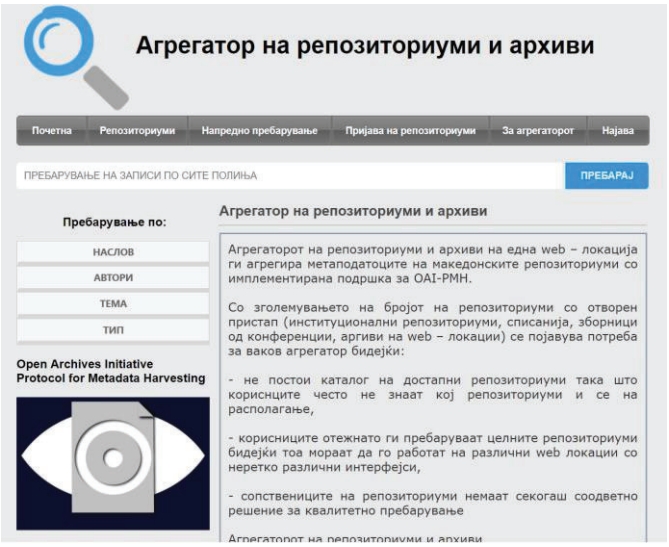


Figure 11. Main interface of the web application.

The application consists of six sections. The section add a repository allows users to add a new repository (name, OAI-PMH URL, description, type admin email, metadata format) which the administrators needs to review before enabling data retrieval. This section is shown on Fig. 12.

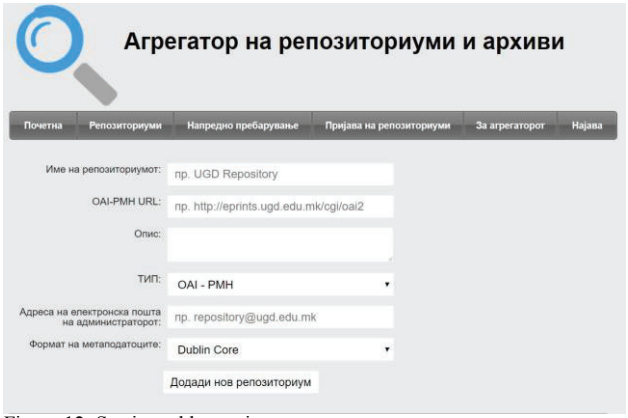


Figure 12. Section add repository.

After login administrators can see a list of reported repositories by users with their respective OAI-PMH URL. Admin page is shown on Fig. 13.

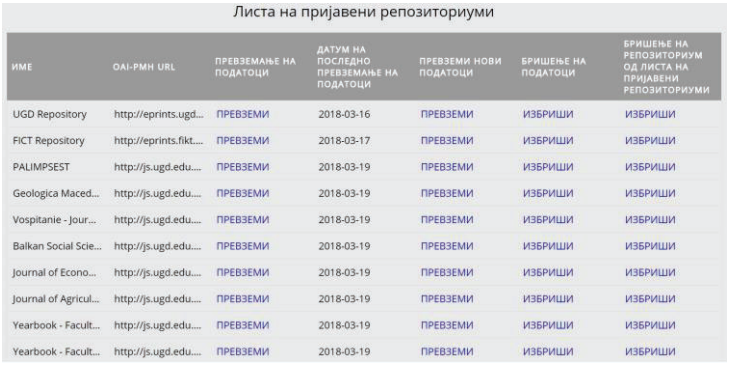


Figure 13. Admin page.

Administrators can delete the OAI-PMH URL of the reported repositories and can also enable data retrieval from repositories. The date of the last harvesting of metadata is shown next to each repository. Administrators can also harvest new metadata that is added to the repositories after the last harvesting. Metadata that is harvested from specific repositories can be deleted by the administrators. The section for repositories displays each repository (name, description) and number of harvested records from it. If the user clicks on a repository all of the harvested records from that repository will be displayed. The section for repositories is shown on Fig. 14.

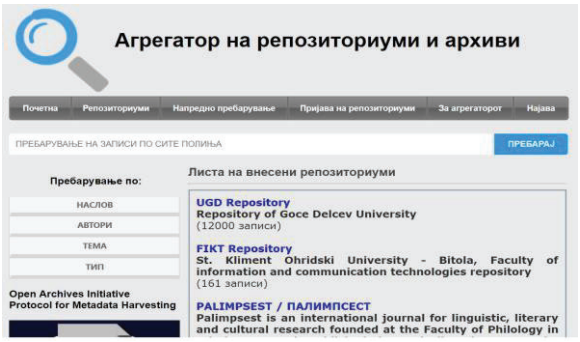


Figure 14. Section for repositories

The advanced search section allows users to search for records by entering values for: repository, name, authors, topic, description, publisher, date, type, format, identifier, source and language. On the main web page of the application users can choose to search the records only by title, authors, topic and type by clicking the corresponding button. On the main web page of the application there is a search form and users can enter search term to search the records by all attributes. For example, if we enter the term "programming" in the search box the web application will show all the records from repositories that contain the term "programming" in some attribute (title, authors, topic, description, publisher, date, type, format identifier, source, etc.). Some of the results of this search are shown on Fig. 15.

As you can see on the image, the application displays the records found in a results list with title, authors and date. If the user clicks on the title of some record or on "view record" from the results list a new page will appear with more information (such as: topic, description, publisher, type, format, identifier, language, URL) about the selected record. By clicking on "view original" new tab will open with the full URL from the repository from where the selected record is harvested.

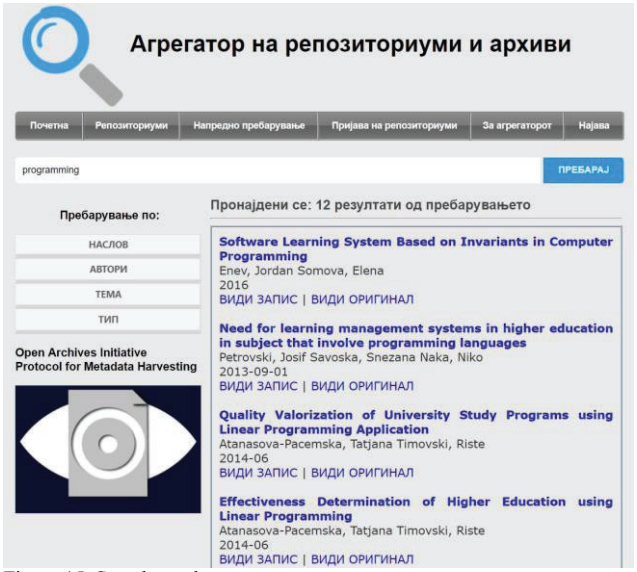


Figure 15. Search results.

V. Discussion and Conclusion

The process of development of this web application was extensive and required a lot of time. The results are very positive because we have made a full implementation of OAI-PMH Service Provider which can harvest all of metadata records from Macedonian repositories and archives and store the records in local database which later is used for search of records from all of these repositories. The testing was done on real Macedonian repositories and archives. The total number of added repositories and archives in this OAI-PMH service provider is 18, and the total number of harvested records is 13,360.

This application should get approval and support from the Government and the Ministry of education of Republic of Macedonia for future work, especially because an application of this type, also written in Macedonian language does not exist. The application can be used at the state level to collect the publications from all Macedonian universities, libraries, journals and archives. The search of publications will be simple through this application because publications from all repositories from across the country can be searched at a one central location. In the future, we hope to improve the web application by adding Linux cron jobs, more metadata formats and more features.

References

[1] OAI – PMH Initiative, https://www.openarchives.org/pmh/.

[2] Pavel Simek, Jan Jarolimek, Jiri Vanek, Michal Stoces, "Implementation of Metadata Harvesting of Scientific and Scholarly Research Journal's Content", Proceedings of the International Conference on Information and Communication Technologies for Sustainable Agri-production and Environment, Skiathos, September 2011, pp 827-829.

[3] Diogo Reis, Nuno Freire, "OAI – PMH implementation and tools guidelines", TEL plus, ECP-2006-DILI-510003, 2008, pp 3-9.

[4] Sem Gebresilassie, "Harvesting Statistical Metadata from an Online Repository for Data Analysis and Visualization", Helsinki Metropolia University of Applied Sciences, 2015, pp 8-14.

Importance of Quality Assurance in Software Products

Aldion Ambari
Faculty of Computer Science and Information Technology
University American College Skopje
Skopje, Macedonia
aldionambari@yahoo.com

Adrijan Bozinovski
Faculty of Computer Science and Information Technology
University American College Skopje
Skopje, Macedonia
bozinovski@uacs.edu.mk

***Abstract*— As software is integrated more frequently into every aspect of our lives, as it grows more quickly in size and function, as its failure in operations causes increasingly devastating consequences, and as schedules and budgets are continually reduced despite the need for high-quality, reliable, and secure software, advanced and innovative technologies must be developed to achieve software quality assurance more effectively and efficiently. Software Quality Assurance (SQA) is defined as a planned and systematic approach to the evaluation of the quality of and adherence to software product standards, processes, and procedures. SQA includes the process of assuring that standards and procedures are established and are followed throughout the software acquisition life cycle. The word assurance means „guarantee". So the Quality Assurance Group's role is to guarantee that the product is of high quality. With this definition, it is imperative that the SQA helps an organization in continuous performance improvement and strive for perfection. This paper presents the impact of software quality assurance and software testing in the lifecycle of software development. It shows how important the implementation of software quality assurance is for applications used every day by every user. What will be presented in this paper here is that the software quality assurance is a part of the agile development process and not just a step that needs to be utilized at only a certain point. The key characteristics and best practices to achieve the quality requirements that are needed for successful software will also be presented.**

Keywords—Quality requirements, Software quality, Quality criteria requirements, Quality management software

I. INTRODUCTION

Software quality assurance (SQA) is a process that ensures that developed software meets and complies with defined or standardized quality specifications. It involves the entire software development process, monitoring and improving the process, making sure that any agreed-upon standards and procedures are followed, and ensuring that problems are found and dealt with. An ongoing process within the software development life cycle that routinely checks the developed software to ensure it meets desired quality measures. Its practices are implemented in most types of software development, regardless of the underlying software development model being used. In a broader sense, software quality assurance incorporates and implements software testing methodologies to test software [1].

The evaluation of the software based on certain attributes is a short explanation of what software quality means when mentioned. Software quality can be explained based on the study of external and internal features of the software (Figure 1).

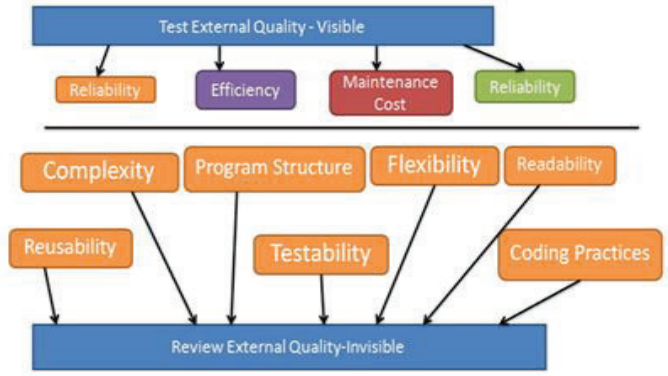


Figure 1: Software Quality

How useful it is for its users and how software performs in real time scenarios in operational mode, are what explain external qualities of the software. On the other hand internal quality focuses on the internal aspects that depend on the value of the written code.

II. BEST PRACTICES FOR SQA

Following are some practices which are currently considered to be the best [2].

A. Measure the Importance of Quality

The importance of the quality of the software application they are going to develop is what the development team needs to be aware of. The QA team is explicitly needed for planning the development approach, high quality design and code. To prevent the programming errors or bugs this practice can become an initial step for the future [3].

B. Quality Benchmarks

An important part of the development cycle as well as quality assurance is gathering requirements. Developers and business owners need to know everything about the application to be developed before documenting them. We can refer this as verification and validation. Verification typically involves review and meeting to evaluate documents, plans, code, requirements and specifications [4]. Validation on the other side involves actual testing and takes place after verifications are completed.

C. Continuous Testing

Performing tests continuously on every minor release in order to check whether each build is successful is what defines continuous testing. If this doesn't work we can still detect them on time and make necessary changes and again take them into the testing process [5]. This is a process which for continuous delivery of the end product every QA team needs to adopt.

D. Implement Automated Testing

The stupendous benefits of automated testing can no longer be unaware in the process of quality assurance. That's why implementing automation tests can be considered as an important recommendation in the quality assurance process. When real world scenarios are taken into consideration, testing software applications with manually written test scripts will make it limited to a great extent [6]. For smaller projects, the time needed to learn and implement them may not be worth it unless personnel are already familiar with the tools. But for larger projects, or ongoing long-term projects they can be valuable.

E. Error Reports

Proper error reports are very important for developers to understand the areas of loopholes. On the other hand, un-organized and unclear error reports, can lead to serious misgivings [7]. Very helpful for tracking errors are the latest automated testing tools which have built-in integration with various tracking tools so error reporting becomes much easier.

F. Time Management

Just the way errors cannot be underestimated if they don't appear in the screen, from becoming a QA expert to finding errors and understanding the risk level of errors needs time for evaluation, reporting, resolution and re-testing those scenarios. While planning the application release dates, quality assurance team has to allocate appropriate time for such unexpected circumstances [8].

III. Proposed Methods to Achieve Quality

The first rule of writing a good program is "Don't write it yet". Instead, look for an existing effort that accomplishes all or most of what you want to do, and make use of it. The tendency of software shops to re-implement existing software poorly has been nicknamed the "Not Invented Here" syndrome. If you can reuse existing code to do what you want, then you will finish much more quickly.

Keep your text and graphic files separate until after the text has been formatted and styled [9]. Do not use hard tabs, and limit use of hard returns to only one return at the end of a paragraph. Do not add any kind of pagination anywhere in the paper. Do not number text heads-the template will do that for you.

A. Code Design

Code design is an initial mental effort done to plan the way the code will evolve into the future. Most software developers when confronted with the question, will agree that it is a good idea to design code well before writing it, as fixing badly designed code may be much more costly in time [10]. Good internal design is indicated by software code whose overall structure is clear, understandable, easily modifiable and maintainable. Good functional design is indicated by an application whose functionality can be traced back to customer and end-user requirements

B. Refactoring and rewriting code

Refactoring is the process of improving the internal quality of the code (or its elegance) without changing its external behavior. Refactoring can be done by applying several well-understood transformations to the existing codebase, while preserving its integrity at every point, requiring very little effort and with relatively little risk of breaking the code [11].

C. Quality throughout the Development Process

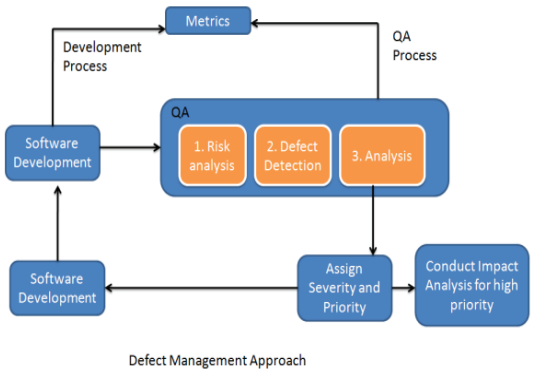
The first step in the approach is the analysis of the software development process itself. Software development, like software itself, is a sequential arrangement of activities which transforms software requirements into software products by way of logical process. Ensuring that the development team is on the same page from the outset is one of the most important facets of quality assurance is [12]. Analysts and developers should understand the needs of the end user. In turn, they will outline the architecture and functionality of the software and choose the proper programming languages, libraries, and tools for the job.

IV. Proposed Methods to Determine Quality

Presently two important approaches are used to determine the quality of the software:

- Defect Management Approach
- Quality Attributes approach

What usually leads to design error is the failure of the development team to understand the requirements of the client.



Defect Management Approach

Figure 2

Other reasons can also be poor functional logic, wrong coding or improper data handling. The defect management approach can be applied in order to keep track of any particular error. In defect management, categories of

defects are defined based on severity as shown above (Figure 2).

Figure 3

Quality Attribute Approach on the other hand focuses on six quality characteristics as shown above (Figure 3) [13].

V. Types of Testing

- Black box testing – not based on any knowledge of internal design or code. Tests are based on requirements and functionality.
- White box testing – based on knowledge of the internal logic of an application's code. Tests are based on coverage of code statements, branches, paths and conditions.
- Unit testing – the most 'micro' scale of testing, to test particular functions or code modules. Typically done by the programmer and not by testers, as it requires detailed knowledge of the internal program design and code. Not always easily done unless the application has a well-designed architecture with tight code and may require developing test driver modules or test harnesses [14].
- API testing – testing of messaging/data exchange among systems or components of systems. Such testing usually does not involve graphical user interfaces. It is often considered a type of 'mid-level' testing.
- Incremental integration testing – continuous testing of an application as new functionality is added, it requires that various aspects of an application's functionality be independent enough to work separately before all parts of the program are completed, or that test drivers be developed as needed, done by programmers or by testers.
- Integration testing – testing of combined parts of an application to determine if they function together correctly. The 'parts' can be code modules, services, individual applications, client and server applications on a network, etc. This type of testing is especially relevant to multitier and distributed systems.
- Functional testing - functional testing black box type testing geared to functional requirements of an application; this type of testing should be done by testers. This doesn't mean that the programmers shouldn't check that their code works before releasing it (which of course applies to any stage of testing.
- System testing - black box type of testing that is based on overall requirements specifications and covers all combined parts of a system [15].
- End to end testing - similar to system testing, the 'macro' end of the test scale, it involves testing of a complete application environment in a situation that mimics real world use, such as interacting with a database, using network communications, or interacting with other hardware, applications, or systems if appropriate.
- Regression testing re-testing after fixes or modifications of the software or its environment. It can be difficult to determine how much re-testing is needed, especially near the end of the development cycle. Automated testing approaches can be especially useful for this type of testing.
- Acceptance testing - final testing based on specifications of the end user or customer, or based on use by end-users/customers over some limited period of time.
- Load testing - testing an application under heavy loads, such as testing of a web site under a range of loads to determine at what point the system's response time degrades or fails.
- Performance testing - term often used interchangeably with 'stress' and 'load' testing. Ideally 'performance' testing (and any other 'type' of testing) is defined in requirements documentation or QA or Test Plans.
- Usability testing - testing for 'user friendliness'. Clearly this is subjective, and will depend on the targeted end-user or customer. User interviews, surveys, video recording of user sessions, and other techniques can be used. Programmers and testers are usually not appropriate as usability testers.
- Security testing - testing how well the system protects against unauthorized internal or external access, willful damage, etc. It may require sophisticated testing techniques [16].
- Compatibility testing - testing how well software performs in a particular hardware, software, operating system, network, etc, environment.
- Exploratory testing - often taken to mean a creative, informal software test that is not based on formal test plans or test cases; testers may be learning the software as they test it

VI. User Satisfaction

A high-quality software application makes its users happy. They enjoy using it and it doesn't get in their ways. It yields the right results quickly, without requiring workarounds for bugs, does not crash or hogs the system, and allows the users to go on

with the rest of their lives. Either the program is invisible to them and they don't think about using it, or it works so well that they enjoy using it and possibly comment about it to their friends.

On the other hand, a software application of poor quality annoys, irritates and/or frustrates its users, or even causes them to lose a lot of time, money or worse. Either it crashes a lot or hogs the system, or it has other bugs like those causing data loss. Whatever its faults are, it fails or partially fails to be a useful tool in the hand of the user.

The user focuses more on how the software works at the external level, but the quality at external level can be maintained only if the coder has written a meaningful good quality code.

Rather than checking for quality after completion, it processes test for quality in each phase of development until the software is complete. With software quality assurance, the software development process moves into the next phase only once the current or previous phase complies with the required quality standards. SQA generally works on one or more industry standards that help in building software quality guidelines and implementation strategies. [17]

As software developers, it is our mission to make sure the software we produce is high quality so it will perform its function properly without inflicting anguish or loss upon the user.

CONCLUSION

Software quality assurance (SQA) is a process that certifies a software application for its quality during the whole software application development phase. Last year, the latest software testing trends were highly useful, which has helped the software testing industry in revamping the current testing strategies. What changes for each application is the software. It is now the software which costs the major proportion of funds from most modern complex computer applications. Software production is currently a labour intensive industry which demands tight control. The nature of software is such that errors can remain undetected for years until the right combination of factors occur to expose them sometimes with unfortunate consequences where the implications on plant operation and safety could be disasterous.

Every piece of software requires changes, updates, and revisions over time. Around launch, these demands are especially high as developers work out last minute tweaks. However, ongoing software maintenance and planning for new versions are all part of the quality assurance challenge.

REFERENCES

[1] C.S., E. (1979). Software Quality Assurance. University of Central Florida, Orlando, Florida.

[2] Fruehauf K., Ludewig J., Sandmayr H. (1986) Software Quality Assurance. In: Güth R. (eds) Computer Systems for Process Control. Springer, Boston, MA (Accessed 12 January 2018).

[3] Schoitsch, E. (1988) 'Software-safety and Software Quality Assurance in Real-time Applications', Austrian Research Center Seibersdorf, [Online]. Available at https://www.academia.edu/17417416/Software_safety_and_software_quality_assurance_in_real-time_applications (Accessed 5 March 2018).

[4] O. Shakiru, A., Ahmad Yusuf, M. (2015) 'Software Quality: Predicting Reliability of a Software Using a Decision Tree', [Online]. Available at http://scialert.net/fulltext/?doi=itj.2014.2755.2759&org=11 (Accessed 10 March 2018).

[5] Yatsyshyn, V., Kharchenko, A., Deren, A., Galay, I. 'Quality Management Requirements in the Process of Creating Software', CSN Department, Temopil Ivan Puluj National Technical University, [Online]. Available at https://www.academia.edu/10224101/Quality_Management_Requirements_in_the_Process_of_Creating_Software (Accessed 2 March 2018).

[6] Heam, J.E. (2016) 'The development of an Automated Testing Framework for Data-Driven Testing Utilizing the UML Testing Profile', Graduate project submitted to Dakota State University, [Online]. Available at http://jameshearn.com/wp-content/uploads/2017/05/Hearn-Development-of-an-Automated-Testing-Framework.pdf (Accessed 20 February 2018).

[7] B. Tomar, A., M. Thakare, V. (2011) 'A Systematic Study of Software Quality Models', International Journal of Software Engineering & Applications (IJSEA), 2(4) [Online]. Avaliable at https://www.academia.edu/34051256/A_SYSTEMATIC_STUDY_OF_SOFTWARE_QUALITY_MODELS (Accessed 22 February 2018).

[8] Garner, B. (2017) 'What is Software Quality Assurance?', Intertech, 7 November, [Online] Available at https://www.intertech.com/Blog/software-quality-assurance/ (Accessed 10 January 2018).

[9] Morse, C.A. (1986) 'Software quality assurance', Sage journals, 1 April, [Online] Available at http://journals.sagepub.com/doi/abs/10.1177/002029408601900302?journalCode=maca& (Accessed 15 March 2018).

[10] Wahyudin, D., Schatten, A., Winkler, D., Biffl, S. (2007) 'Aspects of Software Quality Assurance in Open Source Software Projects: Two Case Studies from Apache Project', Institute for Software Engineering and Interactive Systems, Vienna University of Technology, [Online]. Available at https://www.academia.edu/9309784/Aspects_of_Software_Quality_Assurance_in_Open_Source_Software_Projects_Two_Case_Studies_from_Apache_Project (Accessed 5 February 2018).

[11] Takanen, A., Demott, J., Miller, Ch. (2008). Fuzzing for Software Security Testing and Quality Assurance, Boston.

[12] Walkinshaw N. (2017) Software Inspections, Code Reviews, and Safety Arguments. In: Software Quality Assurance. Undergraduate Topics in Computer Science. Springer, Cham

[13] Test Institute (2014) What Is Software Quality Assurance [Online], Available at http://www.test-institute.org/What_is_Software_Quality_Assurance.php (Accessed 13 January 2018).

[14] Walkinshaw N. (2017) Testing. In: Software Quality Assurance. Undergraduate Topics in Computer Science. Springer, Cham.

[15] Mussa, M., Ouchani, S., Al Sammane, W., Hamou-Lhadj, A. (2009) 'A Survey of Model-Driven Testing Techniques', 2009 Ninth International Conference on Quality Software, 13(1) [Online]. Available at https://pdfs.semanticscholar.org/c70a/da32d6ecccc16bce2148588981101 98f4f54.pdf (Accessed 17 February 2018).

[16] Tian, J. (2005). Software Quality Engineering. Department of Computer Science and Engineering, Southern Methodist University, Texas, Dallas.

[17] Mehdi, Q. (2014). *Software Quality Assurance*. Department of Computer Science, Qassim University, Kingdom of Saudi Arabia.

Netflix - an Example of a Disruptive Innovation Business Model

Kiril Kiroski
Faculty of Computer Science and Engineering, University Ss Cyril and Methodius
Skopje, Macedonia
kiril.kjiroski@finki.ukim.mk

Smilka Janeska-Sarkanjac
Faculty of Computer Science and Engineering, University Ss Cyril and Methodius
Skopje, Macedonia
smilka.janeska.sarkanjac@finki.ukim.mk

Abstract - Internet and its continuous advancement is changing the way we work, communicate, and purchase goods and services. It has great impact on our lives and changes traditional approaches to number of human activities. It destroys obsolete businesses, and promotes new and innovative ones. This paper tries to analyze the business model of Netflix, and the ways it introduces disruptive innovation into the world of television. In this paper, we provide an overview of how Netflix entered already established market, and how they managed, through an innovative use of ICT, to disrupt major players, ultimately creating a completely new market and becoming a major player in it. We will also describe how they are use modern ICT technologies to innovate the ways they create company products and services, work with clients, compete on the market, and create promotional activities.

Keywords - Netflix, disruptive innovation, internet marketing

I. INTRODUCTION

In this paper, we analyze the successful business model and application of ICT in Netflix, an innovative video streaming company. The company, from its foundation till present day has completely changed the way it does business, starting from a company operating in online renting DVDs business, then introducing streaming of video content leased from production and television companies, and all the way through starting their own production of video content.

Today, the success of Netflix and their breakthrough in the world of online television has been copied with different level of success by a number of companies, starting from traditional TV networks, such as HBO (with HBO GO) or CBS (with CBS All Access), all the way through direct online competition, such as Amazon Video Services or Hulu. In this moment, Netflix has the advantage over both groups of competitors, owing it to their dedication and innovation trends, capability of inciting and keeping the interest of their ever-growing user base, as well as the quality of offered content.

Our motivation for analyzing Netflix and their business model is the fact that have taken 7th place on the list of Most Innovative Companies for 2017 and 2nd place for 2018, as well as their ability to engage their users in a captivating experience [1]. What is most fascinating about them is their continuous innovations in internet television experience, the way Netflix's algorithm choses content to present to its users, quality of the service, and the ability to customize to various types of devices. We can also emphasize the way they manage to stay in the focus of their users, how they produce and present new, high-quality content, as well as their business model, where they do not measure and single out viewer ratings for specific content individually, but they are more interested in total subscriber numbers, an approach very much different from traditional television productions.

II. SHORT HISTORY

Netflix is a company founded in 1997 [2] by Reed Hastings and Marc Randolph, with the goal to offer online movie renting, and in 1998 they launch the website netflix.com, through which they provide renting and selling of DVDs. In 1999 they introduce the subscription model, offering unlimited rentals of DVDs for fixed monthly price, a model which in modified form is active today. In 2000 they introduced personalized movie recommendation system, which in its most basic form uses members' movie ratings in precisely predicting their movie choices. In 2002 Netflix appears on Nasdaq, and by 2005 number of their subscribers surpasses 4 million.

Netflix has changed traditional movie renting in 2007 by introducing an online streaming service, enabling their members to immediately watch offered content on their personal computers, and in 2008 they started expanding to other platforms, such as the Xbox 360, Blu-ray players and TV set-top boxes, as well as other Internet connected devices. In 2010 and 2011, Netflix started expanding out of USA, at first in Canada, and then in Latin America and the Caribbean.

In 2013 Netflix starts streaming their original produced television content, starting with TV shows "House of Cards" and "Orange is the new black" complemented by movie production starting with "Beasts of No Nation". At the same time, Netflix continues expanding out of the American continent, and by 2016, their services are offered all around the world. In 2016, Netflix has published 126 original TV shows and movies, and today they have more than 117 million subscribers in 190 countries all over the world. [3]

III. ANALYSIS

A. Innovations

We define innovation as a process of carrying over an idea or invention into goods or services, creating value for clients willing to purchase it. Innovations are incurring great risks, but innovative organizations can provide great advantages over

their competitors, or they can open new markets. Innovations can take the form of an evolution, i.e. gradual improvement of already existing products, or revolution, implying creation of completely new products.

According to Kotler and Armstrong [4], "… Innovation is a messy process—hard to measure and hard to manage. When revenues and earnings decline, executives often conclude that their innovation efforts just aren't worth it. Better to focus on the tried and true than to risk money on untested ideas. The contrary view, of course, is that innovation is both a vaccine against market downturns and an elixir that rejuvenates growth."

According to [5], we recognize four types of innovations:

- Product/service innovations: new or significantly improved goods or services.
- Process innovations: new or significantly improved method for production or delivery.
- Marketing innovations: a new marketing method including significant modifications of product design and package, product promotion or price politic.
- Organizational innovations: a new organizational method of practicing business, workplace organization or external relationships.

Fast Company, the world's leading progressive business media brand, compiling annual lists of the most innovative companies, in the last 10 years are constantly putting Netflix in the most innovative 50 companies in the world. Netflix is pointing out innovations and innovation capability as one of the nine behaviors they found most valuable in their employees [6], meaning they hire and promote people having them. Said document is publicly available on Internet, and it's been viewed more than 17.5 million times. They are defining their employees' capability of innovations as:

- Capability of creating new ideas prove to be useful;
- Re-conceptualization of issues to discover solutions to hard problems;
- Capability of challenging prevailing assumptions and suggesting better approaches;
- Maintaining mobility through complexity minimization and finding time to simplify;
- Thriving on change.

This approach enables them over the years to stay one of the best companies to work for [7] through new concept promotion, such as the one letting their employees use vacation time as they deem appropriate. On the other hand, their organizational approach to employees is "… hire only "A" players to work alongside them. Excellent colleagues trump everything else…" and that "… [if they] wanted only "A" players on [their] team, [they] had to be willing to let go of people whose skills no longer fit, no matter how valuable their contributions had once been."

Netflix innovations are not only organizational. Through the course of the years, they have accomplished product and process innovations. Netflix beginnings are connected to the innovative idea of renting DVDs via mail, where the subscriber can, for a monthly fee smaller than a price of cinema ticket, rent as much DVDs as he can watch. Besides that, they have removed penalties for late movie return, increasing their popularity even more among US subscribers, which puts them as a direct competition to the major video store chain Blockbuster. From 2004 to 2010 they completely succeeded, first to become prevalent on the video content market [8], and after that, using the innovative idea to move from physical medium to online content streaming, to create completely new market where they will be major player for a long time [9]. During this time, Blockbuster is in constant decline, and in 2010 they had to declare bankruptcy.

This great ascent of Netflix does not stem only from following modern trends, but they also start investing tens of millions of dollars every year into streaming technologies research, starting from 2001. During the course of years, they often set up smaller testing groups, offering video content streaming, all the time trying out new approaches and scrapping unsuccessful ideas. Their CEO, Reed Hastings, during these initial years, is often making deals with different content providers, trying to find perfect match for their purpose and always staying ahead of the curve, even in situations when he knew that certain agreement would not prove to be most profitable. They have also tested different price models for streaming content, ending with free offer as a part of DVD subscription. Through this approach, they could get people used to the idea of renting online content, all the while building their content library and not letting to slip up and opening space for the competition. This strategy was based upon four basic principles:

1. Think big;
2. Start small;
3. Abandon quick (not good enough technology);
4. Fast escalation.

This type of innovation that made Netxlix major player in the newly shaped industry is called disruptive innovation, a term widely introduced by Christensen et al. [10], and described on Fig. 1.

On the figure are shown products performance trajectories (red lines describing how products or services are improving over the time), compared to the client requests trajectories (blue lines describing clients willingness to pay for certain performance). As the incumbents are introducing higher quality products or services (upper red line), they are greatly surpassing all of the lower end market client needs and most of

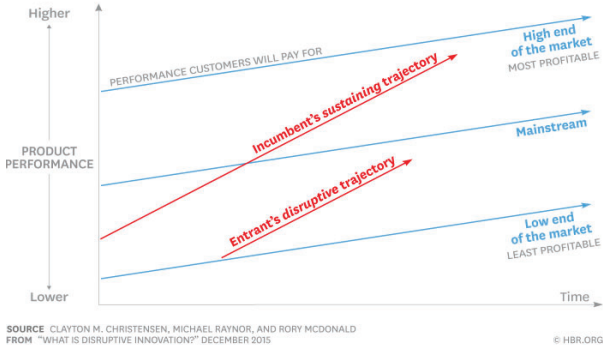


Fig. 1. Disruptive innovation. Taken from hbr.org.

the market mainstream. This enables new players to find their spot in less profitable market slots, ignored by the incumbents. New players with disruptive trajectory (lower red line) improve their products' performances and shift to the upper end of the market (where the profit is highest) and challenge the rule of incumbents.

Disruptive innovation creates new market and value network, displacing established market leaders and alliances, in the end disrupting existing markets and value networks. At the beginning, Netflix focused on a small number of user groups – those not interested in new releases, early adopters of devices intended for DVD reproduction, and on online buyers (relatively small markets in the early 2000s) [11]. One of the main reasons to claim that Netflix is successfully performing disruptive innovation is that they are entering ready for the explosive increase in online streaming demand by 2010, mostly due to their investing in this technology over the years, as well as due to the fact that they provide unlimited access to vast choice of video content, on the user demand, with low price and high quality. Big players such as Blockbuster are completely unprepared for this approach, and they fail, unable to follow Netflix.

Despite their great success, the company does not stop with further innovations. They are constantly improving their offer, start their own production in 2011 and every year investing more and more into it, so in 2017 they create more than 1000 hours of program, investing more than $6 billion. They are promoting this original content in an interesting and humorous way, which also enables them to create their brand and make it more recognizable. Their expansion out of the USA, starting in 2010, is finished by 2016 with global accessibility and more than 117 million subscribers. In 2017 they are innovating the user interface and replace static images with created animated content reviews that are automatically playing as the user moves mouse pointer over the title screen of video content. This enables users to interactively discover what to watch, instead of pointlessly browsing over the video catalogue and wondering if something is worth watching. [12]

B. Competitive Capabilities

Today companies are more competitive than ever. To succeed on today market, companies not only have to become skillful in managing their own products, but also in managing client relations, dealing with their competitors and hostile business environment. Understanding their own customers is of critical importance, but it should not stop there. Innovations help in singling company quality out of the competition, but competitive advantage requires company customers to acknowledge higher value the company offers compared to their competitors.

From a historical perspective, in its beginnings, Netflix had completely different set of competitors than today. Since they are starting as traditional DVD rental service, their historical competitors are video rental stores such as Blockbuster, the alternative DVD exchange systems by Peerflix, and Redbox, automatic kiosk for selling DVDs, Blu-rays and video games.

These historical Netflix's competitors are no longer challenging them, primarily due to their lack of readiness to accept changes. Netflix's success compared to these historical competitors is mostly due to their readiness to adopt new content distribution means, introducing internet video streaming. At the same time, started building their video content library by signing lease contracts with TV and production houses, enabling Netflix to broadcast online their movies and TV series. Using this approach, every customer gets the opportunity to follow his favorite TV series and movies in the way they prefer and in quantity he finds most appropriate. Ken Aulettta [13] is comparing this manner of consuming video content with reading books and he is pointing it out as the main reason that Netflix successfully defeats their historical competitors.

Today, when cable networks and content streaming providers are fighting in every possible way for more subscribers, Netflix's competition is even bigger. Amazon.com Inc. [14] are in the moment their biggest competitors. According to Nielsen, in the second quarter of 2017, 59% of American households have access to at least one subscriber service, mainly split between Netflix and Amazon. Other companies representing potential competitors in this market are Hulu Plus, HBO Now, and CBS All Access, owned by companies long-time present in the content creation market. These Netflix's competitors were its suppliers, being creators of content leased by Netflix and streamed over their online service. Perspective shift is happening when these companies have started their own streaming services, provoking Netflix and other streaming companies, such as Amazon, to involve themselves in creating content for streaming to their customers.

Table 1 shows the overview of advantages and disadvantages of Netflix, compared to their competitors. This overview is showing current service characteristics which, having in mind the field they are competing on, can be drastically and dynamically changing.

Judging by current Netflix investments into creating original and exclusive content, they are considered by already established companies in the market of creating original content as their biggest contemporary competitors. Netflix has in 2017 invested more than $6 billion in creating new content

TABLE I. Overview of Advantages and Disadvantages of Netflix, Hulu, Amazon and HBO Now.

Company	Advantages	Disadvantages
Netflix	Excellent choice of content Optimized interface Good recommendation system	More expensive than most other services Titles are often changing Restricted new content
Hulu	Excellent choice of current TV content Cheap Quality original programming	Frequent commercials Inconsistent choice of older TV seasons Complicated interface
Amazon	Broad choice of content Included other Prime benefits Good original series	Rarely showing newest videos Not available on all platforms Content other than offered is additionally charged
HBO Now	Awarded original programming Blockbuster movies Original news programming	Expensive

[15], and is planning to invest more than $8 billion in 2018 [16]. Of course, their competitors are not idly sitting and watching these attacks, and the company can get into troubles if this aggressive approach fails. Content creation is very expensive, and Netflix is using subscription based model. Creating more content is asking for aggressive battle for more subscribers. Deciding to become content creator has put Netflix on a path that can financially drain them, since producing new and expensive content requires constant subscription growth. Some critics are emphasizing that Netflix is making very little profit in comparison to actual revenues. [17]

C. Customers

Netflix is operating on consumer market. Their potential users include children, people of age from 18 to 60, as well as older, and across all over the world. Today they are operating in single product category, i.e. online content streaming for their subscribers, while in the past they have dealt with DVD renting over mail.

Netflix customers are drastically different among themselves by what they are watching, by their age, by their purchasing power, technology improvements in their environment. The only thing common for all of them is internet access and interest in watching video content. Large part of the subscriber base is used to follow free or pirated content. This fact is giving power to the customers and requires greater involvement of the company trying to attract and retain customer attention. Netflix is using free one-month full subscription model as a basic tool for attracting new customers.

In order to understand, satisfy and service their diversified customer base, Netflix is gathering and maintaining information about their subscribers. Information gathered from their web site and used for following their subscribers pertains to: their e-mail addresses, names, age, credit card and other information, geographic region or country obtained by their IP address, Internet service providers and software used by the customers (type, configuration, unique identifiers), time spent in binge watching movies and TV series and related activities. [18]

Netflix's philosophy while gathering and analyzing their customers' data is based on three key facts [19]:

- Data must be available, easily discovered and easy to process by anyone;
- Visualization should make it easier to explain, no matter the quantity of data;
- The more time is required to discover it, the lower is the value of data.

Netflix is building their user interface visual style based on gathered data, colors and tones used for presenting different content, choice of content displayed on top and their grouping, as well as their innovative recommender system that suggests content to their users, not primarily based on genre, type or duration of video content, but searches and suggests based on more subtle inner characteristics of video content receiving more attention from the customer.

Subscriber base is enabling Netflix's growth and continuous offering new, rich and interesting content. If people cannot find movies to watch, they would soon cancel their subscription. For Netflix, it is of great importance to ensure having precise algorithm providing them with required data, suggesting what next should occupy customer's attention, without letting customers looking for fun in another place. The fact that more than 75% of customers' activity is based on these suggestions [20], is indicative enough of Netlix's success in occupying their customer's attention. Netflix user base is constantly growing during the years, owing to their manner of work and philosophy for providing ever-growing content for every interest group. On Fig. 2 we can see Netflix subscribers [21] at the end of each of last seven years. We have chosen 2011 as starting year to coincide with Netflix spread outside of USA. In 2016 they are available in 190 countries all over the world. During these seven years, the number of subscribers has 400 percent growth rate, with fairly stable growth rate of 25 – 40 percent annual growth. At the end of 2011, they have 23.53 million subscribers, and in 2017 almost 118 million subscribers. In order to be able to keep up with the tempo they are creating and adding new content, they must continue increasing their subscriber numbers in the following years. [22]

In the early 2000s, advance of computer technology, Internet connections and video compression applications, is enabling streaming of big video files over Internet. This development is unfolding in ideal time for Netflix, them being involved in streaming technologies development since 2000, and being prepared to replace renting physical DVDs with a streaming service. That shift starts in 2007, and soon TV series are becoming integral part of their business model. Netflix is offering their customers [23] non-linear television experience, not limiting their product within dedicated timeframe and to a static screen using complicated remote controls. Internet, with its personalization, availability on demand and on any type of

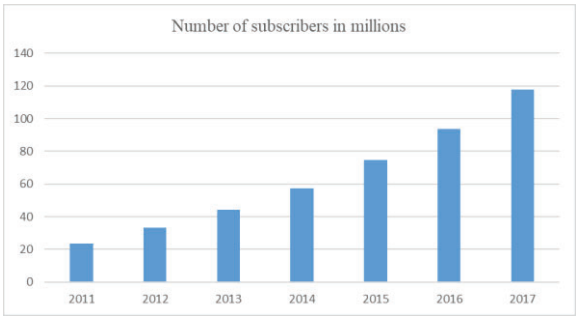


Fig. 2. Number of Netflix subscribers at the end of each of last seven years.

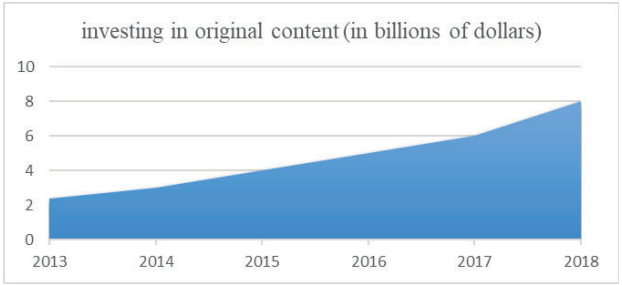


Fig. 3. Netflix investments in original programming, in billions of dollars.

screen is successfully replacing linear television. Today, even leading television houses are offering on-demand programs through applications running on phones or smart TVs. These applications are enabling viewers to catch up with missed programs and binge watch more than one episode in a row. Television channels that understood and accepted this change have become more valuable, while those that stayed behind are bound to lose their viewers and revenues.

Netflix understands that peoples' tastes are very divergent, depending on their age, country, even in the same demographic group. This is the reason their content offer is broad, starting from action blockbusters, Turkish soap operas, animated movies, science fiction, children' programs, and other genres. Before, great part of their offer was represented by movies and TV series leased for a limited period of time, while in the last few years, they are constantly investing in their own original programming.

On Fig. 3 we can observe this Netflix's trend in investing in their own programming, starting in 2013 with $2.4 billion, and continuing growing to $3 billion in 2014, $4 billion in 2015, little less than $5 billion in 2016, reaching $6 billion in 2017, and predicted $8 billion for this year.

While analyzing the product approach Netflix has taken over the years, we can conclude that, starting from their beginning as an innovative video content provider (renting DVDs and their sending via mail), then as a pioneer of streaming service while building their customer base and digital library, and today, as one the biggest creators of original video content, this company is constantly investing in providing better and more creative product, building their brand, and becoming most recognizable name in digital on-demand entertainment industry. Everyone that looks for this kind of product instinctively turns his attention to this company, which becomes synonym for quality and encompassing video content, similar to Google becoming synonym for internet content search or Apple for quality personal phones and computers.

D. Innovative Promotion

Promotional activities of a company are helping promote their products and services, build their name and make it recognizable compared on the market. Although Netflix, as any other company, is not publicly announcing their marketing plan and activities to take, we can draw certain conclusions about their innovative promotional activities based on how they are approaching the offered content and how they increase their target audience. Following are some of the means Netflix uses [24] to promote their products:

1. Building user experience through autonomous products – Netflix does not pay great attention to traditional marketing, but takes an innovative approach, creating new, autonomous products to promote new movie or series. For example, for promotion of the second season of "Stranger Things", Netflix creates new experience using augmented reality, where users can interactively discover different "Easter eggs" from the series. [25]

2. Investing into original content – Netflix is investing so much into original content, itself presenting a promotional activity, with $6 billion invested in 2017 and planned investment of $8 billion in 2018. These numbers are speaking for the company's determination in investing in meeting customers' demands, compared to the competition.

3. Gathered user data – today data about user habits are one of the best tools for predicting their behavior, given the premise that they used right metrics and in the right moment. Netflix is using data gathered to improve user experience, ultimately inspiring the greatest part of their marketing activities. For example, while promoting "House of Cards", Netflix prepares tens different promotional videos, serving them [26] to their subscribers depending on previous user behavior.

4. Viral marketing – today one of the most frequently used marketing terms is "going viral", but Netflix succeeds in doing it in such a way that makes content easier to share, all the while keeping their name as a consistent and honest brand. One such example is promotion of the second season of "Daredevil" [27], shown on Fig. 5, twitting messages from Netflix to people following their twitter account.

5. Not rolling out everything at once – one of Netflix trademarks is giving their customers a chance to binge watch all episodes of whole season or of entire series at once. This trend continues when the company starts producing their own content. Subscribers are satisfied to be able to watch series with their own pace, but this approach shortens their social media presence. For example, media presence of HBO's show "Game of Thrones", broadcasted on weekly basis, is constantly reigniting for the duration of dozens of weeks. Netflix is tackling this problem with issuing additional content and information about their shows few weeks after their initial publishing. An example of this is "Daredevil" where they eleven days after first season publishing are announcing renewing of the show for a second season, and few weeks later announce information about new characters in the second season.

6. Keeping users interested – cable television usually commit their users with long-term contracts, with services that usually take more time and effort to establish, and much harder to get away from. On the other hand, internet television, such as Netflix, do not require special devices or connections, and the user can start using or cancel on a monthly basis. There, the company must invest more effort into staying fresh and relevant for their subscribers. Netflix is doing this by contacting their subscribers about value to deliver in the next month. These messages are usually sent during week before the last in a month, making user's decision about renewing the subscription for one more month easier.

Activities we stated are just a part of the means Netflix uses to present themselves before their customers, and they are varying depending on the market, target group, and even the taste and habits of individual users. Means of promotion are not always transparent to the audience, but they are using creating methods to reach as much population as possible and deliver their message.

IV. Conclusion

In this paper we have presented Netflix, one of the most innovative companies in the last two decades. Since the beginning, the company is distinguished by its own identity, offering diversity and with innovative ways of accepting and putting in use new technologies.

While analyzing the company, we have established that they are introducing innovation in every step of the process of creation and delivery of their products. At first, through online rental of DVDs and their delivery via mail, they are representing interesting alternative to "mortar and brick" shops, enabling users to rent movies from the conformity of their own homes for a fixed monthly fee. With increased internet speed and emerging of new streaming technologies, Netflix is successfully shifting to this field, all the while building up their content library. In last few years, the company starts to create their own content, successfully fighting biggest traditional TV houses. Netflix's innovations are not exclusive to products only. They are implementing innovative processes in video content delivery, implement new strategies for marketing promotion, and are able to reorganize the company depending on current challenges in front of them. All of these innovations have great impact on their competitors, as they are compelled to follow Netflix's pace, or be left behind and forgotten.

Netflix's product go along with their customers, with the company trying and succeeding in creating value for different kinds of user groups, and creating its own voice and brand for every user individually. Their customer base is constantly growing throughout the years, and reaching almost 118 million subscribers all over the world at the end of 2017.

Innovative promotional activities on Internet and social media are helping the company promote their own brand, increasing awareness in their current and target customers, and building up good reputation. Some of the activities taken include: building user experience with the help of accompanying products, investing in new and original content through own production and coproduction with global production houses, improving user experience using gathered user data, viral marketing, keeping interest in certain content, as well as sustaining customer attention and satisfaction.

The analysis of Netflix as a modern innovative company helps us draw a myriad of useful information about their activities, but we must emphasize constant risks they take. The industry they compete requires constant improvement, taking risky steps to stay ahead of the curve, and being able to learn, change and improve.

References

[1] Company profile – Netflix, https://www.fastcompany.com/company/netflix. As seen on 1.2.2018

[2] Netflix timeline - A brief history of the company that revolutionized watching of movies and TV shows (https://media.netflix.com/en/about-netflix). As seen on 1.2.2018.

[3] Financial overview, https://ir.netflix.com/static-files/0c060a3f-d903-4eb9-bde6-bf3e58761712, 22.1.2018.

[4] Philip Kotler, Gary Armstrong, "Principles of Marketing", 14th edition, Prentice Hall, 2012.

[5] Oslo manual, "Guidelines for collecting and interpreting innovation data," https://www.oecd-ilibrary.org/docserver/9789264013100-en.pdf?expires=1523990471&id=id&accname=guest&checksum=045EAD5C8038C6EBCBD24117D4B9037F. As seen on 6.2.2018.

[6] Netflix company culture. https://jobs.netflix.com/culture.

[7] Patty McCord, How Netflix reinvented HR, https://hbr.org/2014/01/how-netflix-reinvented-hr. As seen on 3.2.2018

[8] Megan O'Neill, How Netflix bankrupted and destroyed Blockbuster, http://www.businessinsider.com/how-netflix-bankrupted-and-destroyed-blockbuster-infographic-2011-3. As seen on 8.2.2018

[9] Chunka Mui, How Netflix innovates and wins, https://www.forbes.com/sites/chunkamui/2011/03/17/how-netflix-innovates-and-wins/#55f3165f61f3. Forbes Magazine Online, 17.3.2011

[10] Clayton M. ChristensenMichael E. Raynor, and Rory McDonald, What is disruptive innovation, https://hbr.org/2015/12/what-is-disruptive-innovation%20, Harward Business Review, december 2015 issue.

[11] Toke Kruse, Disruptive innovation: how Netflix revolutionised the video market, http://www.seiercapital.com/disruptive-innovation-how-netflix-revolutionised-the-video-market/, 2.8.2016

[12] Why Netflix is one of the most innovative companies of 2017, https://www.fastcompany.com/3067462/why-netflix-is-one-of-the-most-innovative-companies-of-2017. FastCompany, 13.2.2017

[13] Ken Auletta, How Netflix killed Blockbuster, https://www.youtube.com/watch?time_continue=1&v=QsSsMHV0Kt8, BigThink series, published on 3.4.2014.

[14] [14] Dan Moskowitz, Who are Netflix's main competitors?, https://www.investopedia.com/articles/markets/051215/who-are-netflixs-main-competitors-nflx.asp, Investopedia, 22.1.2018

[15] https://www.cnbc.com/2017/05/31/netflix-spending-6-billion-on-content-in-2017-ceo-reed-hastings.html, published on 31.5.2017

[16] https://www.nytimes.com/2017/10/16/business/media/netflix-earnings.html, The New York Times, 16.10.2017

[17] Robert S., Netflix's uncertain future, https://digit.hbs.org/submission/netflixs-uncertain-future/, Harvard Business School.

[18] Daniel Newman, Improving customer experience through customer data, https://www.forbes.com/sites/danielnewman/2017/04/04/improving-customer-experience-through-customer-data/#1b1c4e374e64, as seen on 9.2.2018

[19] Phil Simon, Big data lessons from Netflix, https://www.wired.com/insights/2014/03/big-data-lessons-netflix/

[20] Tom Vanderbilt, The science behind the Netflix algorithms that decide what You'll watch next, https://www.wired.com/2013/08/qq_netflix-algorithm/, Wired Magazine, 8.7.2013

[21] https://www.statista.com/statistics/250934/quarterly-number-of-netflix-streaming-subscribers-worldwide/, as seen on 9.2.2018

[22] Amanda Lotz, The unique strategy Netflix deployed to reach 90 million worldwide subscribers, http://theconversation.com/the-unique-strategy-netflix-deployed-to-reach-90-million-worldwide-subscribers-74885, as seen on 9.2.2018

[23] Netflix's view: internet entertainment is replacing linear TV, https://ir.netflix.com/netflixs-view-internet-tv-replacing-linear-tv, as seen on 11.2.2018

[24] Dennis Williams, 4 content marketing lessons to learn from Netflix, https://www.entrepreneur.com/article/294050, as seen on 12.2.2018

[25] http://www.businessinsider.com/netflix-is-running-a-first-of-its-kind-snapchat-world-lens-for-stranger-things-2017-10, as seen on 12.2.2018

[26] Ryan Lawler, How Netflix will use big data to push House of Cards, https://gigaom.com/2011/03/18/netflix-big-data/, as seen on 12.2.2018

[27] Chris Kerns, Streaming social: What marketers can learn from Netflix's social strategy, https://marketingland.com/streaming-social-marketers-can-learn-netflixs-social-strategy-171659, Marketing Land, 11.4.2016

Acceptance of the m-commerce among the citizens of Republic of Macedonia

NikolaSkenderov
University for Information Science and Technology
Ohrid
nikola.skenderov@cns.uist.edu.mk

Elizabeta Maneska
University for Information Science and Technology
Ohrid
elizabeta.maneska@mir.uist.edu.mk

Vlatko Grujoski
University for Information Science and Technology
Ohrid
vlatko.grujoski@cse.uist.edu.mk

***Abstract*–Due to the fast development of technology in the recent years, there has been a constant growth of the e-commerce and the m-commerce in the world and in Republic of Macedonia. There are some official statistics about the state of the e-commerce in Macedonia from the State Statistical Office and several researchers but the state of the m-commerce in Republic of Macedonia is limited or not covered at all. The aim of this paper is to show the factors that influence the Macedonian customers' behavior towards using the mobile commerce in Macedonia. The study is based on the m-commerce in Macedonia. More specifically, to which extent the m-commerce is accepted by the Macedonian citizens and what are the factors that influence its development and its acceptance. Also, the satisfaction of the customers after using the m-commerce will be measured, together with its advantages and disadvantages. The result of this research should serve as a further analysis and an improvement in the field of online shopping using mobile devices.**

Keywords— e-Commerce, m-commerce, customer satisfaction

I. Introduction

With all the advancements in technology in the recent years, the electronic commerce has experienced constant growth in the world and in Republic of Macedonia. An Electronic commerce is the process of selling goods and services on the Internet. The beginnings of the electronic commerce date from 1993, just two years after the official launch of the Internet, and since then it has increased exponentially together with the growth of Internet users. Today almost 50% of the world population is connected to the Internet. In Macedonia, the number is even larger with 73.6% of the citizens having access to the Internet. These numbers constantly increase and this is very beneficial for the businesses that sell their products and services on the Internet in order to expand their market presence and increase their profit.

Mobile commerce, also known as m-commerce, is the use of wireless handheld devices, such as cellular phones used to buy or sell goods or services online. According to Clarke, m-commerce refers to any transaction with monetary value that is conducted via a mobile network [1]. This concept is relatively new because at the beginning the commercial transactions were conducted via wired devices such as personal computers. With the development of wireless networking and with the innovations of laptops and cellular phones, the electronic commerce has experienced additional expansion and it has eased the process of making online transactions. So, customers can buy products/services anywhere besides their homes.

M-commerce dates from 1997. At the beginning, it was limited to buying ringtones and paying services through text messages.

The fact that the m-commerce will grow more in the future is the projected statistics from Statista [2]. According to the statistics the revenues from the m-commerce are projected to rise from 96.34 billion U.S. dollars in 2015 to 693 billion U.S. dollars in 2019. According to Statista, in the third quarter of 2017, 58% of the population in South Korea ordered products/services via smartphones in the past month. Up next is Thailand with 52% and United Arab Emirates and Taiwan with 45%. [3]

According to a research from RetailMeNot in 2015, Europeans spent about 45 billion euros via mobile devices. It was an increase of 88% compared to the previous year. In 2015, PayPal did research on the m-commerce in Europe. The results showed that 53% of the online shoppers in Turkey bought a product online via a smartphone before the months the research was done. 21% of the citizens of France and Sweden and 17% of the citizens of the Netherlands also did online shopping via a smartphone. [4]

According to a survey carried by Grouper-MK [5], Macedonian citizens spent 85 million euros on online shopping in 2016 year.

According to the State Statistical Office [6] in 2017, 12.8% of the citizens in the Republic of Macedonia made an order/bought goods/services on the Internet in the last 3 months, while 19.5% of them did that during the last 12 months. Almost one-third of the citizens in the Republic of Macedonia use mobile phone networks for connecting to the Internet.

This statistics promise future increase of the electronic commerce and the m-commerce in Republic of Macedonia.

This paper is divided in several parts: First, a literature review is provided and then an analysis of the survey. The survey consists of a representative sample of age and gender of 163 respondents. At the end of the paper, a conclusion and some recommendations about the next steps that should be taken for a further development of the m-commerce in Republic of Macedonia are given.

II. Literature review

Nowadays, the online shopping is considered to be a faster way of buying something you need. With its fast growing, the m-commerce has become a very interesting topic for the researchers all over the world. We can say that the research articles for m-commerce increased significantly since 2000[7]. Topics of customer satisfaction, security issues and the acceptance of the m-commerce are mostly published by the researches.

According to Europe M-Commerce 2017[8], nearly one-third of the European Internet users completed purchases via a mobile in 2016. In specific countries like China, South Korea, and India, the M-Commerce's share has already exceeded one-half and continues growing.

Several researches have been done for the trends of the e-commerce in Republic of Macedonia. Anita Ciunova-Shuleska Marija Grishin, Nikolina Palamidovska (2011) [9] targeted the attitude of young people towards the online shopping in Republic of Macedonia. TeutaVeseli and Remzije Rakipi(2015) [10] have investigated the relationship between the Internet and the customer satisfaction in Republic of Macedonia. The both researchers made a survey conducting a specific target group and they compared the results with the data from the State Statistical Office.

Other research from "Insider ID" [11] says that these 70% Macedonians are afraid of the quality of the online products and they are also afraid that their personal data or credit cards might be used for bad purposes.

Blaženka Knežević, Mia Delić Ines Dužević (2016)[12], from the neighboring countries have done an analysis of the products/services the young people in Croatia buy via mobile phones, and they also pointed their motivation of buying the products/services as well as the advantages and disadvantages of shopping via a mobile device.

When it comes to researchers from the other parts of the world, Hongjiang Xu (2014) [13] has done an analysis of the usage and the acceptance of the m-commerce among the citizens in China, which is the world's largest market for mobile devices and one of the largest in terms of percentage of online transactions done by mobile devices.

As it was mentioned above, the m-commerce has attracted the attention of researches all over the world, but this is not the case with Macedonia. The m-commerce has very little coverage here and there is still not enough information about it.

III. Research Methodology

An online questionnaire consisting of 19 questions is created as a Google Forms link. The questions are based on the questionnaire conducted by Koen Siers[14] with some additional questions about the demographics of the respondents similar to the data of the State Statistics Office in Macedonia. The first section of the questionnaire consists of questions about the demographics of the respondents in terms of interest of the study, such as age groups, gender, level of education, usage of smartphones and mobile networks. The second part of the questionnaire consists questions concerning the correspondents' opinions about the m-commerce, its advantages and disadvantages, their online shopping habits and predictions about the progress of the m-commerce in future. The questionnaire was administered through the social network Facebook in the period of March 2018 to around thousand potential respondents.

IV. Results

The first part of the analysis consists of descriptive statistics and analysis of the results. The questionnaire is shared from the Facebook accounts of the authors, and is reposted from some of the respondents. From about thousands of potential correspondents on the social network Facebook 163 responses are obtained. This is a small size of data compared to the total population in Macedonia however, it can be used for making conclusions about the state of the m-commerce in Macedonia.

53, 4% of the respondents are females, and 46.6% of them are males. According to the age structure 69, 6% are aged 25-35, 21.1% are aged 18-24, 8.1% are aged 36-50 and 1.2% are 50+. The dominance of the age category 25-35 in this survey is of no surprise since the authors of this paper together with most of their friends on the social network belong to this age category. Also this percentages correspond to the data from the State Statistics Office of Republic of Macedonia (SSO) where it is stated that 94.3% of the citizens aged 15-24 are connected on the Internet on a daily basis, while 70.4% of the citizens aged 25-54 use the Internet daily. The group of higher education is dominant with 90.6% of the responses and this is closely related to the dominance of the age category. The group of secondary education consists only 8.8% of the total responses. Similar dominance is noticed in the category of labor force with 72.7% of the respondents being employed and students making 14.3% of the total number of respondents.

The next set of questions corresponds to the main theme of this survey, the m-commerce. According to the results, 98.8% of the respondents own a smartphone and 94.4% of them use the network from their mobile provider to connect to the Internet. This corresponds to the relatively high percentage of mobile network usage among the citizens in Macedonia (64.6% according to the SSO).

82.5% of the respondents search for products and services through their smartphone devices. This number is high due to the high percentage of smartphone users together with the dominance of the categories mentioned above. 47.1% of the respondents which answered positively on the previous question were interested only in knowing details about the product/service, while only 34.1% actually bought the product from their smartphone. This relatively low number corresponds to the low number of online transactions among the citizens in Macedonia in 2017. Only 24.8% of the total population made a transaction within the last 12 months, from all the types of devices.

When it comes to the type of products/services the respondents buy, 38.9% responded that they buy electronic equipment, 33.3% responded that they buy clothes, and 9.7% responded that they buy household goods. This habits correspond greatly to the data from SSO where 64.1% of the ordered goods/services are clothes, 19.5% are sports goods and electronic equipment and 13.1% are household goods.

In terms of customer satisfaction, on a scale from 1 to 5 according to Likert, (1-unsatisfied, 5 the most satisfied) 35% of the respondents that ordered a product/service via a smartphone rated the transaction high with 4 stars. 33.3% of them rated the transaction with 5 stars, 23.8% gave a rating of 3 stars, 5.6% gave a rating of 2 stars and only 1.6% gave a 1 star rating. These results show a satisfactory outcome when it comes to making a transaction via a smartphone.

In terms of advantages, 47.1% of the respondents answered that the biggest advantage of shopping via a smartphone is that it is always on hand. Also, there are identical responses for the next two categories easy and fast with 25.7% and 23.5% respectively. These 3 advantages are dominant in the process of buying via a smartphone. 36.8% of the respondents stated handiness as the main advantage of a buying process. 24.5 % of them claimed that the buying process is easy and 16.1% of them claimed the process is fast. In terms of disadvantages 9% of the respondents stated that the buying process via a smartphone is clumsy and 7.8% of them found the buying process via a smartphone difficult.

When it comes to the comparison of shopping via a smartphone to shopping via a pc/laptop, 65.6% of the respondents stated that more products can be bought using a laptop/pc, while 34.4% of the respondents opted the smartphone as an option. This relatively high percentage in favor of shopping via a laptop/pc is related to the next question in which 45.5% of the respondents pointed the screen size of the laptop/pc as the main reason of the high percentage. On the other hand, the smartphone as a handier device is favored by 37%, while only 9.1% of them choose the security of the pc/laptop.

In addition, 82.4 % of the respondents that have not made a transaction using a smartphone yet, will mostly do that in future. 43.5% of the respondents said that they will start buying products via a smartphone if the web shops creates mobile design web sites or apps, while 35.1% will do it if the security of the smartphones is improved. According to the official statistics from SSO and NBRM, the fear in terms of online shopping security is generally present among the citizens in Macedonia.

As to the progress of the m-commerce in the future, 62.8% of the respondents think that there will be more mobile apps created and the web sites will be improved to suite the mobile devices which will boost the online shopping on a higher level. 27.6% of them pointed the smartphones as a beginning point of the creation of new technologies. 9.6% were skeptical and thought that there will be no improvements in future. The differences in the results of our survey compared to the data from the State Statistical Office described above are summarized in the following bar plots:

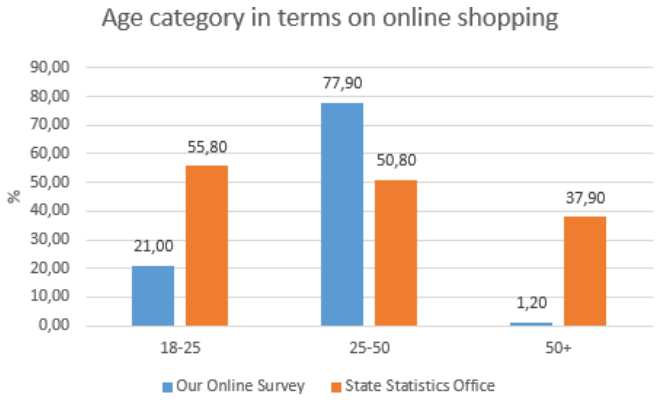


Fig.1 Age category in terms of online shopping from our survey vs the data of the State Statistic Office.

Fig.2 Types of goods/services people ordered online from our survey vs the data of the State Statistic Office.

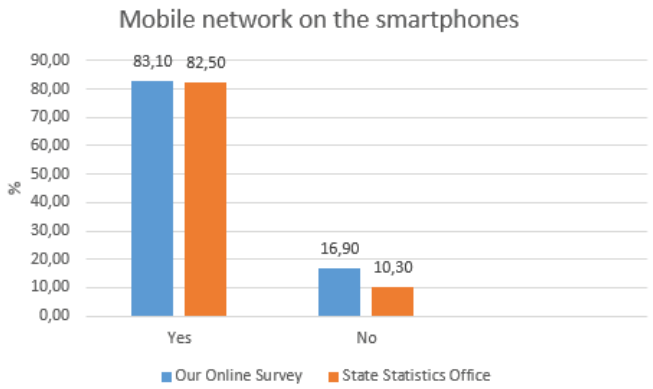


Fig.3 Mobile network on the smartphones from our survey vs the data of the State Statistic Office.

V. Descriptive statistics

For obtaining a relationship between the responses in the questionnaire, inferential statistics is conducted using the programming language R in R Studio. The results from the questionnaire are firstly summarized in an Excel file. Because the mostly used file format for importing data in R studio is CSV (Comma-Separated Value), separate CSV files are created. For the data analysis, the frequency of answers in terms of gender and age category are of particular interest. For that purpose each CSV file consists of 2 columns: the first one being either Gender or Age category and one of the remaining questions related to the topic of m-commerce is in the second column.

The first pairing of the answers is between questions "What is your Gender?" and the question "If you order on the Internet, what kind of goods / service do you buy?" The answers of these 2 questions are extracted from the summarized excel file and they are inserted in a separate blank Excel Worksheet. Then, the results are filtered and any row which has at least one blank response is removed. The Worksheet is saved as CSV file and it is read into the R Studio.
The null hypothesis is formed on the basis of the variables Gender and Kind of goods/service from the both questions:

H_0: "The kinds of goods/services the customer buys is not related to the gender".

To test the validity of the Null hypothesis, because the two variables Gender and Kind of good/service are categorical a Chi-squared test is performed in the R studio.

After the CSV file is read in the studio, a contingency table is created based on the CSV file in order to form the cross table which will consist of the distribution of Kind of good/service across the Gender category.

After creating the contingency table, the next step is to perform the Chi-Squared Test of the obtained table.

Chi-square test examines whether the rows and columns of the contingency table are statistically significantly associated. For calculation of the chi square value an expected value is calculated with the following formula based on the data in the contingency table.

$$e = \frac{row.sum \ * \ col.sum}{grand.total} \quad (1)$$

The Chi-square statistic is calculated as follows:

$$x^2 = \sum \frac{(o-e)^2}{e} \quad (2)$$

- o is the observed value
- e is the expected value

This calculated Chi-square statistic is compared to the critical value (obtained from the statistical tables)
with $df = (r-1)(c-1)$ degrees of freedom and $p = 0.05$.

- r is the number of rows in the contingency table
- c is the number of columns in the contingency table

If the calculated Chi-square statistic is greater than the critical value, then the row and the column variables are not independent of each other. This implies that they are significantly associated.

The results of the chi-square test that are obtained in R using the function chisq.test() on the contingency table are the following:

X-squared = 29.681, $df = 4$, p-value = 5.684e-06

Because the p value has less than 0.05 significance level we reject the null hypothesis and state that there is a relationship between the gender category and the types of products/services respondents buy via a smartphone.

Because of the confirmed association between the two variables in the chi square test we have calculated the Cramer's V to obtain the strength of the association. The Cramer's V value is calculated by the following formula:

$$V = \sqrt{\frac{\varphi^2}{\min(k-1, r-1)}} = \sqrt{\frac{X^2/n}{\min(k-1, r-1)}} \quad (3)$$

where:
φ- is the phi coefficient
x^2- is the calculated chi square statistics
n- is the grand total of observation
k – is the number of columns
r- is the number of rows

In R we use the function cramersV()to calculate the value and the obtained value is 0.481504 which indicates a medium strength of the relationship between the gender category and the types of products/services.

We have plotted the contingency table in r to visualize the results. We have used the library gplots to plot the table with the help of the function baloonplot() The rows represent the gender and the column represents types of products/services
The boxplot is represented in the following figure:

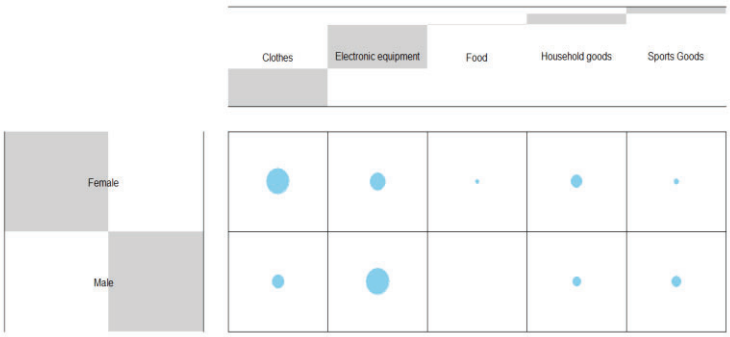


Fig. 4: Boxplot of gender vs product / service type

From the plot we can observe the relationship between the two categories and we can conclude that males shop more electronic equipment and Sports goods while females shop more clothes and household goods.
To find relationships between other variables we have conducted the chi square test of the answers of the questions "What is your age" and the question "What do you use your smartphone for".
The results of the chi-square test are the following:

X-squared = 3.0522, *df*= 9, *p*-value = 0.9622

The p value is higher than 0.05 significance level, so we accept the null hypothesis and we state that there is a independence between the age category and the usage of smartphones.

The entire code in the R studio together with the explanation of the steps is in the following lines:

```
#read the CSV file
res<-
read.csv(file="C:/Users/Nikola/Desktop/trening/1_10.csv",
header=T, sep=";")

#obtain the names of the columns of the variables
names(res)

#creating and displaying the contingency table
tbl<- table(res$Gender, res$Buy)
tbl

#plot the contingency table
library("gplots")
balloonplot(t(tbl), main ="usage", xlab ="", ylab="",
label = FALSE, show.margins = FALSE)

#calculate the chi-square test and display the results
CHI<-chisq.test(tbl, correct=T)
CHI

#calculate the Cramer's V value
library("lsr")
cramersV(tbl)
```

We have tried to test the relationship between the age category with the other questions, and the gender category with the other questions, but in all of the results the obtained p value was higher than the 0.05 significance level. This signifies independence between the variables. This is due to the small sample size of 163 responses, the uneven distribution in terms of age, level of education and labor force. A similar research can be done in future but with a larger sample size, a different structure of the questionnaire and an equal distribution in terms of demographics.

VI. Conclusion

The m-commerce is in a process of steady progress in the entire world and in Macedonia. The data from our research confirms that the citizens of the Republic of Macedonia are afraid to shop online and this confirmation is supported by the data from the State Statistics Office in Macedonia and some similar studies.

The citizens of the Republic of Macedonia spend a lot of time using their smartphones on a daily basis. They have mobile networks on them, but only a small percentage of the citizens uses them for online shopping. However, the majority of them visit e-shops or search for product/service details via their smartphones.

A large number of the citizens think that the m-commerce will eventually grow with responsive design of the online shops and creation of several apps.

The statistical analysis showed medium relationship between the category gender and the category types of goods/services. The subsequent analyses showed little or no relationship between the other categories.

References

[1] I. Clarke III, "Emerging value propositions for m-commerce", Journal of Business Strategies, vol. 18, no. 2, p .133 – 148. (2001)
[2] Statista, "Transaction value of global m-commerce sales from 2014 to 2019(in billion U.S dollars)", 2018. [Online]. Available: https://www.statista.com/statistics/557951/mobile-commerce-transaction-value-worldwide/. Accessed [1-Mar-2018].
[3]Statista, "Share of population who bought something online via phone in the past month as of 3rd quarter 2017, by country", 2018. [Online]. Available: https://www.statista.com/statistics/280134/online-smartphone-purchases-in-selected-countries/. Accessed [1- Mar- 2018].
[4]Ecommerce News Europe, "Mobile commerce in Europe", 2016. [Online]. Available: https://ecommercenews.eu/mobile-commerce-europe/. Accessed [2- Mar - 2018].

[5] Nina Angelovska, “E-commerce grows 36% in 2016 [Infographic by Grouper]”. 11.04.2017. [Online]. Available: https://blog.grouper.mk/kakva-e-sostojbata-so-e-trgovijata-vo-makedonija/. Accessed [26 – Feb - 2018].
[6] Drzhaven Zavod Za Statistika, “Usage of informatics-communication technologies in household and individuals, 2017”. 31.10.2018. [Online]. Available: https://www.stat.gov.mk/PrikaziSoopstenie.aspx?rbrtxt=77. Accessed: [15 – Jan - 2018].
[7] E.W.T. Ngai, A. Gunasekaran, “A review for mobile commerce research and applications”. Decision Support Systems, vol.43, no.1, 3-15. 2007
[8] yStats GmbH & Co.KG, ‘Europe M-Commerce 2017’, 2017.[Online]. Available:https://www.researchandmarkets.com/research/ctclxs/europe_mcommerce. Accessed [1-Mar-2018].
[9]A. Ciunova-Shuleska, M. Grishin, N. Palamidovska, “Assessing young adults’ attitudes toward online shopping in the Republic of Macedonia.”. Ekonomski Pregled. vol.62, p.752-772. 2011
[10] T.Veseli, R. Rakipi,”The Relationship between the attributes of the internet and consumers' satisfaction: A study of e-commerce in Macedonia” 4th REDETE Conference (Research Economic Development and Entrepreneurship in Transitional Economics). Graz, Austria 2015.
[11]Sergej Zafiroski, “Only 3 % of Macedonians are ready to shop online”. 23.11.2017. [Online]. Available: https://www.insider.mk/allwebconference/. Accessed [6- Feb - 2018].
[12] B. Knežević, M. Delić, I. Dužević, “Customer satisfaction and loyalty factors of Mobile Commerce among young retail customers in Croatia”, Revista Eletrônica Gestão & Sociedade, vol.10, no.27, p. 1459-1476, 2016.
[13] H. Xu, “The Effect of Perceived Security on Consumers' Intent to Use: Satisfaction and Loyalty to M-Commerce in China”, Journal of Electronic Commerce in Organizations, vol.11, no .4, p. 37-51, 2013.
[14]K. Siers, “Improvement of acceptance of mobile commerce” 2017. [Online]. Available: http: //essay.utwente.nl/72687/1/Siers_BA_BMS.pdf. Accessed: [27-Jan-2018]

ABSTRACTS

Anti-virus Engine Analysis using Deep Web Malware Data

Igor Mishkovski, Miroslav Mirchev, Milos Jovanovik
Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University
Skopje, R. Macedonia
igor.mishkovski@finki.ukim.mk

ABSTRACT

This template, modified in MS Word 2007 and saved as a "Word 97-2003 Document" for the PC, provides authors with most of the formatting specifications needed for preparing electronic versions of their papers. All standard paper components have been specified for three reasons: (1) ease of use when formatting individual papers, (2) automatic compliance to electronic requirements that facilitate the concurrent or later production of electronic products, and (3) conformity of style throughout a conference proceedings. Margins, column widths, line spacing, and type styles are built-in; examples of the type styles are provided throughout this document and are identified in italic type, within parentheses, following the example.

AntiVirus products and tools are essential in every business deployment connected to the Internet. Nowadays, with the increase in the number and diversity of malware on the Web, there are also more AntiVirus Tools (AVT) becoming available to protect users and/or companies from malware. However, the quarterly growth at around 12\% for known unique malware samples, according to the *Intel Security Group's McAfee Labs Threat Report: August 2015*, and the fact that some AntiVirus companies use same or significantly similar AntiVirus engines leave us in some way vulnerable to the existing security threats.

In this work, using graph analysis and visualization methods, on one hand we will empirically infer detection engine similarity and existing groupings and/or overlapping between them, while on the other hand we will infer which Anti-Virus Tools (AVTs) differentiate from other AVTs and have greater advantage in detecting malware compared to others.

Using the AVT responses to our malware file set we will optimize the combination of AVTs in order to obtain maximum detection rate (i.e. coverage). We strongly believe that this approach can be used by companies who want to implement multi-scanning approach on their email gateways.

Finally, another novelty in this work is that we relate the source of the malware, i.e. the domain name where the malware is found, with AVTs. In this way, we will show the detection rate of AVTs across domains in which potential malware resides. The results will imply that certain AVTs have more detection capabilities on specific domains, whereas, others might have detection rate spread across multiple domains. All the analysis will be done on a malware file set provided by F-Secure and the AVTs responses on this file set obtained using the Virus Total API.

Based on the dataset we measure the similarity between different AVTs in order to see if there are some clusters or communities that share similar ``reaction`` to a certain malware files. Thus, we construct the *similarity network* $G^1 = (V, E, W^1)$ in order to characterize the similarity between different AVTs based on the shared files which they labeled them as malwares. The node set V consists of AVTs which were reported by Virus Total and the undirected edges set E contains the links between the AVTs that have labeled at least one common malicious file, with an edge weight w_{ij}^1 being defined through Jaccardi score of the sets of malware files detected by the two AVTs i and j. Thus, here we define the similarity between V_i and V_j as the co-occurrence strength. Let us assume that Fi and F_j denote set of files, labeled as malware by V_i and V_j, then we can define the Jaccardi similarity measure as a co-occurrence strength as follows.

$$sim(V_i, V_i) = \frac{|F_i \cap F_j|}{|F_i \cup F_i|} = w_{ij}^1 = w_{ji}^1, \ , \quad (1)$$

where |F| indicates the size of the set F. The value of wij1 is between 0 and 1 (where "0" indicates no co-occurrence relationship between two AVTs and "1" indicates a full co-occurrence).

The results show high similarity between certain AVT in their malware detection. Some of the AVT groups that show high similarity are i) **BitDefender, F-Secure, Emsisoft, MicroWorld-eScan** and **Ad-Aware**; ii) **Arcabit, eTrust-InoculateIT, UNA** and **T3**. This results clearly show that there might exist grouping in sense of structural communities and/or clusters between different AVTs. This kind of clustering or grouping might be as a consequence of the fact that different AVTs are specialized for certain type of malwares (Trojans, Adwares, Exploits, Rootkits, etc.), or malwares written for a given platform (such as Win32,

OSX, Android, etc.) or simply due to the fact that some companies use engines from other AV companies, such as *F-Secure* and *BitDefender*, *AVWare* and *VIPRE*.

Keywords— malware; community detection; anti-virus engines; data science; multi-scanning approach

ACKNOWLEDGMENT

Authors gratefully acknowledge the CyberTrust research project and F-Secure for their support. I.M. work was partially financed by the Faculty of Computer Science and Engineering at the University 'Ss. Cyril and Methodius' as part of the project "*AVADEEP: Anti-virus analysis using Deep Web malware files*".

REFERENCES

[1] M. Lindorfer, M. Neugschwandtner, L. Weichselbaum, Y. Fratantonio, V. v. d. Veen, and C. Platzer, "Andrubis - 1,000,000 apps later: A view on current android malware behaviors," in Proceedings of the Third International Workshop on Building Analysis Datasets and Gathering Experience Returns for Security (BAD-GERS), 2014, pp. 3-17.

[2] M. K. Bergman, "White paper: the deep web: surfacing hidden value," Journal of electronic publishing, vol. 7, no. 1, 2001.

[3] A. Mohaisen and O. Alrawi, "Av-meter: An evaluation of antivirus scans and labels," in Proceedings of the 11th International Conference on Detection of Intrusions and Malware, and Vulnerability Assessment, ser. DIMVA '14. Springer International Publishing, 2014, pp. 112-131.

[4] "VirusTotal: Free service to analyze suspicious files and URLs," https://www.virustotal.com/en/, online; accessed 14 July 2016.

[5] I. Gashi, V. Stankovic, C. Leita, and O. Thonnard, "An experimental study of diversity with off -the-shelf antiVirus engines," in Proceedings of the 8th IEEE International Symposium on Network Computing and Applications, 2009.

[6] I. Gashi, B. Sobesto, V. Stankovic, and M. Cukier, "Does malware detection improve with diverse antivirus products? an empirical study," in Proceedings of the 32nd International Conference on Computer Safety, Reliability, and Security, ser. SAFECOMP '13. Springer Berlin Heidelberg, 2013, pp. 94-105.

[7] J. Canto, M. Dacier, E. Kirda, and C. Leita, "Large scale malware collection: lessons learned," in Proceedings of the 27th International Symposium on Reliable Distributed Systems, ser. SRDS '08, 2008.

[8] Q. Jerome, K. Allix, R. State, and T. Engel, "Using opcode-sequences to detect malicious android applications," in 2014 IEEE International Conference on Communications (ICC), 2014, pp. 914-919.

[9] M. Zheng, P. P. Lee, and J. C. Lui, "Adam: an automatic and extensible platform to stress test android anti-virus systems," in International Conference on Detection of Intrusions and Malware, and Vulnerability Assessment, Springer, 2012, pp. 82-101.

[10] F. Maggi, A. Bellini, G. Salvaneschi, and S. Zanero, "Finding non-trivial malware naming inconsistencies," in Proceedings of the 7th International Conference on Information Systems Security, ser. ICISS '11. Springer Berlin Heidelberg, 2011, pp. 144-159.

[11] A. Kantchelian, M. C. Tschantz, S. Afroz, B. Miller, V. Shankar, R. Bachwani, A. D. Joseph, and J. D. Tygar, "Better malware ground truth: Techniques for weighting anti-virus vendor labels," in Proceedings of the 8th ACM Workshop on Artificial Intelligence and Security, ser. AISec '15. ACM, 2015, pp. 45-56.

[12] A. Kantchelian, M. C. Tschantz, S. Afroz, B. Miller, V. Shankar, R. Bachwani, A. D. Joseph, and J. Tygar, "Better malware ground truth: Techniques for weighting anti-virus vendor labels," in Proceedings of the 8th ACM Workshop on Artificial Intelligence and Security. ACM, 2015, pp. 45-56.

[13] J. Chang, K. K. Venkatasubramanian, A. G. West, and I. Lee, "Analyzing and defending against web-based malware," ACM Computing Surveys (CSUR), vol. 45, no. 4, p. 49, 2013.

[14] H. S. S.L., "Virustotal public api."

[15] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," Journal of statistical mechanics: theory and experiment, vol. 2008, no. 10, p. P10008, 2008.

[16] H. T. T. Truong, E. Lagerspetz, P. Nurmi, A. J. Oliner, S. Tarkoma, N. Asokan, and S. Bhattacharya, "The company you keep: Mobile malware infection rates and inexpensive risk indicators," in Proceedings of the 23rd International Conference on World Wide Web, ser. WWW '14. ACM, 2014, pp. 39-50.

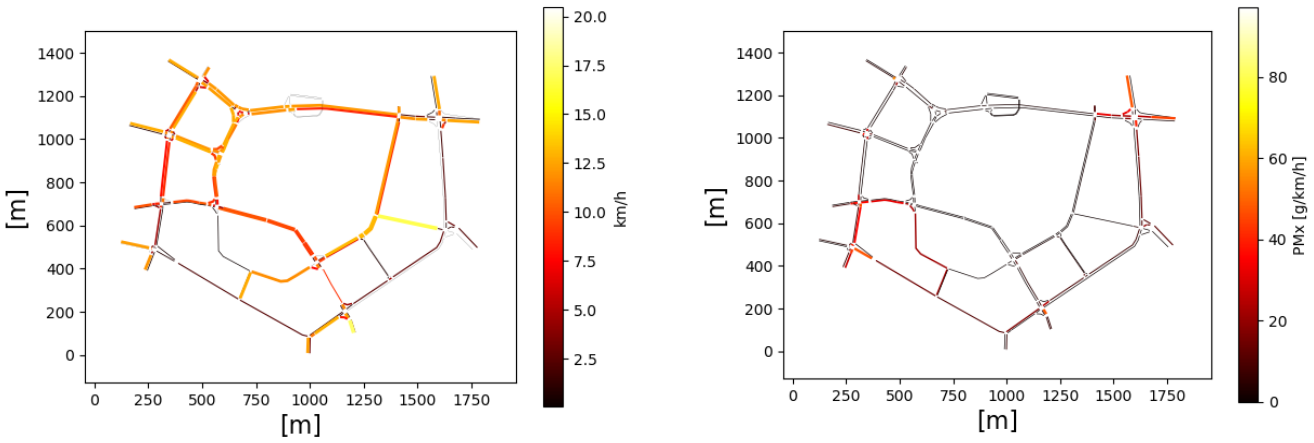
Urban traffic simulation for smarter and greener cities

Sasho Gramatikov, Igor Mishkovski and Milosh Jovanovik
Faculty of Computer Science and Engineering
Ss. Cyril and Methodius University
Skopje, Macedonia

ABSTRACT

Transportation is one of the main concerns of the big cities because vehicles are one of the top sources of air pollution. Despite this fact, many big cities follow a trend of growth of the number of vehicles that circulate on the road networks. In order to meet the demand, the authorities are obliged to increase the roads capacity or build new road infrastructure. However, this solution does not significantly contribute to pollution reduction and traffic jams, especially in the weekdays peak hours. One solution to respond to the travel demand and reduce the vehicle number is the concept of car sharing where more passengers with similar travel route share a vehicle.

The goal of our work is to estimate the effect that the car sharing service would have on the overall traffic in urban areas by means of simulation. As an initial point, our goal is to create a simulation environment based on real traffic data data that can be further used for analysis of the effect of different solutions. We use the SUMO, a traffic simulator which uses a network of interconnected edges, traffic lights and traffic demand to create a microscopic simulation of vehicle movement. We use some of the features of the simulator to download an urban area from Open Street Map and convert it into network compatible for traffic simulation. As a traffic demand we use data obtained data from inductive loops in the central area of the city of Skopje. The data contains a number of vehicles that pass each hour during 24 hours during the month of November 2017. We choose the data from the first Monday of the month since the traffic is most dense in weekdays. The data is used as input of a DFROUTER tool which generates possible routes for the network and flows of vehicles that meet the input induction-loops



(a) ***(b)***

Figure 1: (a) Out-bound and (b) in-bound traffic for different packet loss probabilities for segments with duration of 2 seconds

ACKNOWLEDGMENT

The author thanks the Faculty of computer science and engineering at the Ss. Cyril and Methodius University in Skopje, under the IACT “Influence of Autonomous Connected Transport in Urban Areas” project for financial support

Foreign Direct Investment Net Inflows: Determinants and Prediction Models

Ana Gjorgjevikj, Kostadin Mishev and Dimitar Trajanov
Faculty of Computer Science and Engineering
Ss. Cyril and Methodious University
Skopje, Macedonia
ana.gorgevic@gmail.com, {kostadin.mishev, dimitar.trajanov}@finki.ukim.mk

ABSTRACT

The OECD Benchmark Definition of Foreign Direct Investment - 4th edition (BMD4), defines foreign direct investment (FDI) as a cross-border investment taken in order to gain a lasting interest (at least 10% of the voting power) in an enterprise resident in an economy that is different from the economy of the investor. As the document indicates, foreign direct investments are usually motivated by the benefits from establishing a long-term relationship with the foreign affiliate, such as influencing its management or gaining access to its resident economy [1][2]. Due to the liberalization of the market access, as well as the rapid technological development, over the past decades the foreign direct investments have become an important form of international capital transfer and driver for creating stable and long-term relations between the countries [1]. Foreign direct investments have positively influenced many countries economic growth by providing them with a stable capital and knowledge inflow. FDI net inflows of an economy represent the net value of the inward direct investments made by non-resident investors to that economy. Modeling and reliably predicting FDI net inflows per country, as well as identifying the influencing factors, is very important for the multinational enterprises or any other enterprise with intention to invest abroad, as well as for the countries aspiring to attract foreign investors. Understanding what drives FDIs has also attracted the attention of the research community, resulting in proposal of variety of prediction models based on financial, economic or political stability determinants [3][4][5][6][7]. Although variety of theoretical models have been proposed, no consensus has been achieved over one model [8]. It has become evident that the importance of one determinant for a country deepens on the country's level of development, resulting in separate analysis of the more and less developed countries. Although the main types of indicators that influence one country's investment climate have been identified in the available literature on the subject and included in the proposed models, today's technology advancements allow more seamless integration of larger number of heterogeneous data sources and development of more advanced prediction models. Today we are able to accurately analyze textual data (e.g. news articles), infer its sentiment, recognize the entities it refers to and their relations, as well as perform many other kinds of information extraction tasks that would later enable integration of this extracted data with the available and well-structured financial and economic country data. Aiming at development of a model that accurately predicts the level of risk for investing in a country based on a large number of heterogeneous data sources (e.g. World Bank Open Data[1], OECD[2], GDELT[3] and other), as well as on the state-of-the-art machine learning techniques, we start by providing an overview of the available literature on the subject and discussing the most important determinants that attract foreign investors. We further make a comparison of the available sources of country financial, economic and political stability data, while discussing the challenges we have identified. By making an assumption that one country's FDI net inflows value is a sufficiently representative indicator for that country's attractiveness to foreign investors, we further analyze the data available for this indicator and its relation to other indicators.

Keywords—Foreign direct investment; FDI net inflows; Predictive modeling; Big data

REFERENCES

[1] *OECD Benchmark Definition of Foreign Direct Investment: 4th Edition (BMD4),* Organisation for Economic Cooperation and Development, 2008.

[2] *Balance of Payments and International Investment Position Manual: Sixth Edition (BPM6),* International Monetary Fund, 2009.

[3] J. H. Dunning and S. M. Lundan, *Multinational enterprises and the global economy*. Edward Elgar Publishing, 2008.

[4] Q. Li and A. Resnick, "Reversal of fortunes: Democratic institutions and foreign direct investment inflows to developing countries," *International organization*, vol. 57, no. 1, pp. 175–211, 2003.

[1] https://data.worldbank.org/

[2] https://data.oecd.org/

[3] https://www.gdeltproject.org/

[5] A. Bénassy- Quéré, M. Coupet, and T. Mayer, "Institutional determinants of foreign direct investment," *The World Economy*, vol. 30, no. 5, pp. 764–782, 2007.

[6] M. Busse and C. Hefeker, "Political risk, institutions and foreign direct investment," *European journal of political economy*, vol. 23, no. 2, pp. 397–415, 2007.

[7] K. Dellis, D. Sondermann, and I. Vansteenkiste, "Determinants of fdi inflows in advanced economies: Does the quality of economic structures matter?" 2017.

[8] I. Faeth, "Determinants of foreign direct investment–a tale of nine theoretical models," *Journal of Economic Surveys*, vol. 23, no. 1, pp. 165–196, 2009.

Big data-based platform for country economic stability

Kostadin Mishev, Ana Gjorgjevikj, Dimitar Trajanov
Faculty of Computer Science and Engineering
Ss. Cyril and Methodious University
Skopje, Macedonia
kostadin.mishev@finki.ukim.mk, ana.gorgevic@gmail.com, dimitar.trajanov@finki.ukim.mk

ABSTRACT

Political risks in each country are an important factor in determining the economic stability of a given country [1]. Considering the far-reaching effects of geopolitical instability, to the everyday influences of changing laws and social unrest, political factors are increasingly influencing investors' decisions whether the country is ideal for investment [2].

In recent years, due to the development of technology and the ability to store and process data in large quantities from different sources, there exist a lot of opportunities and platforms that provide building a statistical model that present the internal and external policies which affected the development of the country itself. Since these political factors are difficult to identify and even more complicated to measure, they are often left out by investors when making investment decisions in a country [3]. Analysts are unable to incorporate policy into their allocation strategies and risk management, relying on short-term reports and static indicators to make critical decisions in real time.

The idea behind the proposed tool for predicting the economic stability of a country is to identify appropriate macroeconomic and geopolitical indicators that have cascaded the country's economy over a longer time period, taking data from multiple sources and aligning them according to the time frame. There are multiple data sets that can be used as a series of data from the past to determine the characteristics how an appropriate prediction model will be built.

So far, we have identified several data sets that we will use in feature vector extraction. GDELT [4], Thomson Reuters News archive [5], IBM Watson Discovery News [6] represent textual news databases that are updated daily with additional information about the sentiment that is expressed in the content of the news. WorldBank, EuroSTAT, Quandl represent the largest database of financial data for each country individually, primarily taking the financial indicators: GDP per country, inflation, interest rate and unemployment.

The prediction model will be based on a machine learning technique. As an input, it will be used the sentiment from historical events, the state of the world stock market and a stock market, coupled with measures that show the macroeconomic and financial indicators for each country individually. Predictions will indicate the country investment index stability, considering the current values from the indicators used to train the model. To achieve that, it is necessary to build a robust software solution which can use previously trained neural network in combination with the latest data obtained for all parameters, to provide a virtual index for investment convenience. Due to the heterogeneity and data size, it will be used lambda-based architectures that implement the modern concept of data lakes. Such approach will provide reads, writes and indexing of the data coming from different sources to be realized in real-time by using advanced message brokers.

Keywords— Big data, computer modelling, economic stability of countries, machine learning, data acquisition framework, data ingestion flows, data lakes

REFERENCES

[1] *Bevan, Alan A., and Saul Estrin. "The determinants of foreign direct investment into European transition economies." Journal of comparative economics 32.4 (2004): 775-787.*

[2] *Jiménez, Alfredo. "Political risk as a determinant of Southern European FDI in neighboring developing countries." Emerging Markets Finance and Trade 47.4 (2011): 59-74.*

[3] Faeth, Isabel. "Determinants of foreign direct investment–a tale of nine theoretical models." Journal of Economic Surveys 23.1 (2009): 165-196.

[4] GDELT Project, https://www.gdeltproject.org/

[5] Thompson Reuters News Archive, https://www.thomsonreuters.com/en.html

[6] IBM Watson Discovery News, https://www.ibm.com/watson/services/discovery-news/

Human Activity Recognition Using Unobtrusive and Wearable Sensors

Gjorgji Madjarov, Dejan Gjorgjevikj
Faculty of Computer Science and Engineering
Ss. Cyril and Methodius University in Skopje
Skopje, Macedonia
gjorgji.madjarov@finki.ukim.mk, dejan.gjorgjevikj@finki.ukim.mk

ABSTRACT

The advances in electronics and computation and communication technologies increase the number and the different types of sensors that are used in everyday life. The number of different applications that use sensors increase every day, due to the low prices and the high availability of the different sensors. Unobtrusive human activity monitoring using extremely cheap and widely available sensors are the future for human activity recognition. It will support the extensive penetration of new applications in Ambient Assisted Living (AAL), Smart Homes (SH), Smart Cities (SC) and Health Monitoring (HM).

AAL, SH and HM have gained a lot of attention for the provision of enhanced quality of life within the home especially considering the worldwide increase of an aging population. In order to support AAL, SH and HM significant research has been directed toward recognizing different activities of a person using sensors. Most of the research in this area was directed towards using wearable motion sensors in a form of mobile phone, wristband or wearable passive RFID tags. Very little research in this area has been conducted using unobtrusive sensors like Pyroelectric infrared (PIR) and radar sensors.

The biggest challenges in these applications are the automatic processing and analyzing the large amounts of sensory data as well as building machine learning models for monitoring, detection, recognition and prediction of an activity, movement, state or an event. The aim of this research is to develop a system for automated human activity detection and monitoring using low-cost, unobtrusive PIR and radar sensors.

Our sensor node is composed of Arduino (Uno or nano) microcontroller, SD card module, real time clock module utilizing extremely accurate DS3231 RTC, modified microwave radar sensor RCWL-0516, modified passive InfraRed HC-SR501 (PIR) sensor. Both the RCWL-0516 and the HC-SR501 are extremely low cost sensors for general application mainly used for motion detection light switching. Although relaying on different type of signals (infrared emitted by warm objects as animals and humans in the case of the PIR sensor, and microwave reflected by metallic and water reach moving objects i.e. cars or humans) the processing of the raw signal obtained by the sensors is performed by the same (or two different but functionally almost identical) IC- IS0001 in the case of the PIR sensor and either BIS0001 or RCWL-9196 in the case of the RCWL-0516 radar sensor. These IC chips typically include Bi-directional level detector and several high input impedance operational amplifiers used for amplification and filtering that can be configured using outside passive components. The main task of this IC is to detect movement and output digital signal that can be used for switching different devices. However, the analog signal from the preprocessing and conditioning stages is available on one pin of the IC that we have patched on in order to use it not for just detecting presence or absence of motion but to try to distinguish between several different types of motion in order to utilize it for human activity recognition.

Also, a system for capturing the sensor data from common readily available cheap sensors around the Arduino platform is developed. The software for the microcontroller in the Arduino environment can measure several (up to 4 analog 10-bit, and 4 digital sensor inputs) with rates of up to 200 samples per second + the current temperature and logs the measurements on a SD card including a precise timestamp of the measurements so that the logs of several nodes can later be fused (joined) accurately aligned in time. All the measurements are interrupt driven utilizing an accurate clock signal from the RTC modules that are internally temperature compensated to avoid the instability of the microcontroller crystals. Since the nodes are independent and have no communication among each other the synchronization relays on the accuracy of the RTC included in each node. Therefore, before the measurement all RTC modules are synchronized to millisecond precision. To perform this, we have developed a setup around an ESP8266 microcontroller that can connect up to 4 RTC modules at once, can connect to internet timeserver, read the current time and synchronizes all the modules. All RTC modules incorporate a coin cell battery for time keeping even when not powered on.

Also, for this research a platform for measuring and capturing data from all sensors of Microsoft band 2 is available out of the box. It is used for pre-processing and annotation of the data collected from the unobtrusive PIR and radar sensors and also as a ground truth for the human activities.

The experiments in a control environment with volunteers are conducted and the collected data from the sensors are pre-processed and labelled for further analysis. The continue of this research will be designing and implementing machine learning based approaches that would allow automatic recognition and monitoring of human activity using the cheap unobtrusive sensors supported by the wearable motion sensors.

Keywords—human activity; recognition; sensors; wearable; unobtrusive

STUDENT PAPERS

Evaluating Techniques for Building a University Course Recommendation Engine

Bozhidar Stevanoski
Faculty of Computer Science and Engineering
University Ss. Cyril and Methodius
Skopje, Macedonia
stevanoski.bozidar@gmail.com

Alisa Krstova
Faculty of Computer Science and Engineering
University Ss. Cyril and Methodius
Skopje, Macedonia
krstova.alisa@gmail.com

***Abstract*—The wide and heterogeneous course spectrum at the university level introduces the need for applications that analyze educational records to assist the students in developing a detailed and well-informed study plan. The goal of this paper is to make a contribution to this area of research in two ways: 1) we propose a prototype of a course recommendation engine based on the SimRank algorithm, 2) we develop a method of calculating the most probable grade for the recommended courses based on course-course and student-student similarity. Furthermore, using the aforementioned approach, we seek to identify groups of related courses that would further facilitate the process of making an informed choice.**

Keywords— course recommendation engine; study plan development; student performance prediction

I. Introduction

While having a large number of available university courses to choose from can be the basis for developing a study plan tailored to every student individually, the scope of fields and topics can often be overwhelming. One way of tackling this problem is making relevant, personalized recommendations that would facilitate the process of course selection and guide students in their desired career path. However, building a recommendation engine for an educational environment differs from recommending music, news or products - such a system should support the learning process by incorporating domain-specific information, for example students' academic records, learning strategies, interests etc. We aim to translate this idea into a prototype of a system that recommends courses and makes corresponding grade predictions by finding semantic relationships between the courses and the students that require assistance in the enrollment process.

II. Related Work

Careful investigation of past attempts to construct an accurate course recommendation engine is a valuable starting point that allows us to identify potential problems and aspects that require more attention. Over the past few years, a remarkable amount of research has been devoted on methods to make relevant course recommendations in order to improve the quality of higher education. Many of the proposed approaches propose different variants of collaborative filtering algorithms that exploit similarities between students based on the courses they have taken. The authors of [1] show that there is not much difference in accuracy between user-based and item-based collaborative filtering methods when it comes to recommending elective courses and predicting the possible outcomes (grades). However, the dataset used in this research is rather small in size, so this statement needs further elaboration. The collaborative recommendation systems described in [2], [3] additionally employ association rules to find related courses. An attempt to use fine-grained information related to the student's past performance and engagement in making recommendations is described in [4].

Within the context of developing models for grade prediction, some of the most interesting approaches include neighborhood-based methods in combination with matrix factorization. The research presented in [5] offers a comparison between several course-specific approaches relying on linear regression and on sparse linear and low-rank matrix factorization. The results indicate that a student-course-specific approach using the same techniques leads to poorer performance due to the course diversity. The authors of [6] evaluate Probabilistic Matrix Factorization [7] and Bayesian Probabilistic Matrix Factorization using Markov Chain Monte Carlo [8] to predict grades for newly enrolled courses based on past student records. The experiments show that the two methods demonstrate varying levels of precision for the different grades - the PMF algorithm is better suited for predicting lower grades, whereas the higher grades are more accurately identified by the BPMF method. A combination of these methods leads to nearly 80% of the grade predictions to fall within a deviation of ± 1.

III. Dataset

The experiments were conducted using an internal dataset of fully anonymized student records at the authors' institution. This dataset gives an overview of student performances at the end of an exam period. Each record provides information about the student-course-grade relation, i.e. which student (identified solely by an integer ID) obtained which grade in the corresponding course. There are 4053 students from different semesters of undergraduate studies and 369 courses belonging to different study programs offered at the university level. The grades range from 5 to 10, with 6 being the lowest passing grade.

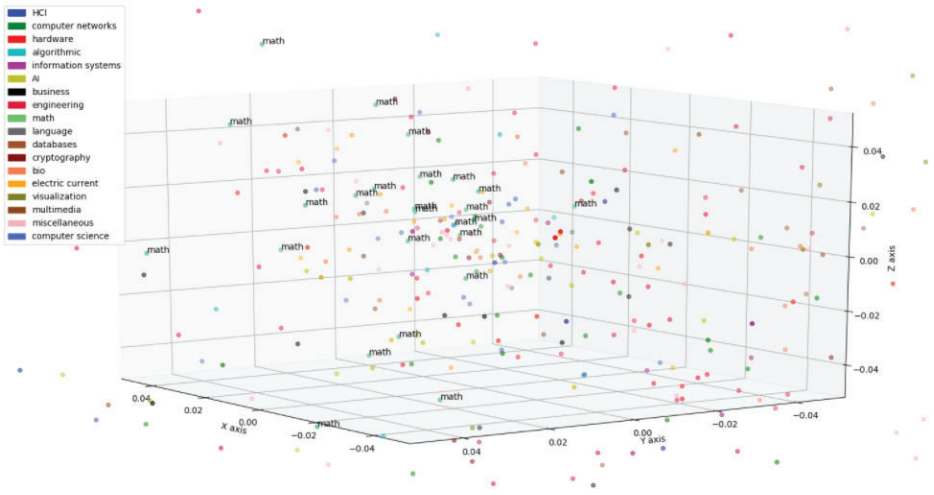


Fig. 1. Three most important principal components of mathematical courses.

Identifying relationships between courses is crucial to solving the problem of making accurate recommendations. We attempt to achieve this using a representation of the student-course data in the form of 40-dimensional vectors obtained using the Alternating Least Squares (ALS) algorithm for matrix factorization. Recognizing patterns of 369 courses over 40 dimensions is a rather impractical task; therefore we examine them using Principal Component Analysis (PCA). In order to recognize possible correlations, we visualize the 3 most important resulting components by assigning a category to each course, i.e. grouping the courses into 18 logical units (mathematics, software engineering courses, computer science courses etc.). This, unfortunately, does not provide us with meaningful insights about the similarity between the courses - the data is scattered across all 3 dimensions and no explicit patterns can be observed. A weak grouping illustrated in Fig. 1 is formed by the courses from the field of mathematics, however the comparatively large values for the intra-cluster distances mean that these results may not be useful for improving recommendations.

IV. Methodology

In order to group similar courses in one cluster, first we define course similarity. Intuitively, two courses are similar if similar students have enrolled in both of them, and two students are similar if they are enrolled in similar courses.These definitions are cyclic. The intuition behind student and course similarity resembles the one concerning ranking websites by their importance. Therefore, we approach the similarity problem with an algorithm that shares the idea of Google's PageRank, and hence the name SimRank [9].

A. SimRank

We construct a directed bipartite graph whose nodes represent the students and the courses from our dataset. A directed edge from student s to course c exists if and only if s has enrolled c.

We distinguish the similarities of the two types of nodes in our graph. Specifically, the similarity of students s_1 and s_2, and the one of courses c_1 and c_2 are recursively calculated as

$$s\left(s_1, s_2\right) = \frac{C_1}{\left|O\left(s_1\right)\right|\left|O\left(s_2\right)\right|} \sum_{i=1}^{|O(s_1)|} \sum_{j=1}^{|O(s_2)|} s\left(O_i\left(s_1\right), O_j\left(s_2\right)\right) \tag{1}$$

$$s\left(c_1, c_2\right) = \frac{C_2}{\left|I\left(c_1\right)\right|\left|I\left(c_2\right)\right|} \sum_{i=1}^{|I(c_1)|} \sum_{j=1}^{|I(c_2)|} s\left(I_i\left(c_1\right), I_j\left(c_2\right)\right) \tag{2}$$

where C_1 and C_2 are hyperparameters in the range $(0, 1)$ that can be thought of as decay factors, $I\left(v\right)$ and $O\left(v\right)$ are the set of in-neighbors and out-neighbors of v, and $I_i\left(v\right)$ and $O_i\left(v\right)$ is an individual in-neighbor and out-neighbor respectively. As originally proposed in [9], we take $C_1 = C_2 = 0.8$.

B. Louvain Algorithm

Once course-course similarities are acquired, we build a new undirected weighted graph in which the courses are denoted as nodes, and there is an edge between them if and only if their similarity is larger or equal to a threshold t. The weight of each edge is equal to the similarity of the courses it connects.

In this graph, we identify clusters or communities of courses with many intra-cluster, and only a few inter-cluster edges. In other words, we partition the graph into groups of vertices, such that we maximize the modularity of the partition. Modularity is a measure of the quality of the clustering the graph vertices in the range $[-1, 1]$, or more formally defined as $Q = \frac{1}{2m}\sum_{i,j}\left[w_{i,j} - \frac{d_i d_j}{2m}\right]\delta\left(c_i, c_j\right)$, where m is the total weight of all edges, $w_{i,j}$ - the weight of the edge between vertices i and j, d_i - sum of the weights of the edges incident to node i, and δ - the Kronecker delta function.

In order to find such partition, we utilize the Louvain algorithm - a heuristic method that is shown to outperform other similar modularity optimization methods [10]. The algorithm consists of two phases that are repeated iteratively. Initially,

every vertex is assigned to a separate community. During the first phase, for each vertex v_i it considers all vertices v_j that are connected to v_i, and calculates the gain in modularity if v_i is removed from its current community and is moved to the community of v_j. The maximal gain over all neighbors is selected, and if positive, v_i is moved. This procedure is applied to each vertex. In the second phase, a new graph is built in which the nodes belonging to the same cluster are merged together. The algorithm terminates when no vertex is moved to other cluster in the first phase.

C. Similarity-based Course Recommendation

We define an appropriateness of course c for student s as

$$A(s,c) = \sum_{q_c \in N(s,c)} g_{q_c} s_{c,q_c} \tag{3}$$

where $N(s,c)$ is the set of courses previously taken by the student s which contain the course c in their k_c closest neighbors and g_{q_c} is the grade obtained by the student s in the course q_c.

As a recommendation for new r courses for student s to enroll in, we propose taking the r courses with the largest appropriateness measure for s.

D. Similarity-based Course Grade Prediction

On the basis of the aforementioned SimRank scores we develop an algorithm to make a grade prediction for a given user for a course she would like to enroll. First, we extract the k_c-nearest course neighbors of the target course, based on the SimRank scores. For these courses we compute the following value:

$$S_q = \frac{1}{k_c} \sum_{i=1}^{k_c} f(g_i)\, s_i, \tag{4}$$

where S_q is the cumulative SimRank score for the query student, n is the number of courses she has taken, g_i is the obtained grade for the i^{th} course, f is a piecewise function that maps a grade to a value between 0 and 1, and s_i is the SimRank similarity score between the i^{th} course and the query course. From a semantic point of view, the grades from the training set are not equal in terms of distribution and frequency. Therefore, we transform the six possible grade values using different mapping functions f, such as the min-max normalization and tangent-like defined in (5) and (6) respectively.

$$f(g) = \begin{cases} 0 & \text{if } g = 5 \\ 0.2 & \text{if } g = 6 \\ 0.4 & \text{if } g = 7 \\ 0.6 & \text{if } g = 8 \\ 0.8 & \text{if } g = 9 \\ 1 & \text{if } g = 10 \end{cases} \tag{5}$$

$$f(g) = \begin{cases} 0 & \text{if } g = 5 \\ 0.3 & \text{if } g = 6 \\ 0.45 & \text{if } g = 7 \\ 0.55 & \text{if } g = 8 \\ 0.7 & \text{if } g = 9 \\ 1 & \text{if } g = 10 \end{cases} \tag{6}$$

The tangent-like function f captures the difference between the failing grade 5 and the lowest passing grade 6, the two highest grades 9 and 10 and observes the differences between the intermediate grades as small. To the best of our knowledge, this is a novel approach in this particular context. In the next step, we choose the k_s-nearest student neighbors of the student of interest who have been enrolled in the query course (also calculated based on SimRank scores), and we compute their cumulative value by selecting the k_c courses they have taken that are most similar to the query course and determining the final score using a formula analogous to (4). We predict the grade as mean of the grades gotten by the k_n students whose cumulative SimRank score is closest to the one of the query student.

V. Results and Discussion

In this section we analyze the results from our experiments with different parameters for the aforementioned algorithms and compare the modularity of the course clusters and accuracy of the grade predictions. For the task of grade prediction, we randomly choose 90% of the data for the training and the remaining 10% for test set. An 80-20 ratio leads to significantly worse results.

A. Course Cluster Modularity Results

The approach described in Section IV-B allows us to experiment with different values for the threshold t to find the most optimal grouping of the courses in our dataset. As shown in Table I, the best clusters are found when the similarity threshold is set to 0.07 for which we get modularity of **0.875**. The three subsequent values for the threshold give comparable results. A visualization of the clusters for one of the most optimal values, $t = 0.075$ is presented in Fig. 2. It is important to note that there is a tradeoff between the value for the modularity of the clusters and the number of courses covered by the clusters.

TABLE I
Dependency Between SimRank Threshold and Cluster Modularity

Threshold	Modularity
0.07	0.875101472
0.065	0.873507831
0.075	0.872919737
0.08	0.870902711
0.085	0.868282379
0.06	0.866736672
0.09	0.862220643
0.055	0.860249929
0.05	0.852247659

Table II presents some of the identified clusters for $t = 0.07$. This version of the algorithm yields a total of 19 clusters; for brevity we list four. Courses grouped together have a strong semantic connection - the members of the first cluster focus on topics on hardware and embedded systems; the second cluster represents courses that are offered as part of the Computer Networks Technologies study program, whereas the third cluster groups together courses related to advanced areas of software engineering and its theoretical foundations. The courses listed in the fourth example cluster are connected because they are offered across all study programs as part of the first year of studies.

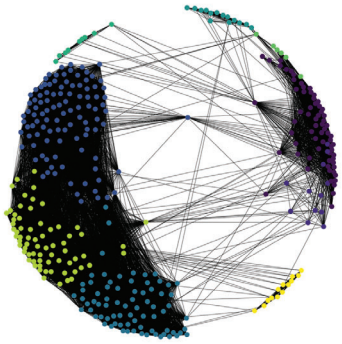


Fig. 2. Cluster visualization for $t = 0.075$

TABLE II
EXAMPLES OF COURSE CLUSTERS

No.	Courses
1	Computer Processors, Mobile and Embedded Systems, Software Development for Embedded Systems, Software Reliability, Integrated Circuit Design, Digital System Design using HDL, Design with HDL Languages, Design Techniques for Microchip Systems
2	Wireless and Ad-Hoc Computer Networks, Monitoring and Performance of Computer Networks, Network Virtualization and Cloud Computing, Advanced Computer Networks, Advanced Modelling and Simulation, Sensor Networks, Modern Methods for Network Analysis, Computer Network Administration, Public Mobile Services and Applications, E-Marketing
3	Architecture, Models & Design, Advanced Topics in Software Engineering, Software Engineering for Critical Systems, Software Engineering for Database Systems, Software Engineering for Large-Scale Database Systems, Formal Methods in Software Engineering, Software Quality and Testing, System Integration
4	Structured Programming, Introduction to Computing, Introduction to Web Design

B. Course Grade Prediction Results

We conduct several experiments with different values for the parameters of the algorithm described in section IV-D, i.e. with the number of closest student neighbors k_s and course neighbors k_c and the grade transformation function. The results show that the best performance is achieved by taking the 10 most similar students, 10 most similar courses to the student-course query pair and 5 students with closest cumulative SimRank scores, i.e. $k_c = 10, k_s = 10, k_n = 5$. It is important to note that the tangent-like function outperforms the min-max normalization in all cases, regardless of the values for the hyperparameters. This combination of parameters leads to **74.03%** of the predictions to fall within a ± 1 range of the actual grade. Increasing the number of nearest courses and students taken into consideration (example values: $k_c = 20, k_s = 15$) leads to a decline in accuracy to 72.54%. A more significant deterioration in performance is observed when the considered neighborhood is small ($k_c \leq 5, k_s \leq 5$) - the accuracy drops below 68%. Also, changing k_n to 10 or 3 gives 70.75% and 72.09% respectively.

VI. CONCLUSION AND FUTURE WORK

In this paper, we partitioned the university courses at our institution into clusters to achieve a significant modularity based on the SimRank algorithm. We also extended to evaluate several strategies to construct a university course recommendation and grade prediction engine. The results of our explorations show that it is possible to make reasonably accurate recommendations and predictions by exploiting the similarity between courses and students.

The accuracy on both tasks can be further improved by applying the described algorithms on an increased number of training instances relevant to the current study programs offered at the authors' institution, perhaps even by using methods of incremental learning, since as we noted, the size of the dataset is border-line sufficient. Our current research efforts are directed to optimizing the current procedures by including the data about the professors in each student-course-grade record, as well as some background information about the student, such as the high school she graduated from.

ACKNOWLEDGMENTS

This work is a result within the project ISISng (Integrated Study Information Systems of the Next Generation) [11], which is currently ongoing at the Faculty of Computer Science and Engineering. The authors would also like to thank Ljupcho Rechkoski for the provided materials.

REFERENCES

[1] S. Ray and A. Sharma, "A collaborative filtering based approach for recommending elective courses", In: Dua S., Sahni S., Goyal D.P. (eds) Information Intelligence, Systems, Technology and Management. ICISTM 2011. Communications in Computer and Information Science, vol 141. Springer, Berlin, Heidelberg, 2011

[2] A. Al-Badarenah and J. Alsakran, "An automated recommender system for course selection", International Journal of Advanced Computer Science and Applications (IJACSA), Vol. 7, No. 3, 2016

[3] Y. Lee and J. Cho, "An intelligent course recommendation system", Smart Computing Review, vol. 1, no. 1, October 2011

[4] A. Elbadrawy, R. S. Studham, and G. Karypis, "Collaborative multi-regression models for predicting students' performance in course activities", In Proceedings of the 5th International Conference on Learning Analytics and Knowledge, LAK 2015 (Vol. 16-20-March-2015, pp. 103-107)

[5] A. Polyzou and G. Karypis, "Grade prediction with models specific to students and courses", International Journal of Data Science and Analytics, 2016

[6] L. Rechkoski, V. Ajanovski, and M. Mihova, "Evaluation of grade prediction using model-based collaborative filtering methods", In Global Engineering Education Conference (EDUCON), 2018 IEEE (accepted)

[7] R. Salakhutdinov and A. Mnih, "Probabilistic matrix factorization", In Proceedings of the 20th International Conference on Neural Information Processing Systems, Vancouver, British Columbia, Canada, 2007

[8] R. Salakhutdinov and A. Mnih, "Bayesian probabilistic matrix factorization using Markov chain Monte Carlo", In Proceedings of the 25th International Conference on Machine Learning (ICML '08). ACM, New York, NY, USA, 2008

[9] G. Jeh and J. Widom, "SimRank: A measure of structural-context similarity", In Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 538-543. ACM Press, 2002

[10] V. D. Blondel, J. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks", Journal of Statistical Mechanics: Theory and Experiment. 2008

[11] Integrated Student Information System of the Next Generation – Official project website. (2009–2018) https://develop.finki.ukim.mk/projects/isis

Detecting Customer Sentiment in Amazon Review Data

Alisa Krstova, Lodi Dodevska, Sonja Gievska
Faculty of Computer Science and Engineering
Ss. Cyril and Methodius University
Skopje, Macedonia
krstova.alisa@students.finki.ukim.mk, dodevska.lodi@students.finki.ukim.mk, sonja.gievska@finki.ukim.mk

***Abstract*—The benefits of online reviews are twofold: a) providing companies with insights into customers' preferences and satisfaction, b) allowing customers to take into account other people's opinions when they make informed decisions related to product purchases. This paper aims to contribute to the research on deep sentiment analysis. It reports an experiment carried out on a dataset of millions of reviews from Amazon customers. Several architectures of convolutional neural networks (CNN), recurrent neural networks (RNN) and hybrid models have been trained. The results of the performance analysis of different models and the impact of tuning various hyper-parameters have been discussed and compared.**

Keywords—sentiment analysis, deep learning, convolutional neural networks, recurrent neural networks

I. INTRODUCTION

A central commitment of companies and service providers is to ensure customer satisfaction. Automatic sentiment analysis of online users' reviews, posts and comments have been emphasized as providing opportunities for getting insights into users' preferences and customers' satisfaction. The analysis can take on different meanings depending on the application domain - this research focuses on identifying the polarity of a sentiment expressed in someone's opinion i.e., classifying it as a positive, negative or neutral opinion towards the product or item that is targeted in the review.

The problem of sentiment analysis of text has received a considerable research attention in the intelligence community. Drawing on theoretical research in computational linguistic, a number of studies have been conducted to investigate either correlations between various text features and text sentiment, or their predictive power regarding automatic sentiment analysis [1]. More recently, deep neural networks have become a guiding method for learning data representation that are essential for complex natural language processing (NLP) tasks, such as machine translation, question answering, or opinion mining. Research show that while many new deep learning practices for tackling the open problems in NLP are emerging [2],[3], coherent perspectives on the type of architectures and parameter tuning are still missing.

The purpose of this study was to examine various deep neural architectures for detecting sentiment polarity that can perform well across several categories of products included in the Amazon product reviews dataset. After a brief discussion of relevant research that has investigated sentiment analysis using similar deep neural models, we highlight the primary findings of our research. In what follows, we report on experiments carried out to investigate the differences in performance achieved by CNN-based, RNN-based and hybrid neural models as they relate to polarity detection of users' opinion expressed in online reviews. Issues that need further attention will also be discussed.

II. RELATED WORK

This study was guided by a large body of evidence that documents the power and performance advantage of deep neural architectures for variety of NLP tasks. Of particular relevance to the authors of this paper were the research works that pertain to opinion mining presented in [4], [5],[6]. The research presented in [4] has pointed out that deep, narrow RNN architectures employed for opinion mining in a news dataset perform better than shallow, wide RNNs with the same number of parameters. Word-level and character-level embeddings have been used to better capture the complex semantic structures. The superiority of hierarchical bidirectional RNNs that have achieved a precision as high as 94% on the task of aspect-specific opinion mining of DBS Text Mining Challenge 2015 dataset have also been stressed in [5].

A novel hybrid model ConvLstm that uses both, recurrent and convolutional layers on top of pre-trained word vectors have shown an accuracy of 88.3% on the Stanford Sentiment Treebank [6]. Inspired by this research, we have explored the idea of using hybrid architectures that are based on RNNs and it variants, such as Gated Recurrent Neural Networks (GRU), as a substitute for pooling layers.

III. DATASET

Our deep learning models for sentiment analysis were trained on Amazon dataset[1], which is significantly larger than other available datasets. It includes 142.8 millions of reviews (text, ratings, helpfulness votes) and product metadata (e.g., brand, product descriptions, price) for variety of products. For our particular task, we have used only the following attributes: ID of a product, the text of a review, and the product rating.

A subset of the Amazon dataset that includes three related categories of products, namely, Android application, cell phones and electronic devices were used in our experiments. A total of 2,636,564 reviews were distributed as follows: 752,937 reviews for Android applications, 194,439 reviews for cell phones and accessories, and 1,689,188 reviews for electronic devices.

IV. PREPROCESSING

In the absence of appropriate labels for the sentiment polarity of a review, we have applied the following label assignment. The reviews with 1- and 2-star ratings were assigned as negative reviews, those with 4- or 5-star ratings were marked as positive, and the 3-star ratings were considered neutral. The distribution of the three classes of sentiment (positive, negative, neutral) was highly unbalanced, the number of positive reviews being at least three times larger than the number of neutral and negative reviews. We have opted for removal of almost half of the positive reviews chosen randomly, resulting in a more balance dataset containing 1,024,722 positive reviews, 338,305 negative reviews and 248,817 neutral reviews.

A tokenizer was used to identify the 20,000 most frequent words in the reviews that will serve as a vocabulary V; each word has an associated vocabulary index k. The length of each review text was set to 300 words resulted in a word embedding matrix $M \in \mathbb{R}^{n \times |V|}$, where n=300. One-hot encoding was used to encode the categorical class features.

V. MODELS AND ARCHITECTURES

Our research objective was to empirically validate different deep learning structures suggested for their suitability and performance in the context of sentiment analysis. The relationship between the model parameters and performance metrics has been a topic of interest in the field, and this section describes the models we have experimented with.

A. CNN Architectures

Gaining popularity in domains such as image recognition and speech recognition, convolutional neural networks have been applied to various NLP tasks. A CNN architecture for text classification with a kernel and max-pooling layer have been especially successful in determining discriminative phrases in text [7]. A carefully chosen size of the window i.e., kernel is of crucial importance - a large window may lead to a complex network and longer training times, whereas too small window may not be suitable for capturing the relationships between words in a longer sentence.

Models 1-4 are four different architectures based on convolutional neural network model that were evaluated for the performance benchmarks. Model 1 has an embedding layer with an input dimension of 20,000 (equivalent to the size of our vocabulary) and input length of 300 words as the first layer. This layer outputs 256 units which are then passed to a sequence of three one-dimensional convolutional layers with 64 filters and a kernel i.e., a sliding window of size 5. Each convolutional layer has LeakyReLU as activation function and is followed by a max pooling layer and a regularization layer with 20% dropout rate. Finally, we use a fully connected layer and an output layer with three neurons. A sigmoid activation function was used to calculate the respective probability for each of the three classes of reviews. Model 2 is a modification of Model 1 with a dropout rate set to 30%. The difference between Model 3 and Model 1 is in the alpha parameter of the LeakyReLU activation function that was set to 0.3. Model 4, uses softmax instead of a sigmoid as activation function. The first layer of Models 2, 3 and 4 outputs 128 units.

B. RNN Architectures

The advantages of recurrent neural networks is their capability of preserving the relationships between distant words in a given text in a fixed-size hidden layer; closer words being more important than the distant ones. In many languages, the most important relationships can be found between words at the very beginning and at the very end of a sentence, making RNNs not the most suitable choice. To remedy this issue, a more powerful version of RNNs known as Long-Short Term Memory (LSTM) network has emerged as an effective model for handling sequential data [8], in our case sentences. The central idea behind these models is a memory cell (a neural network within the neural network) which maintains its state and captures the temporal dependencies, and non-linear gates which regulate the amount of information the memory cell can hold.

A more recent and computationally more efficient variant of LSTMs are the GRUs [9]. In comparison to the

[1] http://jmcauley.ucsd.edu/data/amazon/

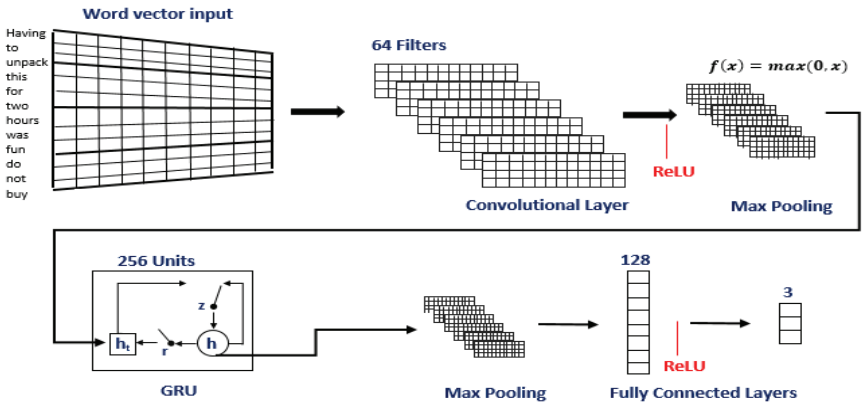


Fig. 1. Hybrid architecture of Model 7

LSTMs which have three gates (input, output and forget gates), GRUs merge the input and output into one update gate enabling faster training time of this type of networks.

Model 5 consists of a convolutional layer which has an input dimension of 64 and a kernel of size 5 inserted between the embedding layer and an LSTM unit with 128 input dimensions. Model 6 has an LSTM unit of size 256, to better memorize temporal dependencies in the sentences of the training corpus. GRU was used instead of LSTMs with a recurrent dropout of 0.2 in Model 7. The overall architecture of Model 7 is illustrated in Fig. 1. Models 5-7 use local max pooling layers, ReLU as activation function and Adam as a weight optimization algorithm.

VI. Results and Discussion

To address the question of whether or not, and to what extent, a particular architecture can lead to a better performance, we have evaluated the architectures described in the previous section. While aiming for achieving high accuracy, avoiding overfitting was also an objective to ensure the model is robust enough to handle previously unseen data. There are several distinct fronts through which we have addressed the problem of overfitting: 1) by experimenting with the hyper parameters during the training process, 2) exploring the effects of variable input dimensions, 3) using more aggressive regularization dropout rates and 4) increasing the number of training epochs. The results demonstrating the effects of such experiments are summarized in Table 1. The core of models 5 to 7 are the RNNs, so different factors come into play when optimizing the overall architecture.

The results revealed four broad patterns related to the architectural parameters that were varied across the seven models:

1) By comparing Model 1 and Model 4, we may conclude that traditional sigmoid function performs slightly better compared with softmax activation function, even though it is often recommended as a good choice for dealing with multiple classes and categorical cross entropy. Varying the value of the alpha parameter for LeakyReLu (from 0.1 in Model 1 to 0.3 in Model 3) did not result in any significant performance changes.

2) The performance was improved for the hybrid models by increasing the input dimension from 128 in Model 5 to 256 in Model 6. However, the most successful architecture that uses GRU have 128 input levels.

3) By contrasting the performance results of Model 1 and 2, it is evident that a dropout rate of 0.3 might be too aggressive and prevents the network from learning the data well.

4) The results did not show significant changes in the performance measures when the number of epochs (e.g., 3,5) for the proposed CNN-based architectures was varied.

There are a few interesting observations concerning the outcomes of our experiments. First, it is evident that the *hybrid structure* which implements recurrent layers and memory cells on top of the convolutional layer significantly *outperforms standard CNN architecture* leading to an increase in accuracy of up to 10% depending on the model. There is also a dramatic decrease in the training and validation loss, which means that RNN-based networks make better and more accurate predictions of the sentiment polarity of customers' reviews.

The performance analysis is consistent with other research which have pointed out the differences between two architectures - RNNs are well suited to capture ordered information and long-term context dependencies, whereas CNNs are considered to be good at encoding local features and important segments in shorter texts [10].

TABLE I. PERFORMANCE RESULTS FOR MODELS 1-7

Model	Loss	Accuracy	Val. loss	Val. acc.
1	0.6443	0.7466	0.6437	0.7470
2	0.6627	0.7400	0.6507	0.7440
3	0.6483	0.7456	0.6469	0.7465
4	0.6515	0.7416	0.6510	0.7383
5[a]	0.4262	0.8275	0.4216	0.8295
6	0.3801	0.8478	0.4335	0.8293
7	0.3785	0.8521	0.4183	0.8349

[a.] Shaded data refer to the hybrid architectures.

Considering that the average length of single review in the dataset is above 100 words, CNNs fail to encode all context-sensitive sentiment indicators in a review. There is, however, a caveat to this - RNN-based architectures are much slower to train than CNNs.

One explanation for the performance advantage of hybrid models is because detecting sentiment polarity is based on understanding the semantic meaning of longer reviews, rather than detecting subtle syntactic or stylistic clues indicating a certain language phenomenon (e.g., sarcasm, topic, deception). The reason behind the superior performance of GRUs is that they are designed to store long-range semantic dependencies rather than focusing on isolated parts of the sentence. For example, there might be a review which contains lots of positive words, but carries an overall negative sentiment, such as the following: "I thought this would be an improvement over the previous model... Yeah right!". Although the words "improvement" and "right" are usually associated with positive sentiment, this is an example of a negative opinion about a particular product.

It was noted that the gap between training and validation accuracy (and the respective losses) is rather small, which means that the networks do not overfit. However, during the different training epochs in some of the models we observed that the validation accuracy is higher than the training accuracy. Despite the fact that training accuracy is not an objective measure of the goodness of the model, such results might be another confirmation that the networks really do not learn the data "by heart". This phenomenon could be explained by the fact that we use dropout layers during the training process, which means that random neurons are turned off in each forward pass to improve robustness. The validation process, on the other hand, uses the whole network as it is and evaluates the model on data which has not been trained upon.

VII. CONCLUSION AND FUTURE WORK

This article contributes to the conversation of how deep learning can support sentiment classification of massive review datasets. This paper reflects upon the challenges surrounding the efforts in identifying sentiment polarity in Amazon customers' reviews. Furthermore, results in our research indicate the necessity to explore different deep neural models to address the problem at hand. Our findings highlight the performance advantage of hybrid architectures compared with traditional convolutional neural networks, although we should note that the accuracy improvements come at the cost of slower training time.

At the moment, more experiments using word2vec word representations trained on domain-specific and cross-domain data are in place. Our future research efforts are directed toward exploring the impact of emotional content of reviews on a fine-grained multi-level classification of opinions.

REFERENCES

[1] B. Liu, Sentiment Analysis and Opinion Mining. Morgan & Claypool Publishers, 2012

[2] P. Liu, X. Qiu, and X. Huang, "Recurrent neural network for text classification with multi-task learning", Fudan University, Shanghai, China, 2017

[3] E. Sygkounas, G. Rizzo, and R. Troncy, "Sentiment polarity detection from Amazon reviews: an experimental study", In: Sack H., Dietze S., Tordai A., Lange C. (eds) Semantic Web Challenges. SemWebEval 2016. Communications in Computer and Information Science, vol 641. Springer, Cham

[4] O. Irsoy and C. Cardie, "Opinion mining with deep recurrent neural networks", 2014 Conference on Empirical Methods on Natural Language Processing (EMNLP 2014), Doha, Qatar.

[5] A. Chaudhuri. and S.K. Ghosh, "Sentiment analysis of customer reviews using robust hierarchical bidirectional recurrent neural network", In: Silhavy R., Senkerik R., Oplatkova Z., Silhavy P., Prokopova Z. (eds) Artificial Intelligence Perspectives in Intelligent Systems. Advances in Intelligent Systems and Computing, vol 464. Springer, Cham, 2016

[6] A. Hassan and A. Mahmood, "Deep learning approach for sentiment analysis of short texts", 3rd International Conference on Control, Automation and Robotics (ICCAR), Nagoya, 2017, pp. 705-710.

[7] S. Lai, K. Liu, L. Xu and J. Zhao, "Recurrent convolutional neural networks for text classification", Proceedings of Association for the Advancement in Artificial Intelligence (AAAI), 2015

[8] K. Greff, R. K. Srivastava, J. Koutník, B. R. Steunebrink, and J. Schmidhuber, "LSTM: a search space odyssey", IEEE Transactions on Neural Networks and Learning Systems, vol. 28, no. 10, pp. 2222-2232, October 2017.

[9] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling", NIPS 2014 Workshop on Deep Learning, 2014

[10] W. Yin, K. Kann, M. Yu, and H. Schütze, "Comparative study of CNN and RNN for natural language processing", 2017

Detecting Emotions in Tweets Based on Hybrid Approach

Ivona Najdenkoska
Faculty of Computer Science and Engineering
Ss. Cyril and Methodius University
Skopje, Republic of Macedonia
najdenkoska.ivona@students.finki.ukim.mk

Frosina Stojanovska
Faculty of Computer Science and Engineering
Ss. Cyril and Methodius University
Skopje, Republic of Macedonia
stojanovska.frosina.1@students.finki.ukim.mk

Sonja Gievska
Faculty of Computer Science and Engineering
Ss. Cyril and Methodius University
Skopje, Republic of Macedonia
sonja.gievska@finki.ukim.mk

***Abstract*—The research that has been conducted and presented in this paper highlights the key ideas on emotion detection that has been under our empirical investigation. This paper presents a hybrid approach for emotion recognition in tweets that is based on linguistic analysis and machine learning. A categorical model distinguishing between four emotions, namely, anger, fear, joy, sadness with an addition of a neutral class was adopted. The comparative performance analysis of four machine learning algorithms points out to their suitability and limitations in recognizing the emotional content in tweets.**

***Keywords*—emotion detection, tweets, hybrid approach, natural language processing, machine learning**

I. Introduction

Proliferation of intelligence technologies is closely related to current influential technological megatrends, such as smart, ubiquitous technologies and social networking. A gradual shift in research interest toward multimodal natural interaction have drawn attention to human affect recognition, a field that that is part of the wider field of sentiment analysis. Emotions play an important role and drive everything we do, from reasoning to making decisions to learning. A large body of research in psychology focuses on the human perception of emotions that shows a universal consistency but also great variability across modalities, contexts and individuals. The goal of computational emotion analysis of user-generated unstructured text, audio, images or video is to detect and identify the polarity, intensity and/or classify human emotional states [1].

Over the past two decades, research on affective analysis has emerged, blending the advances in computational linguistics, natural language processing (NLP) and machine learning (ML) [2]. At the same time, modelling human cognitive processes and reasoning with high-level constructs such as emotional states, moods and interpersonal stance is a new field of research. Much remains to be done in the interdisciplinary research towards a digital experience that is deeply human-centred.

Theories of emotions provide suitable frameworks for modelling and studying the phenomenon of human affect. There are two general models for emotion representations: categorical and dimensional. A categorical emotional model classifies emotions in a number of distinct classes, which in the well-known Ekman's Facial Action Coding System [3] is set to six: anger, disgust, fear, happiness, sadness, and surprise.

On the other hand, there are other models that represent the emotions in a dimensional form where each emotion occupies a certain part of the space. That is the approach of Russell with his Circumplex model (shown in Fig. 1), where the emotions are represented in space with two dimensions: the valence of emotion, which indicates whether the emotion is positive or negative, and the arousal of the emotion, indicating the level of energy associated with the emotion [4], [5].

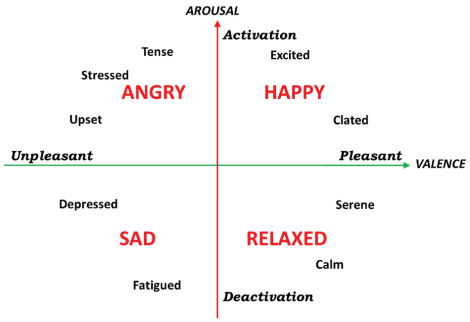


Fig. 1. Representation of the Circumplex model.

II. Methodology

A. Dataset

The dataset that we are using for the experiments is provided for the Workshop on Computational Approaches to Subjectivity, Sentiment, and Social Media (WASSA), by the researchers in computational linguistics Saif M. Mohammad and Felipe Bravo-Marquez [6]. There are four different training and test datasets, for the four emotions respectively. The four emotions that are used for the classification of the tweets are anger, fear, sadness, and joy. These four types of emotions are the focus of this experiment. Every dataset consists of four attributes: tweet id, tweet, emotion type and intensity, shown in Table I. Firstly, we decided to merge the four separate datasets into one, because our goal is to classify each tweet to an emotion. After merging the datasets, we have one training dataset which is consisted of 3613 tweets, labeled with their

emotional intensity and emotion type as a class attribute. On the other hand, the test dataset consists of 3142 tweets labeled with the same attributes.

TABLE I
DATASET ATTRIBUTES AND THEIR TYPES

Attribute	Type
Tweet ID	Integer
Tweet	String
Emotion Intensity	Numeric
Emotion type	Nominal

After reviewing some of the provided tweets and their labeled emotion and intensity we decided to remove a certain number of them and add a neutral emotion. This decision is due to our notice that sometimes the emotion label does not provide the proper emotion for a specific tweet. For example, the following tweet: "Don't #worry if you're not the best, if you are doing something you #love, you're heading in right direction ..." is labeled as fear, but clearly, this sentence is wisdom quote and it would be more suitable to label it with a neutral emotion. The intensity of the labeled emotion for this tweet is 0.104. Because of the low intensity and the meaning of the sentence, we can easily conclude that it is better to classify it as a default class, i.e. neutral emotion, and so we are not taking this kind of tweets in consideration. To manage this challenge, we define a threshold of 0.34 for the emotional intensity. Each tweet that has an intensity above this threshold is retained in the dataset, and each tweet that has an intensity below this threshold is omitted. Adding an additional neutral emotion will help us to build more accurate classifier because if the tweet does not provide enough information for the detection of emotion, the classifier will choose the default class.

This dataset does not have a label for neutral emotion, so we decided to find proper tweets from some other dataset, and append them to the existing dataset. In addition, we found a dataset for sentiment classification from Sentiment140[1], provided by Alec Go, Richa Bhayani, and Lei Huang [7]. This dataset consists of tweets with positive, negative and neutral class for the sentiment of the tweet, so for our purpose, we get the tweets from the test set with neutral class and add them to the initial dataset. After this, we ended up with a dataset containing 5348 tweets labeled with five different classes for the emotional state shown in Fig. 2 with their distribution.

B. Hybrid Approach

As mentioned before, usually concepts from the big fields of Natural Language Processing (NLP) and Machine Learning (ML) are used for text analyzing and identifying emotions in text. The main idea of the hybrid approach is to combine methods from these two big fields and make a model that will join the advantages and try to improve the individual disadvantages from both separate approaches. Using NLP techniques, we are going to produce initial features of the tweets, and then after the feature selection, we will use the selected features to build a model with ML algorithms that can classify emotions of tweets. Fig. 3 demonstrates a visual representation of the steps in our proposed approach.

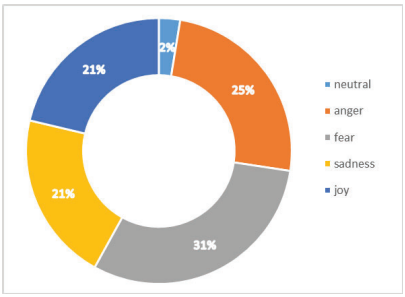


Fig. 2. Distribution of emotion classes in the final dataset.

C. Lexicon

Emotion lexicons are lists of words which have been labeled according to their emotional connotation. The label can simply be an emotion category the word is thought to belong to, or it can be a value representing the strength of a given emotional dimension reflected in the word [8].

After we finished our research in the field of emotion lexicons, we decided to use Warriner et al. extended ANEW (Affective Norms for English Words) lexicon[2]. This lexicon consists of affective norms for valence, arousal and dominance for 13,915 English words (lemmas), which were collected by Amazon Mechanical Turk [9]. The reason for choosing this lexicon is because it uses three dimensions (valence, arousal and dominance) for annotating the words. These dimensions are represented in the PAD (Valence - Pleasure, Arousal, Dominance) model [10]. Valence (also referred to as the pleasure dimension) refers to whether an emotion is positive or negative. Arousal refers to the intensity of which the emotion is experienced or expressed. Both dimensions are independent, in that the valence of an emotion does not affect its activation and vice versa. Dominance is a dimension that represents the controlling and dominant versus controlled or submissive one feels. In this paper, we will use the first two dimensions for generating the features.

D. Preprocessing

The first step in our approach is the preprocessing the text from the tweets. The tweets are presented as sentences that have mis-spellings and casual language used in Twitter. So, to clean up the text we preprocessed the tweets with the following rules:

- Tweets often contain usernames, words that start with the @ symbol. These words are removed from the tweet because they do not provide valuable information within our approach.
- Hash tags in tweets in most of the cases are representative of the emotion expressed in the text, so we decided to

[1]http://help.sentiment140.com/home

[2]http://crr.ugent.be/archives/1003

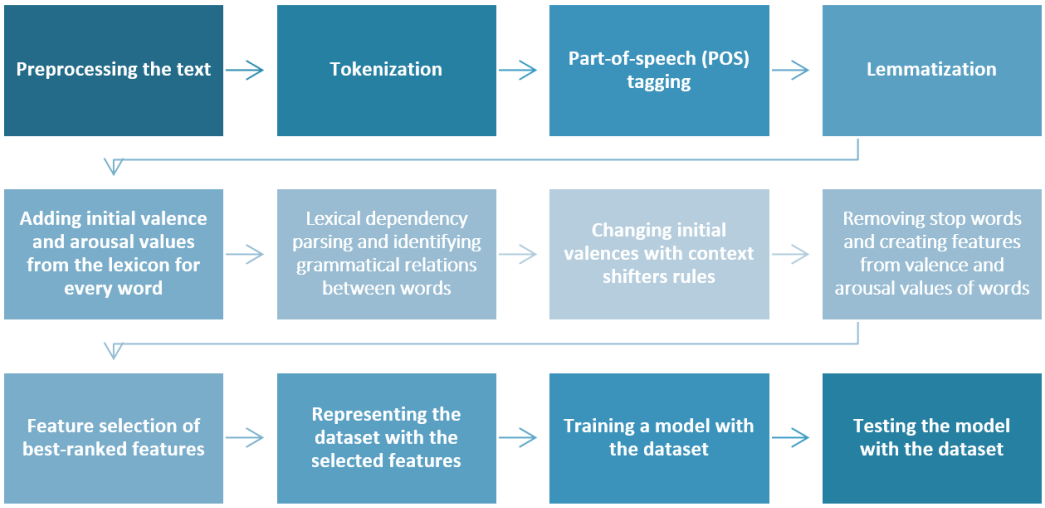


Fig. 3. Fundamental steps in the hybrid approach.

remove the # symbol from the hashtag and keep the rest of the hashtag.

- Another common incorrectness in tweets are words with repeated letters such as ”loveeeee”. Any letter occurring more than two times consecutively is replaced with one occurrence. For instance, the word ”loveeeee” would be changed into ”love”.

Some of the tweets include several types of emojis which are annotated with special characters. We decided to ignore the emojis because they are not applicable to our approach.

E. Feature Extraction

To train a model that would be able to classify emotions from tweets, we represent each tweet with a vector of features. This vector needs to capture the emotion expressed by each tweet. Therefore, in this paper, we explore the usage of valence and arousal dimensions of individual words, obtained from our chosen lexicon, as features in the vector. For the preprocessing of the dataset and feature extraction and selection, we are using the programming languages Java and Python. Also, we use few additional libraries. For example, to work with our dataset we are using the Pandas[3] library for data analysis in Python [11].

The steps of feature extraction from the tweets are:

1) Tokenization
2) Part-of-speech (POS) tagging
3) Lemmatization
4) Adding initial valence and arousal values from the lexicon for every word
5) Lexical dependency parsing and identifying grammatical relations between words
6) Changing initial valences with context shifters rules
7) Removing stop words and creating features from valence and arousal values of words

[3]http://pandas.pydata.org

In this subsection, we are going to describe these steps individually and then explain the process of selecting the best features, i.e. feature selection.

Often natural language processing tools require their input to be divided into tokens. So firstly, we applied tokenization [12] of the tweets. We divided each tweet into tokens using the tokenize method from the class *TweetTokenizer* in *nltk.tokenize* module[4] from the NLTK Python library [13]. Separation of tokens is made by separating commas, quotation marks from words and disambiguating end-of-sentence punctuation (period, question mark, etc.). Before applying the lemmatization, we needed to tag the words with Part-of-speech (POS) tagging [14]. A POS-tagger processes a sequence of words and attaches a part of speech tag to each word. For this part, we use the method *pos_tag*[5] from NLTK library.

After that, the following step was the lemmatization of each word. In linguistic morphology and information retrieval, stemming is the process of reducing inflected (or sometimes derived) word to their word stem, base or root form. Lemmatization usually refers to doing things properly with the use of a vocabulary and morphological analysis of words, normally aiming to remove inflectional endings only and to return the base or dictionary form of a word, which is known as the lemma [12]. This step was also achieved with NLTK Python library using the function lemmatize from the *WordNetLemmatizer*[6] class. This lemmatizer requires the tokens and their POS tag as parameters and returns the root of the words (lemmas).

With the three previously explained steps we had tokens and lemmas for each tweet. Using the lemmas, we assigned each word with the values of dimensions from the lexicon that we have chosen before. Initially, the lexicon had a scale with

[4]http://www.nltk.org/api/nltk.tokenize.html
[5]http://www.nltk.org/book/ch05.html
[6]http://www.nltk.org/api/nltk.stem.html

range 0-9 for the two dimensions. We shifted this scale to a range (-4.5)-4.5 to fit with our approach.

After the initial valence is assigned to the words from the tweets, we started with the process of identifying the grammatical dependencies and modifying the initial valence. This was important for implementation of the rules for contextual valence shifters. The base attitudinal valence of a lexical item is modified by lexical and discourse context, so we need to implement predefined rules as valence context shifters [15]–[17]. The Stanford parser[7] [18] was utilized for performing the lexical dependency parsing and identifying grammatical relations between words. These dependencies were used to modify the initial valence of words with rules for contextual shifters, consisted of:

1) *Negatives (negation)* - If a word is in relation to negatives (e.g. not, never, nothing), then the initial valence of the word is shifted, i.e. is multiplied by -1.
2) *Intensifiers* - Intensifiers are adverbs or adverbial phrases that strengthen the meaning of other expressions and show emphasis. The words that are used in this paper as intensifiers are: deeply, always, absolutely, completely, extremely, highly, rather, really, so, too, totally, utterly, very, extraordinarily etc. If there is an intensifier in the tweet, then the valence of the word that is in relation to the intensifier is increased by multiplying the initial valence with 1,5.
3) *Mitigators (Downtoners)* - Mitigators or downtoners are words that reduce the force of another word or phrase in the sentence. The words that are used in this paper as mitigators are: fairly, somewhat, rather, quiet, lack, least, less, slightly, a little etc. If there is a mitigator in the tweet, then the valence of the word that is in relation to the mitigator is decreased by multiplying the initial valence with 0,5.
4) *Conjunctive adverbs* - A conjunctive adverb is a word that connects two sentences together, making a new sentence. It is like the word and, but it adds more meaning to the sentence. If there is a conjunctive adverb in the tweet, then the valences are neutralized by multiplication with 0.
5) *Negative words* - If the word is in a relation to a negative word, then it is multiplicated with -1 only if the word has positive valence. Otherwise, the valence remains the same.

After implementing these rules and calculating the valence of the words, we generate a vector of initial features. These features are valence and arousal values of every word appearing in every tweet. After removing the stop words using the set of stop words from NLTK library, we ended up with 8989 different words appearing in the tweets, so the vector was of length 17978, containing valence and arousal values for every word. The valence value was shifted by 10 units and the arousal value by 5 units, so these values are positive in every case. If some tweet doesn't include a particular word, then the valence and arousal values for this word are set to -1.

F. Feature Selection

Feature selection is a process for automatically selecting those features in the data that contribute most to the prediction variable [19]. To perform feature selection we are using the *sklearn.feature_selection*[8] module from the Scikit Learn Python Library [20]. The choice of attributes is one of the most significant processes for reducing the dimensionality of data by ranking all possible attributes and selecting those with the highest value.

There are three general classes of feature selection algorithms: *filter methods*, *wrapper methods* and *embedded methods*. Tree-based estimators can be used to calculate features importance, which in turn can be used to dismiss unnecessary features. We are using this kind of estimator to compute importance of every valence feature and retain the most ranked words according to their importance. After the ranking of the words, we defined a threshold of 0.0005 and kept the words that have importance value above this threshold. The number of words that we had after the elimination was 290. Then for these words, we kept the valence and arousal features, and so we ended up with 580 features plus the class. Fig. 4 displays the first 25 top-ranked words with their corresponding importance.

G. Training The Model

After the feature selection, we have feature vector for every tweet. The feature vector is of length 581 including the class (290 valence features and 290 arousal features to the corresponding words). The next step is the training of the model for classification of emotions. In this paper, we will use Weka, a software for knowledge mining, for building the model [21]. To test the built model, we decided to use cross-validation with 10 folds.

For the classification, we used four different classifiers including, Linear SVM (Support Vector Machines), Multilayer perceptron with one hidden layer (approximate sigmoid as activation function and squared error as loss function), Random Forest and LDA. We decided to work with these classifiers because they are capable to handle high dimensional feature vectors and obtain good performance with them.

III. Evaluation and Results

The results obtained from the evaluation of the hybrid approach with the dataset are given in this section. As mentioned in the previous part, we are going to test the models with cross-validation with 10 folds. After the testing, we are evaluating the models and comparing their performances. The metrics that we use to evaluate our models are Precision, Recall, and F-measure. In Table II we can see the number of correctly and incorrectly classified instances of the models. In Table III-V are presented the results from the model evaluation according

[7]https://nlp.stanford.edu/software/lex-parser.shtml

[8]http://scikit-learn.org/stable/modules/feature_selection.html

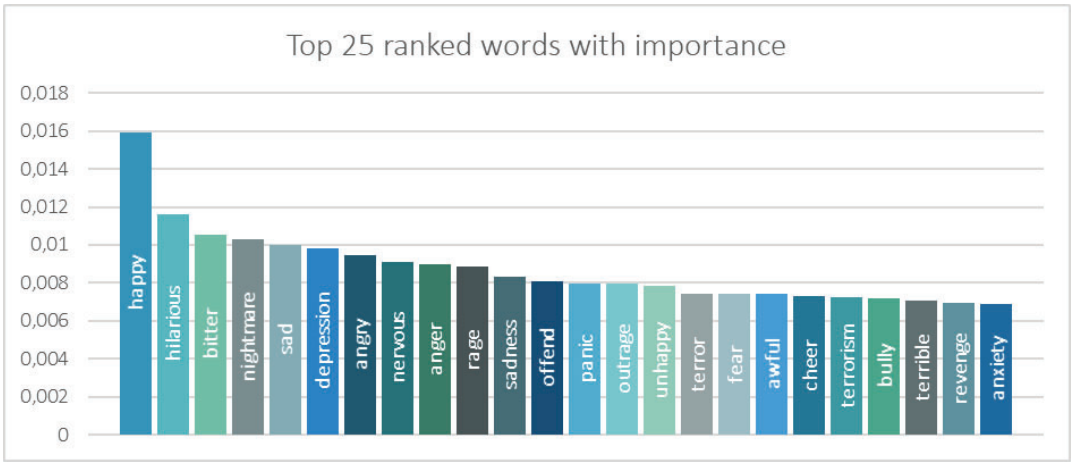


Fig. 4. Top 25 ranked words as features.

to the metrics. Additionally, in Fig. 5 we present a visual representation of the confusion matrices of every model.

TABLE II
NUMBER OF CORRECTLY AND INCORRECTLY CLASSIFIED INSTANCES OF THE MODELS

	Correctly classified instances	Incorrectly classified instances
Linear SVM	4415	933
Random Forest	4358	990
LDA	4453	895
Multilayer perceptron	4231	1117

TABLE III
PRECISION METRIC OF THE MODELS

	Joy	Neutral	Anger	Fear	Sadness
Linear SVM	0.927	0.805	0.884	0.729	0.840
Random Forest	0.880	0.845	0.844	**0.769**	0.784
LDA	**0.971**	**1.000**	**0.935**	0.703	**0.844**
Multilayer perceptron	0.849	0.000	0.814	0.752	0.768

TABLE IV
RECALL METRIC OF THE MODELS

	Joy	Neutral	Anger	Fear	Sadness
Linear SVM	0.836	0.783	0.793	0.889	0.766
Random Forest	**0.882**	**0.790**	**0.799**	0.843	0.726
LDA	0.833	0.710	0.781	**0.916**	**0.786**
Multilayer perceptron	0.856	0.000	0.784	0.834	0.768

TABLE V
F-MEASURE OF THE MODELS

	Joy	Neutral	Anger	Fear	Sadness
Linear SVM	0.879	0.815	0.836	0.801	0.801
Random Forest	0.881	0.816	0.821	**0.804**	0.754
LDA	**0.897**	**0.831**	**0.851**	0.795	**0.814**
Multilayer perceptron	0.853	0.000	0.799	0.791	0.768

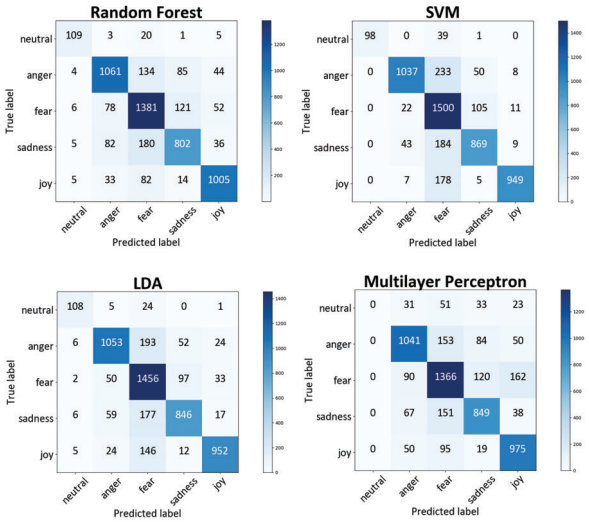


Fig. 5. Confusion matrices of the models.

Generally, all the classifiers have a problem with classifying anger as fear furthermore, fear as sadness and vice versa. Also, surprisingly for us, there is a problem when the models classify joy as fear. This could be due to the dominance of the fear class in the dataset.

With the results obtained by the measures, we can conclude that in general SVM, Random Forest and LDA have approximately equivalent performances. Random Forest has the best performance for classification of the neutral class, and all the models have a difficulty in separating the negative emotions especially anger and fear. From all the classifiers LDA has the highest overall accuracy. Furthermore, the Multilayer perceptron model ignores the neutral class and has the lowest performance from all the models.

IV. Conclusions and Future Work

Emotion detection is one of the most attractive topics of research and experimentation in recent times. Although the definition of emotion is fuzzy, it is still evident that it has been recognized and proven that emotions affect people, especially in their reasoning and decision making about their actions.

In this paper, we have studied the problem of emotion detection from Twitter posts known as tweets. We present a hybrid approach which combines elements from the big fields Natural Language Processing and Machine Learning to classify emotions. The first step in our approach is the preprocessing of the tweets. This is a key step that cannot be excluded, especially because tweets are written with the casual language used in Twitter and so have misspellings and incorrect word writings. Also, tweets have their unique characteristics like usernames and hashtags that need to be handled. After the preprocessing, the feature extraction was made with several concepts from NLP, and then the dimension of the created vector with initial features was reduced with feature selection. For training a model and its evaluation we used four ML classifiers: Linear SVM (Support Vector Machines), Multilayer perceptron, Random Forest and LDA, and compared their performance. The results show that in general SVM, Random Forest and LDA have approximately equivalent results, and LDA has the highest overall accuracy. On the other hand, the Multilayer perceptron model ignores the neutral class and has the lowest performance.

We consider that this approach can be improved if we take the emojis and emoticons into consideration. Also, because tweets have extremely irregular language we can further explore approaches for fixing these irregularities. Another thing that could be applied to improve the performance is to split the hashtags. Hashtags are not always consisted of one word. They can have multiple joint words in them that need to be separated. We take the hashtags, that may consist multiple words, for calculating our features and by separating them we could end up with words and features that are important for the emotion of the tweet.

References

[1] M. Thelwall, D. Wilkinson, and S. Uppal, "Data mining emotion in social network communication: Gender differences in myspace," *Journal of the Association for Information Science and Technology*, vol. 61, no. 1, pp. 190–199, 2010.

[2] L. Canales and P. Martínez-Barco, "Emotion detection from text: A survey," in *Proceedings of the Workshop on Natural Language Processing in the 5th Information Systems Research Working Days (JISIC)*, 2014, pp. 37–43.

[3] P. Ekman, "An argument for basic emotions," *Cognition & emotion*, vol. 6, no. 3-4, pp. 169–200, 1992.

[4] J. A. Russell, "A circumplex model of affect," *Journal of personality and social psychology*, vol. 39, no. 6, p. 1161, 1980.

[5] S. Buechel and U. Hahn, "Emotion analysis as a regression problem-dimensional models and their implications on emotion representation and metrical evaluation," in *ECAI*, 2016, pp. 1114–1122.

[6] S. M. Mohammad and F. Bravo-Marquez, "Wassa-2017 shared task on emotion intensity," *arXiv preprint arXiv:1708.03700*, 2017.

[7] A. Go, R. Bhayani, and L. Huang, "Twitter sentiment classification using distant supervision," *CS224N Project Report, Stanford*, vol. 1, no. 12, 2009.

[8] F. Vaassen, "Measuring emotion. exploring the feasibility of automatically classifying emotional text," Ph.D. dissertation, Dissertation, Antwerpen. http://www. cnts. ua. ac. be/sites/default/files/frederikvaassen_phdpreprint. pdf, 2014.

[9] A. B. Warriner, V. Kuperman, and M. Brysbaert, "Norms of valence, arousal, and dominance for 13,915 english lemmas," *Behavior research methods*, vol. 45, no. 4, pp. 1191–1207, 2013.

[10] A. Mehrabian, "Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in temperament," *Current Psychology*, vol. 14, no. 4, pp. 261–292, 1996.

[11] (2017) Pandas, python data analysis library. [Online]. Available: http://pandas.pydata.org

[12] C. D. Manning, P. Raghavan, H. Schütze *et al.*, *Introduction to information retrieval*. Cambridge university press Cambridge, 2008, vol. 1, no. 1.

[13] (2017) Nltk, natural language toolkit. [Online]. Available: http://www.nltk.org

[14] D. Jurafsky and J. H. Martin, *Speech and language processing*. Pearson London, 2014, vol. 3.

[15] L. Polanyi and A. Zaenen, "Contextual valence shifters," in *Computing attitude and affect in text: Theory and applications*. Springer, 2006, pp. 1–10.

[16] S. Gievska, K. Koroveshovski, and T. Chavdarova, "A hybrid approach for emotion detection in support of affective interaction," in *Data Mining Workshop (ICDMW), 2014 IEEE International Conference on*. IEEE, 2014, pp. 352–359.

[17] K. Koroveshovski, "System for emotion recognition in real time based on user written text analysis," Master's thesis, University of Ss. Cyril and Methodius, Skopje, Republic of Macedonia, 2014.

[18] M.-C. De Marneffe and C. D. Manning, "Stanford typed dependencies manual," Technical report, Stanford University, Tech. Rep., 2008.

[19] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *Journal of machine learning research*, vol. 3, no. Mar, pp. 1157–1182, 2003.

[20] (2017) Scikit-learn, python machine learning library. [Online]. Available: http://scikit-learn.org/stable/index.html

[21] (2017) Weka 3: Data mining software in java. [Online]. Available: http://www.cs.waikato.ac.nz/ml/weka

[22] S. M. Mohammad and F. Bravo-Marquez, "Emotion intensities in tweets," *arXiv preprint arXiv:1708.03696*, 2017.

[23] D. Preoţiuc-Pietro, H. A. Schwartz, G. Park, J. Eichstaedt, M. Kern, L. Ungar, and E. Shulman, "Modelling valence and arousal in facebook posts," in *Proceedings of the 7th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, 2016, pp. 9–15.

[24] J. Posner, J. A. Russell, and B. S. Peterson, "The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology," *Development and psychopathology*, vol. 17, no. 3, pp. 715–734, 2005.

[25] M. Taboada, J. Brooke, M. Tofiloski, K. Voll, and M. Stede, "Lexicon-based methods for sentiment analysis," *Computational linguistics*, vol. 37, no. 2, pp. 267–307, 2011.

[26] P. D. Turney and M. L. Littman, "Measuring praise and criticism: Inference of semantic orientation from association," *ACM Transactions on Information Systems (TOIS)*, vol. 21, no. 4, pp. 315–346, 2003.

[27] S. Baccianella, A. Esuli, and F. Sebastiani, "Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining," in *LREC*, vol. 10, no. 2010, 2010, pp. 2200–2204.

Discovering API Related Functions with Spectral Clustering

Martina Toshevska
Faculty of Computer Science and Engineering
Ss. Cyril and Methodius University
Skopje, Macedonia
martina.tosevska.95@gmail.com

Slobodan Kalajdziski
Faculty of Computer Science and Engineering
Ss. Cyril and Methodius University
Skopje, Macedonia
slobodan.kalajdziski@finki.ukim.mk

***Abstract*—Recommending API related functions is a problem of finding related functions for a particular function. Two functions are related if they are used together in specific user scenarios, appear in cross-references in "see also" sections in the API documentation, are called by the same functions etc. For the purpose of this paper, two functions are considered to be related if the data they access overlap.**

In this paper, an algorithm for discovering related functions based on structural information is proposed. It introduces few changes to the existing tool, *ALTAIR*. This algorithm is divided into different stages including augmented access graph creation from the source code, converting it to a bipartite graph, overlap matrix computation, and spectral clustering. Functions belonging to the same cluster are considered to be related.

The algorithm is applied to *Apache HTTP Server* and is evaluated on *Apache Portability Runtime (APR)*. From the API, *5,524* functions, *744* types and *365,250* function $\Rightarrow$ type relations were extracted. Experiments with different number of clusters are performed. For each experiment, evaluation metrics at different recommendation set size cutoff are computed. We used F1-measure as an evaluation metric. It is computed for each function from APR. The highest average F1-measure is *0.114* and it is achieved when partitioning into *250* clusters at recommendation set size of *20*. This algorithm is compared with an algorithm based on random walks (*FRAN*). According to the *F1-measure*, *spectral clustering* outperforms *FRAN* for *53%* of evaluation functions.

***Keywords*—spectral clustering, function recommendation, graph analysis, API analysis**

I. Introduction

Software systems are complex systems consisted of hundreds or even thousands of functions. These systems are provided as Application Programming Interfaces (APIs) to developers. Instead of writing code from the very beginning, developers often use API functions in their software systems. In order to provide better experience and decrease the time for development process, APIs need to be well documented.

Many APIs are composed of functions grouped by a specific concept i.e. are related. Related functions may refer to functions which are used together in specific user scenarios, functions that call or are called by the same functions, functions that access the same data, functions with similar purpose etc. For example, functions apr_pool_create and apr_pool_clear, from Apache Portability Runtime, both operate with memory pools. Functions apr_file_open and apr_file_close are frequently called together since working with files requires opening the file, operating with it and then closing the file. Consequently, we can infer that these functions are someway related. These associations between functions are placed in cross-reference sections (i.e. "see also") in the software code documentations.

When developing an API, developers can annotate related functions using specific syntax leading to easier cross-references generation. However, tracking all related functions is generally difficult with the API evolving and growing. This is where artificial intelligence comes to help. Discovering related functions can be modeled as a machine learning problem: ***given a particular function, find all related functions***.

An example of an API is Apache HTTP Server[1]. It is an open-source HTTP server for modern operating systems including UNIX and Windows. Its source code is fairly well documented, but, as stated in [1], only a small portion of all functions have cross-referencing annotations. Tools for automatically discovering related functions can be used to fill missing cross-references.

There are different approaches for addressing the problem of automatically discovering related functions, separated into two different groups. Approaches that consider how API functions are used are in one group, while those considering how API functions are implemented belong to another group. In this paper we use an approach that takes into account how API functions are implemented. We consider two functions as related if they share state and access the same data. The main purpose is to discover related functions in version 2.2.31 of Apache HTTP Server using the methodology from ALTAIR [1] with few changes.

The rest of the paper is organized as follows. In Section II we provide an overview of the related work. The methodology is described in detail in Section III. In Section IV we describe different experiments for finding related functions and show experimental results, and then conclude in Section V.

II. Related Work

Analysis of the software code has been widely used in the fields of machine learning, artificial intelligence, data mining

[1]https://httpd.apache.org/ (last visited: 22.02.2018)

etc. One application is sample code generation. DeepAPI [2] uses deep learning approaches for generating usage examples from natural language queries. An algorithm for mining and synthesizing succinct and representative human-readable documentation of program interfaces is presented in [3]. It is based on a combination of path sensitive dataflow analysis, clustering, and pattern abstraction. APIMiner 2.0 [4] relies on association rules algorithms to extract usage patterns and to include more useful examples.

There is large amount of work for the task of finding related functions. Mining frequent usage patterns in client code is one way to discover related API functions [5] [6] [7] [8]. The FRIAR [5] algorithm uses sets of functions that were commonly called together to predict functions related to a particular query function. PR-Miner [6] automatically extracts general programming rules from software code using data mining techniques. XSnippet [9] and Strathcona [10] find relevant code examples from a repository, which focus on how to accomplish a specific task composing several API functions from existing examples.

All of the previously discussed algorithms work with client code. This approach might be sensitive to the availability of the client code or the way query function is used in specific client code. If the client code calls neither the query function nor the functions related to it, it could not find the corresponding patterns, which leads to missing and unreliable results. In situations when client code is unavailable (for example when the API is relatively new), algorithms relying on it cannot be applied.

Unlike operating with client code, structural approaches work with the structure of the API source code. Suade [11] uses a technique based on an analysis of the topology of structural dependencies in a program to automatically propose and rank program elements that are potentially interesting to a developer investigating source code. The FRAN (Finding with RANdom walks) [5] algorithm recommends API functions using random walks on call graphs. This algorithm returns ranked list of relevant functions for a given function. FRAN uses the call graph structure for creating four different function sets: parent (callers of the query function), child (callees of the query function), sibling (union of functions called by the functions in the parent set) and spouse (functions which call the functions in the child set). Next, FRAN applies HITS algorithm on the subgraph formed by these sets and uses authority to rank the functions. Evaluation results of this algorithm are compared with the evaluation results of the algorithm proposed in this paper. In [12] an approach for clustering object-oriented software system using spectral graph partitioning is proposed. Software system is represented as a graph in which nodes stand for the classes and the edges stand for the discrete messages exchanged between the classes.

In this paper we propose an algorithm for discovering related API functions based on structural approach, as defined in ALTAIR [1]. This method introduces few changes in the methodology. The details of the algorithm, as well as differences from ALTAIR are described in the following section.

III. Methodology

Two functions are related if the data they access overlap. The information about functions and the data they access is extracted from the source code and modeled as an augmented access graph. Next, the graph is converted into a bipartite graph. Hence, based on overlap coefficient, functions are partitioned into a set of clusters using spectral clustering. Functions belonging to the same cluster are considered to be related i.e. similar to each other. Each of these steps is explained in detail in the following subsections.

A. Augmented Access Graph

Before applying any technique for information extraction from source code, preprocessing is required. Preprocessing for this task is denoted as modifying the source code. Macro definitions are a problem for further code analysis. Therefore, all keywords representing macros are removed in the preprocessing step.

Augmented access graph is a directed heterogeneous graph consisted of different types of edges and nodes. The graph used in [1] contains three types of nodes (*functions*, *data types* and *composite types*) and four types of edges (*access*, *call*, *allocation* and *composite*). The graph used in this project has an additional *return* edge, ending with five different types of edges:

- Access (function $\Rightarrow$ data type): f loads from/stores to z
- Call (function $\Rightarrow$ function): f calls g
- Allocation (function $\Rightarrow$ composite type: f (de)allocates t
- Composite (composite type $\Rightarrow$ data type): t has a field z
- Return (function $\Rightarrow$ data type): f returns z

An example augmented access graph is represented in Fig. 1. The construction of the graph works as follows. It is extracted from XML representation of the source code. Such representation for the source code is created using srcML[2] [13]. After that, necessary parts from this representation are extracted using XPath and XQuery. For each load/store statement, e.g., one in function f that accesses data x of composite type A, a corresponding access edge from f to x is added, as well as a composite edge from A to x. For each call statement, e.g., g calls g_0, a corresponding call edge from g to g_0 is added. ALTAIR adds call edge only if the callee is private. On the contrary, here, call edge is added for each call statement (for both private and public callee functions). If function e invokes an allocation function for type A, a corresponding allocation edge from e to A is added. For each return statement, e.g., h returns data w, a corresponding return edge from h to w is added.

B. Bipartite Access Graph

Bipartite graph is a graph which nodes can be divided into two disjoint sets. For the purpose of this project, the augmented access graph is converted into a bipartite access graph containing only two types of nodes: *functions* and *data*

[2]http://www.srcml.org/ (last visited: 22.02.2018)

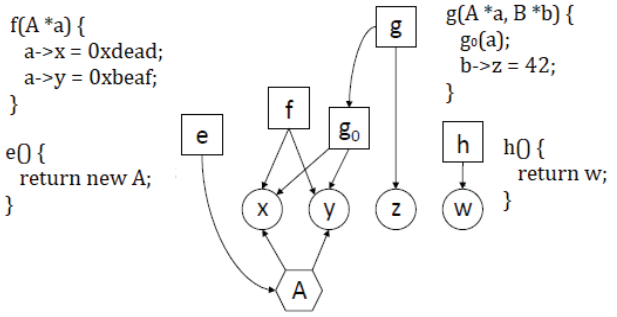


Fig. 1. Example of an augmented access graph [1] where e, f, g, g_0, h represent functions, x, y, z, w represent data, and A represents a composite type (definition for h is modified).

types, and one type of edge: *access* edge (function $\Rightarrow$ data type). An example bipartite access graph is shown in Fig. 2.

When g calls g_0, it implicitly accesses the data that g_0 accesses. Therefore, the effect of g_0 (i.e. the set of data it accesses) should be merged to g. Allocation functions (such as $malloc$ and $free$ in C) do not explicitly touch any field of a given object, but they do affect all fields since they behave as constructors and destructors. For example, function e implicitly accesses fields of A. Thus, all fields from A should be added to the set of fields e accesses.

One approach for determining the set of data each function may access is by computing transitive closure (graph reachability) [14] [15]. The graph created with transitive closure has the same nodes as the original graph. If a path between two nodes in the original graph exists, an edge is added in the new graph. For this project, bipartite graph is created using the algorithm for transitive closure[3] in NetworkX [16].

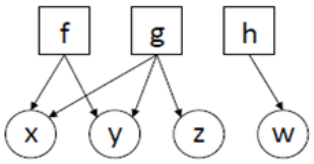


Fig. 2. Example of a bipartite access graph [1]. f, g, h represent functions, while x, y, z, w represent data. A directed edge from f to x means that function f may access some data x.

C. Overlap Coefficient

Overlap measure of a function f with function g is the proportion of f's data that is shared with g. The overlap with function g, for a function f, is computed as:

$$\pi(g|f) = \frac{|\mathcal{N}(f) \cap \mathcal{N}(g)|}{|\mathcal{N}(f)|} \tag{1}$$

where $\mathcal{N}(f)$ and $\mathcal{N}(g)$ are the set of data types accessed by functions f and g respectively. Note here that $\pi(f|g)$ and $\pi(g|f)$ are not the same. For the example in Fig. 2, $\pi(f|g) = 2/3$ and $\pi(g|f) = 1$. The overlap coefficient, which is symmetric variant of overlap measure, between two functions f and g measures how the two functions share in common. It is defined as follows:

$$\begin{aligned}\pi(f,g) &= \frac{|\mathcal{N}(f) \cap \mathcal{N}(g)|}{min(|\mathcal{N}(f)|, |\mathcal{N}(g)|)} \\ &= max(\pi(f|g), \pi(g|f))\end{aligned} \tag{2}$$

where $\mathcal{N}(f)$ and $\mathcal{N}(g)$ are the set of data types accessed by functions f and g respectively. For the example in Fig. 2, $\pi(f,g) = max(\pi(f|g), \pi(g|f)) = max(2/3, 1) = 1$. Overlap coefficient can be easily calculated from the bipartite access graph.

D. Spectral Clustering

For the task of partitioning the functions into modules, Ng-Jordan-Weiss [17] algorithm for spectral clustering is used. Steps of the algorithm are shown in Fig. 3.

The overlap matrix (affinity matrix) constructed of overlap coefficients over all functions is defined as:

$$\Pi = \begin{cases} \Pi_{fg} = \pi(f,g) \\ \Pi_{ff} = 0 \end{cases} \tag{3}$$

In this matrix, all diagonal elements are zero. Let D be the diagonal matrix of Π. We define symmetric normalized Laplacian matrix (L^*) as:

$$L^* = I - L \qquad L = D^{-\frac{1}{2}} \Pi D^{-\frac{1}{2}} \tag{4}$$

Since both matrices L and L^* have the same eigenvectors, matrix L is used in the subsequent steps. First, a $n \times k$ matrix $X = (\mu_1, \ldots, \mu_k)$ formed by the first k eigenvectors of L is calculated. New matrix, Y, is computed by normalizing each row of X, where $||Y_i|| = 1$ for each row i. Then, a classical k-means algorithm [18] is applied to group the n vectors into k clusters, and accordingly cluster the n functions into k modules.

Another difference from the methodology used in [1] is the number of clusters (i.e. modules). ALTAIR applies recursive bi-partitioning [19] according to the second smallest eigenvalue to determine the optimal number of modules. On the contrary, we compare results for different values of k (number of clusters).

1: $D_{ff} \leftarrow \sum_g \Pi_{fg}$, given overlap matrix Π.
2: $L \leftarrow D^{-\frac{1}{2}} \Pi D^{-\frac{1}{2}}$.
3: Compute eigenvectors μ_k of L.
4: $X \leftarrow (\mu_1, \ldots, \mu_k)$.
5: Normalize X to Y, where $Y_i = X_i / ||X_i||_2$.
6: Apply k-means to cluster Y.
7: Cluster Π_f into k modules accordingly.

Fig. 3. Steps for k-module partitioning [1].

[3] https://networkx.github.io/documentation/networkx-1.11/reference/generated/networkx.algorithms.dag.transitive_closure.html (last visited: 22.02.2018)

E. Evaluation

In information retrieval, three measures of performance are used repeatedly: precision, recall and the F1-measure. For the purpose of this paper, we define these measures as follows. Precision is the fraction of relevant functions among the retrieved functions. Recall is the fraction of relevant functions that have been retrieved over the total amount of relevant functions. Precision and recall are calculated as follows:

$$precision = \frac{|rel \cap ret|}{|ret|} \qquad recall = \frac{|rel \cap ret|}{|rel|} \tag{5}$$

where ret (for a specific function) is the set of retrieved functions i.e. functions belonging to the same cluster (determined by the spectral clustering algorithm) and rel (for a specific function) is the set of relevant functions i.e. functions belonging to the same APR module (explained in the next section). The F1-measure is the equally-weighted harmonic mean of the recall and precision measures. It is defined as:

$$F1 = \frac{2 * precision * recall}{precision + recall} \tag{6}$$

We use F1-measure as an evaluation metric for comparison between our algorithm and FRAN algorithm.

IV. Experimental Results

The proposed algorithm was applied to version 2.2.31 of Apache HTTP Server. Its source code is written in the programming language C. For the source code parsing we used srcML. In this step, we had to face a problem implied by errors made by this tool. When a function is defined as a macro, srcML parses it as two separate macros, where the signature is one macro and the body is another. In the resulting XML representation of the source code, these two macros are not related. But, the second macro is the body of a function which signature is represented by the first macro. Information for these functions was implicitly removed. However, we do not want to miss this information. A solution is to remove all keywords representing macros in the source code before creating XML representation.

From the source code we extracted 7,870 functions and 1,184 types. For the relations we obtained 22,553 function calls, 156 type allocations, 12,318 accesses to types, 2,708 type compositions and 5,534 return types. Therefore, the augmented access graph was composed of 9,054 nodes and 41,938 edges. This values are shown in Table I (middle column). Note here that a function f, for example, can allocate a specific type t and then return the same type. Hence, the graph has two edges between function f and type t. The number of edges between two nodes does not play a significant role for transitive closure computation. As a result, for a pair of nodes only one edge is added. This is the reason why total number of edges of the augmented access graph is not equal to the sum of different edge types.

As a result of techniques for graph creation, graph contained different nodes representing the same data. For example, *longint* and *long int* are characterized as different data although they are the same. Consider two functions, f which in the bipartite access graph has relation to *longint* and g which has relation to *long int*. Their overlap coefficient will be 0, while the real value should be 1. Note that *longint* is not used as type in the source code, it appears during the graph creation. Another example are types such as *char* and *const char*. Both are representing the type *char*, except that the second indicates an immutable variable. However, they are declared as two different nodes in the graph. Their influence when calculating overlap between functions is similar as in the previous example. All these issues need to be solved before proceeding with further steps. As a solution, we modified such erroneous types. We converted *longint* to *long int*, *const char* to *char* etc.

TABLE I
Graph Statistics for Augmented Access Graph and Bipartite Access Graph. Total Number of Edges of the Augmented Access Graph is not Equal to the Sum of Different Edge Types Since for a Pair of Nodes Only One Edge is Added Even if There are More.

	Augmented access graph	**Bipartite access graph**
Nodes	9,054	6,268
Functions	7,870	5,524
Types	1,184	744
Edges	41,938	365,250
Access	12,318	365,250
Call	22,553	/
Allocation	156	/
Composite	2,708	/
Return	5,534	/

After computing transitive closure, the number of nodes decreased, while the number of edges increased. The bipartite access graph was consisted of 365,250 edges and 6,268 nodes (Table I (right column)). For each pair of functions in the bipartite graph we computed overlap rank and created the overlap matrix. Because the graph contained 5,524 functions, the matrix was of size 5,524 x 5,524.

The overlap matrix had zero rows (i.e. rows where all entries are 0). The explanation is that a function associated with such row does not overlap with any other function. When calculating the matrix L, this produces division by zero problem which results in having NaN entries. Those entries are a problem for further analysis. One solution to this problem is to set all NaN entries to 0. Another possible solution is to divide with a number close to 0, for example 10^{-10}. The first approach achieved better evaluation results and therefore is used as a solution for this issue.

We performed 5 experiments with different number of clusters (also number of eigenvectors). The affinity (overlap) matrix is the same for all experiments. For a specific experiment, number of used eigenvectors is equal to the number of clusters. To illustrate, for 250 clusters, 250 eigenvectors are used. Thus, we have n x 250 matrix of eigenvectors where n is the number of functions. Since we deal with 5,524 functions, this matrix size is 5,524 x 250.

Apache has a portability layer[4], consisted of 32 separate groupings of 330 functions, or portability layer modules. It is a low-level API called Apache Portable Runtime (APR) that the Apache HTTP Server is built on top of. Most APR function names are with prefix "apr_". Each module incorporates a closely related set of functions that perform different tasks such as file and socket operations, memory pool and thread management, locking operations etc. These modules are used for evaluation. Functions belonging in the same module are considered as related.

For each function f in the APR, functions from the same module are taken as relevant, while functions from the same cluster are taken as retrieved set. Retrieved functions i.e. functions from the same cluster as f are ranked according to the overlap coefficient. Then, F1-measure is calculated at five recommendation set size cutoffs (top-5, top-10, top-15, top-20 and top-25). The highest F1-measure is 1 and is achieved at size 5 cutoff for 200 clusters. The second highest f1-measure is 0.67. It is achieved for both recommendation set size 10 for 250 clusters and recommendation set size 5 for 300 clusters. Minimum and maximum values for each pair of recommendation set size cutoff and number of clusters are displayed in Table II and Table III, respectively. Because minimum value for each pair is 0 these values are omitted and second smallest values are shown. These values are 0.03 for recommendation set size 5 and 10, and 0.04 for recommendation set size 15, 20 and 25.

TABLE II
MAXIMUM F1-MEASURE FOR EACH PAIR OF RECOMMENDATION SET SIZE CUTOFF AND NUMBER OF CLUSTERS.

Maximum F1-measure		**Recommendation set size cutoff**				
		5	**10**	**15**	**20**	**25**
Number of clusters	**100**	0.50	0.59	0.45	0.50	0.42
	150	0.60	0.59	0.50	0.50	0.42
	200	1.00	0.67	0.63	0.63	0.63
	250	0.60	0.59	0.52	0.43	0.36
	300	0.67	0.59	0.48	0.38	0.36

TABLE III
MINIMUM F1-MEASURE FOR EACH PAIR OF RECOMMENDATION SET SIZE CUTOFF AND NUMBER OF CLUSTERS. SINCE MINIMUM VALUE IS ALWAYS 0, THESE MINIMUM VALUES REPRESENT SECOND SMALLEST VALUES.

Minimum F1-measure		**Recommendation set size cutoff**				
		5	**10**	**15**	**20**	**25**
Number of clusters	**100**	0.04	0.04	0.03	0.03	0.03
	150	0.04	0.04	0.03	0.03	0.03
	200	0.04	0.04	0.03	0.03	0.03
	250	0.04	0.04	0.03	0.03	0.03
	300	0.04	0.04	0.03	0.03	0.03

Average F1-measure for different number of clusters and different recommendation set size cutoffs is shown in Fig. 4, from where we can conclude that partitioning into 250 modules achieves best result. It can be considered as (local) maximum. F1-measure increases as the number of clusters increases up to 250, and decreases when the number of clusters is above 250.

For the recommendation set size cutoffs, we can infer that best results are obtained using top-15 and top-20 functions for all values for number of clusters. Average F1-measure, for 250 clusters, at 15 and 20 recommendation set size cutoff is 0.113 and 0.114 respectively. The second value (0.114) is the highest average F1-measure among all experiments. F1-measure when taking only top-5 functions is significantly lower than others. APR contains modules with much more than 5 functions, which means that functions have more than 5 related functions. Consequently, by considering only top-5 functions, we miss more relevant functions. From the other side, for functions with less than 25 related functions, when using top-25 functions we take into account functions that are not relevant.

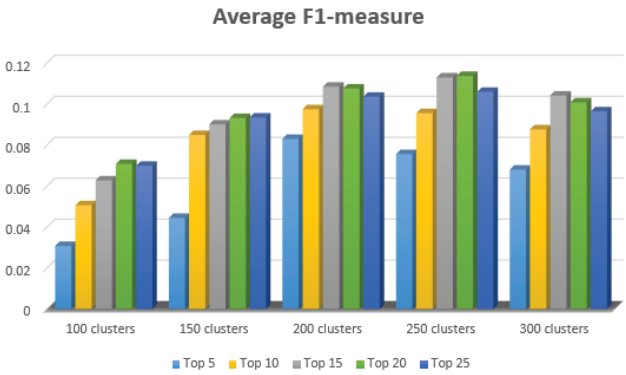


Fig. 4. Average F1-measure for different number of clusters and different recommendation set size cutoffs.

Next, these results are compared to FRAN [5]. Such comparison is reasonable since both algorithms are based on structural approach and operate with API source code. The two algorithms need to be evaluated on the same evaluation data i.e. the same API and the same version. The FRAN algorithm was compiled from its source code and executed on version 2.2.31 of Apache HTTP Server. The comparison between FRAN and spectral clustering with 250 clusters and recommendation set size of 20 is displayed in Fig. 5. This is the result for which spectral clustering algorithm achieves highest average F1-measure (0.114). To save space, other comparison results are omitted.

The graph is separated into three parts. In the first part are functions for which spectral clustering achieves higher value for F1-measure than FRAN. Functions with the same F1-measure for both FRAN and spectral clustering are in the second part. The last part contains functions for which FRAN achieves better results. From the graph, we can conclude that for 175 out of 330 functions (53%) spectral clustering outperforms FRAN and for 114 out of 330 functions (34%) FRAN outperforms spectral clustering. For 13% (41 out of 330) of evaluation functions, both algorithms achieved the same F1-measure.

[4]http://apr.apache.org/docs/apr/0.9/index.html (last visited: 22.02.2018)

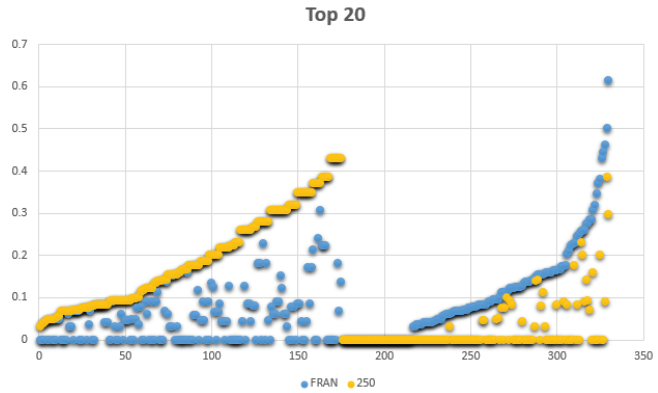


Fig. 5. Comparison of spectral clustering with 250 clusters and FRAN. Vertical axis represents F1-measure. Horizontal axis represents functions in the APR.

V. Conclusion

Discovering API related functions is a problem of finding related functions for a particular function. Two functions are related if they are used together in specific user scenarios, appear in cross-references in "see also" sections in the API documentation, call or are called by the same functions etc. We consider two functions as related if the data they access overlap.

There are different approaches for addressing the problem of automatically discovering related functions, separated into two different groups. Approaches that consider how API functions are used are in one group. The algorithms from this group work with client code. Algorithms considering how API functions are implemented belong to another group. These algorithms work with the API source code and can be applied even when there is no client code for the API, for example when the API is relatively new, which is not the case with algorithms from the first group.

In this paper, we proposed an algorithm for discovering related functions based on structural information. It introduced few changes to the existing tool, ALTAIR. This algorithm is divided into different stages including augmented access graph creation from the API source code, converting it to a bipartite graph, computing overlap matrix based on overlap coefficient, and spectral clustering. Functions belonging to the same cluster are considered to be related.

We applied the algorithm to Apache HTTP Server and evaluated it on Apache Portability Runtime (APR). We performed experiments with different number of clusters and computed evaluation metrics (precision, recall and F1-measure) at different recommendation set size cutoffs. We compared our algorithm (spectral clustering) with an algorithm based on random walks on call graphs, FRAN. According to the F1-measure, spectral clustering outperforms FRAN for a little more than one half (53%) of evaluation functions. The highest average F1-measure is 0.114 and it is achieved when partitioning into 250 clusters at recommendation set size of 20.

The tool used for creating XML representation of the source code, srcML, is susceptible to errors. One error is incorrectly parsing functions defined as macros or macro calls. We found a possible solution to this problem, but this does not guarantee that the problem is resolved. In future, we can try other tools for source code parsing that may lead to more accurate representation of the source code as augmented access graph and possibly better evaluation results.

References

[1] F. Long, X. Wang, and Y. Cai, "Api hyperlinking via structural overlap," in *Proceedings of the the 7th joint meeting of the European software engineering conference and the ACM SIGSOFT symposium on The foundations of software engineering*, pp. 203–212, ACM, 2009.

[2] X. Gu, H. Zhang, D. Zhang, and S. Kim, "Deep api learning," in *Proceedings of the 2016 24th ACM SIGSOFT International Symposium on Foundations of Software Engineering*, pp. 631–642, ACM, 2016.

[3] R. P. Buse and W. Weimer, "Synthesizing api usage examples," in *Proceedings of the 34th International Conference on Software Engineering*, pp. 782–792, IEEE Press, 2012.

[4] S. B. Hudson, "Extracting examples for API usage patterns," Master's thesis, Federal University of Minas Gerais, Department of Computer Science and Engineering, 2014.

[5] Z. M. Saul, V. Filkov, P. Devanbu, and C. Bird, "Recommending random walks," in *Proceedings of the the 6th joint meeting of the European software engineering conference and the ACM SIGSOFT symposium on The foundations of software engineering*, pp. 15–24, ACM, 2007.

[6] Z. Li and Y. Zhou, "Pr-miner: automatically extracting implicit programming rules and detecting violations in large software code," in *ACM SIGSOFT Software Engineering Notes*, vol. 30, pp. 306–315, ACM, 2005.

[7] H. Zhong, T. Xie, L. Zhang, J. Pei, and H. Mei, "Mapo: Mining and recommending api usage patterns," in *European Conference on Object-Oriented Programming*, pp. 318–343, Springer, 2009.

[8] M. A. Saied, O. Benomar, H. Abdeen, and H. Sahraoui, "Mining multi-level api usage patterns," in *Software Analysis, Evolution and Reengineering (SANER), 2015 IEEE 22nd International Conference on*, pp. 23–32, IEEE, 2015.

[9] N. Sahavechaphan and K. Claypool, "Xsnippet: Mining for sample code," *ACM Sigplan Notices*, vol. 41, no. 10, pp. 413–430, 2006.

[10] R. Holmes and G. C. Murphy, "Using structural context to recommend source code examples," in *Software Engineering, 2005. ICSE 2005. Proceedings. 27th International Conference on*, pp. 117–125, IEEE, 2005.

[11] M. P. Robillard, "Automatic generation of suggestions for program investigation," in *ACM SIGSOFT Software Engineering Notes*, vol. 30, pp. 11–20, ACM, 2005.

[12] S. Xanthos and N. Goodwin, "Clustering object-oriented software systems using spectral graph partitioning," *Urbana*, vol. 51, no. 1, pp. 1–5, 2006.

[13] M. L. Collard, M. J. Decker, and J. I. Maletic, "srcml: An infrastructure for the exploration, analysis, and manipulation of source code: A tool demonstration," in *Software Maintenance (ICSM), 2013 29th IEEE International Conference on*, pp. 516–519, IEEE, 2013.

[14] P. Purdom, "A transitive closure algorithm," *BIT Numerical Mathematics*, vol. 10, no. 1, pp. 76–94, 1970.

[15] E. Nuutila, "Efficient transitive closure computation in large digraphs, mathematics and computing in engineering series no. 74 phd thesis helsinki university oftechnology," 1995.

[16] A. Hagberg, P. Swart, and D. S Chult, "Exploring network structure, dynamics, and function using networkx," tech. rep., Los Alamos National Lab.(LANL), Los Alamos, NM (United States), 2008.

[17] A. Y. Ng, M. I. Jordan, and Y. Weiss, "On spectral clustering: Analysis and an algorithm," in *Advances in neural information processing systems*, pp. 849–856, 2002.

[18] S. Lloyd, "Least squares quantization in pcm," *IEEE transactions on information theory*, vol. 28, no. 2, pp. 129–137, 1982.

[19] F. R. Chung, *Spectral graph theory*. No. 92, American Mathematical Soc., 1997.

Automatic Detection of Computational Complexity of Dynamic Programming Algorithms

Zorica Stefanovska
Faculty of Informatics
FON University
Skopje, Macedonia
zstefanovska@yahoo.com

Vekoslav Stefanovski
Sourcico Macedonia

Skopje, Macedonia
wekoslav@sourcico.com

***Abstract* — Exact determination of the computational complexity of a given algorithm is a difficult problem in computation. However, there are several classes of complexities that can be estimated and calculated empirically. This paper takes this approach to test several classic problems, which have dynamic programming solutions. By defining the metrics of the calculations needed in the problem itself, we sidestep the problem of unreliable environment inherent in the empirical determination of algorithm performance. The determined results empirically confirm the accepted theoretical complexity results.**

Keywords — complexity, empirical solution. dynamic programming, fibonacci, dijkstra

I. Introduction

Dynamic programming is first mentioned by Richard E. Bellman, when, in the early to mid-nineteen fifties he laid the foundation on a solid mathematical basis. The approach was revolutionary new, at the time, and contributed to the development of a new method for solving problems that have their own distinctive hierarchy: problems that contain simpler subproblems of the same type.

The process of solving the "main" problem begins with solving the simplest, even trivial, problems contained within it. The problem and the values of those solutions are stored in order to use them further up the pipeline, to obtain the solutions of the next level of problem, up to and including, the solution of the main problem. This method for solving problems is a powerful tool for designing appropriate algorithms. It is also known in the literature as a method of reverse induction, that is, the process of thinking in the reverse order (from the end to the beginning of the solution). The purpose of this method is to find an optimal sequence of steps to solve the main problem. To be able to apply a dynamic programming method to a problem, the problem should have an appropriate hierarchy, that is, it possesses two key features:

- Optimal substructure
- Overlapping sub-problems

One problem has an optimal substructure if the solution to this problem can be obtained through the solutions of its subproblems.

There are two approaches to solving dynamic programming problems: bottom-up and top-down.

A top-down approach uses a technique called memoisation. Memoisation is a notion typical of computer science and it is first defined by Donald Michie, 1968. Its purpose is to memorize the result of resolving a given subproblem, so when we need to re-solve that exact same subproblem, the solution will only be looked up in a table, instead of being resolved again. Most often, this technique is applied to recursive defined problems, by solving its subproblems (if these subroutines are repeated).

The bottom-up approach consists in formulating a complex problem as a series of recursive and simpler problems. This series of subproblems is designed in such a way that each subsequent solution is obtained as a combination of one or more previously solved subproblems. Therefore, based on the previously obtained subproblem solutions and their combinations, solutions are obtained for larger and larger subproblems until a solution for the initial problem is obtained. The solutions obtained in the meantime are stored in a table to avoid unnecessary repetition of an already solved problem.

In this paper, we'll compare several different solutions to common dynamic programming problems using an empirical measurement of their computational complexity.

II. Definition of a Problem and a Solution Algorithm

In principle, a single problem can be solved with several different algorithms. To make an accurate evaluation of the different algorithms that solve the same problem, we need to define an appropriate metric. A good metric on this issue is their time of execution, as well as the amount of resources that the algorithm uses. Usually, by resources we refer to the memory allocations that the algorithm needs. These measurements are direct and independent of the algorithm, because they measure externally observed effects of execution.

However, the environment in which the algorithm is executed can be variable during the execution of the algorithm, that is, the time and memory as consumable resources are variables dependent on the load on the system where the algorithm is being executed. Because of this dependence, most

modern systems consider this approach to be unreliable. An acceptable or standard solution for this performance measurement is to run the algorithm with the same initial state multiple times, and after the extreme values are disposed of, the average or median resource consumption value is taken as the valid value. While this approach is practical, it does not give explicit values for the complexity of a given algorithm, instead we get an approximate (empirical) value. We can get the information that an algorithm is faster than another algorithm for a given set of input data, but with this approach we will not get information on why this is so.

In this paper, the used approach is different from the default. While the time taken to execute an algorithm is measured, the main emphasis is placed on certain metrics defined by the problem itself. For example, if you need to compare in place array sorting algorithms, the algorithm (whatever and however it does the sorting) will have to perform some operations, in particular, it will have to access some element of the array, it will have to compare two elements from the array and will have to change the places of the two elements of the array. These operationscan be defined as metrics of the problem itself, and in this paper each algorithm for solving a problem, mustcalculate and submit values according to these metrics for analysis. In the execution of algorithms, they are executed many times, but not with the same set of input parameters. Input parameters are strictly ordered relative to some criterion, for example, to sort an array in place, the input parameter is a string of data ofincreasing length.

III. Implementation of the Definitions

The specific implementation of the described definitions is made through the TypeScript interfaces.

The problem definition (interface **ProblemOptions,** Fig. 1) consists of:

- The name of the problem
- The names of the metrics to be submitted by the algorithms that will solve this problem
- A series of progressively more complex values that will be used as input parameters for the algorithms that solve the problem. Two options are possible, namely to specify explicitly specific values, or to assign a function (*type InputFunction<T> = (index: number) => T;*) to which the sequence number of the execution will be provided, that will return the required input parameters.
- The number of tests to be performed (optional parameter if values are specified explicitly). If the parameter is not specified, the default value is 100.

```
interface ProblemOptions {
    name: string;
    metricNames: string[];
    inputs: any[] | InputFunction<any>;
    inputLength?: number;
}
```

Fig. 1

The algorithms themselves are defined using the interface shown in Fig. 2:

```
interface Algorithm {
    name:string;
    problemName:string;
    metrics:MetricMap;
    run(input:any):any;
    reset():void;
}

type MetricMap = { [key:string]:Metric };
type Metric = () => number;
```

Fig. 2

The **Algorithm** consists of the following fields:

- **name** is the name of the algorithm itself, and can be an arbitrary value, unless it is the same as the name of another algorithm that solves the same problem
- **problemName** is the name of the problem that this algorithm solves and must be the same with the name used to define the problem.
- **metrics** is a map of functions that can be used to retrieve the metrics after executing the algorithm for a given input value
- **run** is a function that will be used in the execution of the algorithm, and which from our perspective encapsulates the algorithm itself.
- **reset** is a function that will be called after the algorithm is executed and the metrics taken in order to re-initialize the algorithm to the initial state, and to ensure that the metrics cannot be confused between subsequent executions of the algorithm.

These definitions of problems and algorithms are used by the class **AlgorithmComparer** whose constructor accepts the definition of the problem as a parameter (Fig 3).

```
class AlgorithmComparer {
    ...
    constructor(privateoptions:ProblemOptions)
    {
        ...
```

Fig. 3

After we have defined the problem, we can register the algorithms that solve the problem, with the help of the **registerAlgorithm** function. When registering, the function runs the following checks:

- whether the algorithm solves the correct problem
- whether the algorithm exports any metrics
- whether the metrics that the algorithm exports fully correspond with the metrics required by the problem
- whether the name of the algorithm is unique to that problem

If these checks are successful, the algorithm is registered.

After the successful registration of all algorithms, we can trigger their execution using the **runAlgorithms** method, which successively loads or generates data, and then executes the algorithms. After the execution of a particular algorithm, its metrics are retrieved and stored in the results field.

IV. Computational Complexity and Its Evaluation

The simplest definition of complexity from the aspect of algorithms would be a simple relationship between the size of the input parameters of an algorithm and the resources needed to execute the algorithm, i.e. the required number of steps (time) or the required number of memory locations to perform.

In this relation, most often the only focus is on the size of the input itself, with all additive and even multiplicative constants being overlooked. The reason for this is that as a common case in the execution of algorithms, the size of their inputs can vary with several orders of magnitude. For example, one algorithm can be linearly dependent on the input, and require 1,000 steps for each input unit (for one input, performs 1,000 steps, for two inputs, perform 2,000, etc.). Let's compare this algorithm to another, which for each unit input, performs 1 step for the cube of the input size, i.e. for one input it performs 1 step, for two, it executes 8 steps, three inputs- 27 steps, etc.

For small numbers, the advantage in the number of steps is completely on the side of the second algorithm, but if the input increases by several lines of magnitude (which often happens in practical scenarios), the situation changes dramatically. For example, for entry 1,000, the first algorithm will execute 1,000,000 steps, while the second 1,000,000,000 steps, i.e. the first algorithm will be a thousand times faster than the other one.

The complexity of the algorithms is estimated in the asymptotic sense, i.e. which would be the nature of dependence if the size of the entrance tends to infinity, and the so-called: big-o, big-omega and big-theta notations are used. For example, it has been shown that the linear string search is proportional to the number of members of the array n, or has O(n), i.e. is executed in the linear time, while the binary search is proportional to the logarithm of the number of members, i.e. has O($log(n)$) or is executed in logarithmic time. Additionally, in general, the number of steps that will be taken depends not only on the size of the input data, but also on their values. Since, in principle, values cannot be controlled, in most cases, these algorithm complexity notations give an upper limit, i.e. worst case for the input for the algorithm.

V. Implementation of the Metrics-Based Compexity Analysis

After the algorithms have executed to completion, and made their metrics available, an analysis of their complexity can be made. The analysis is performed by the implementation of the **IAnalyzer** interface, shown in Fig. 4

```
interface IAnalyzer {
   analyzeMetrics (metrics:number[]):Complexity;
}
```

Fig. 4

where **Complexity** is an enumeration with the following values

```
enum Complexity {
    Constant="constant",
    Linear="linear",
    Square="square",
    Logarithmic="logarithmic",
    Exponential="exponential",
    LinearLog="linear-log",
    Unknown="unknown"
}
```

Fig. 5

If the user does not assign their own implementation of the **IAnalyzer** interface, the default **Analyzer** class implementation will be used. This implementation uses elementary numerical differentiation to evaluate the trend of values, giving an appropriate result.

VI. Practical Examples

A. Fibonacci Numbers

Fibonacci numbers are numbers with a recursive definition, namely, the first two Fibonacci numbers are 1 and 1, and each remaining Fibonacci number is a sum of the previous two Fibonacci numbers. The function in Fig. 6 is the naive implementation of this definition:

```
public fibbonaci(value: number): number {
  if ((value === 0) || (value === 1)) {
    return 1;
  }

  return fibbonaci(value-1)+fibbonaci(value-2);
}
```

Fig. 6

Accordingly, the problem to be solved is to determine the corresponding Fibonacci number for a given non-negative integer. Due to the recursive nature of the definition itself, a metric that can be used in evaluating the complexity of a particular algorithm to solve this problem will be the number of recursive calls that are needed to get a solution.

Two algorithms for solving are made, with the former using the naïve version (BruteForceFibbonaci), and the latter (DynamicFibbonaci) uses dynamic programming. In it, during the initialization of the algorithm, a map of solutions is created, and then, during execution, if the input parameter is contained as a key in the map, the function immediately returns the value for that key. If the input parameter is not contained in the map, its value is calculated (using the naïve approach), and then the value obtained is saved in the map before it returns as a result.

```
private cache: NumberMap= {
  0: 1,
  1: 1
};
public fibbonaci(value: number) {
  if (this.cache[value] !==undefined)
```

```
        return this.cache[value]
    const result= this.fibbonaci(value - 1) +
                  this.fibbonaci(value - 2);
    this.cache[value] = result;
    return result;
}
```

Fig. 7

The input data used are natural numbers from 1 to 40. The obtained results by comparing these two algorithms are given in Table 1:

TABLE 1

Algorithm	**Brute Force**	**Dynamic**
Total Run Time	12.6 s	0.00 s
Call Count Metric	Exponential	Linear

The actual values for the used metric for the last few input values are given in Table 2:

TABLE 2

Algorithm Name	**Brute Force**		**Dynamic**	
Input	***Execution Time***	***Call Count***	***Execution Time***	***Call Count***
37	1.896 s	78,176,337	<1ms	73
38	2.984 s	126,491,971	<1ms	75
39	4.665	204,668,309	<1ms	77

B. Dijsksra's Shortest Path Algorithm

Many programming problems come down to graph analysis, i.e. objects that are composed of nodes, and edges between these nodes. Many of these problems do not have known solutions with polynomial complexity, i.e. belong to the class of NP problems. One of them is the problem of the shortest path between a particular start and end node. The naive implementation of a solution to this problem is by generating all possible paths and comparing their total costs. This approach, however, has an extremely high complexity, that is, it depends exponentially on the number of input nodes and branches between them.

On the other hand, graph problems are resolved well with the help of dynamic programming, because they satisfy the conditions for the existence of dynamic algorithms. For example, the structure of the graphs itself is such that a subset of nodes and branches is a graph again. These features were used in 1956 by Edsger Dijkstra to design a dynamic algorithm for the shortest path through the graph. The analyzed complexity of this algorithm is O (e + v*log*v), where e is the number of branches, and v the number of nodes.

Correspondingly, two algorithms for finding the shortest path through a graph are implemented, with, respectively, naive implementation, and using the algorithm of Dykstra. The metric that is set is simply counting the access to information for a particular node. The results obtained correspond to the analytically expected values (Table 3):

TABLE 3

Algorithm	**Naïve**	**Dijkstra**
Total Run Time	26.4 s	0.10 s
Node Access Metric	Exponential	Linear-Log

C. Matrix Chain Multiplication

The optimal multiplication algorithm for two matrices is still one of the unsolved programming problems, but in any case, multiplication of two matrices is a process that takes up considerable resources. It is clear that all matrix multiplication algorithms depend on the size of the matrices, with a degree that (most likely) is strictly greater than 2, with the naive approach for multiplication of matrices having a cubic complexity. In the analysis of the problem, we will assume that the simple approach will be used for the multiplication of matrices.

A common case in working with matrices is the need for sequential multiplication of more than two matrices. Since multiplication of matrices is associative, if we have to multiply three matrices, A * B * C, the end result will be the same, whether multiplication will be implemented as (A * B) * C or as A * (B * C). Similarly, if we have 4 matrices, we have 5 different ways in which we can multiply (((A * B) * C) * D, (A * B) * (C * D), (A * (B * C)) * D, A * ((B * C) * D) and A * (B * (C * D))). Not all of these alternatives will be identical from the perspective of the number of operations that need to be performed. For example, if we have three matrices in the sizes 2x10, 10x3 and 3x4, the result in both alternatives will be a 2x4 matrix, but if we calculate it using (A * B) * C, we will need to run 2 * 10 * 3 + 2 * 3 * 4 = 84 multiplications, and if we calculate it using A * (B * C) we will need to perform 10 * 3 * 4 + 2 * 10 * 4 = 180 multiplications. It is obvious that the first approach would be more than twice more efficient than the latter, regardless of the values of the matrix itself and the algorithm used to multiply them.

The problem in this case is to find the most optimal way to put brackets, i.e. the order of multiplication of a given array of matrices. The naïve approach is to generate all possible combinations (essentially, all the permutations of the order of operations), to evaluate them, and to choose the best ones. Since the number of permutations is dependent on the factorial of the number of matrices, the complexity of this approach is exponential.

An alternative, but equivalent (with the equivalent complexity) is the next recursive approach

- Split the matrix array to two sub-arrays
- Find the minimum cost for multiplying each sub-array.
- The cost of multiplication is the sum of the prices for each subset plus the multiplication cost of the two results.
- Repeat the algorithm for each position where we can split the matrix array.

This variant of naïve approach has been implemented. Dynamic improvement of this variant has also been implemented, since the upper algorithm is subject to

memoization. Namely, many of the subproblems are repeated, i.e. the optimal multiplication for a given sub-set will be determined again and again. If these values are calculated only once, and are remembered, the complexity of the algorithm drops dramatically, which can be seen from the results obtained by execution:

TABLE 4

Algorithm	Naïve	Dynamic
Total Run Time	116.3 s	3.42 s
Call Count Metric	Exponential	Cube

The metric used is the number of recursive function calls. It is worth noting that the cubic complexity of the dynamic algorithm is still relatively high, but it is obviously dramatically lower than the naive implementation.

D. Array Sorting

Array sorting is one of the most common and best-studied programming problems. There are many sorting algorithms with a variety of complexities. Two comparisons of three sorting algorithms were executed. The algorithms chosen were Bubble Sort, Quick Sort and Merge Sort, and the metrics are

- The number of array element accesses
- The number of comparisons between different array elements
- The number of swaps between different array elements.

All three algorithms are implemented by sorting in place, so that no auxiliary arrays with copies of the elements are created, but the input array is sorted itself.

The execution comparisons differ in the nature of the arrays to be sorted, namely the first run uses an array with randomly generated elements, while the second comparison is sorting a pre-sorted array.

It should also be noted that a different analyzer is used to analyze the results of these metrics with greater tolerance because the results of randomly generated elements are not completely deterministic.

Using arrays with randomly selected values, we got the following results (Table 5):

TABLE 5

Algorithm	Bubble Sort	Quick Sort	Merge Sort
Total Run Time	15.9 s	0.014 s	0.010 s
Comparisons Metric	Square	Linear-Log	Linear-Log
Swaps Metric	Square	Linear-Log	Linear-Log
Accesses Metric	Square	Linear-Log	Linear-Log

For the second comparison of sorting algorithms we will use an already sorted array, because it is known that some algorithms perform with worst-case efficiency on that specific case. That can indeed be seen from the results (Table 6)

TABLE 6

Algorithm	Bubble Sort	Quick Sort	Merge Sort
Total Run Time	5.91 s	1.52 s	0.010 s
Comparisons Metric	Square	Square	Linear-Log
Swaps Metric	Constant	Linear	Constant
Accesses Metric	Square	Square	Linear-Log

By comparing the results, it can be seen that the complexity of the sorting algorithms depends on the data being sorted, specifically the second case is intentionally chosen to be the worst case for the Quick Sort algorithm. That algorithm, although in the average case is an O(n*log*n) algorithm, in the worst case is a square algorithm. In contrast, the Merge Sort algorithm, in the worst case, still retains its complexity of O (n*log*n)

VII. Conclusion

From the values obtained by comparing the algorithms with and without dynamic programming, we can conclude that dynamic programming is a powerful technique by which we can perform effective execution of algorithms with extremely large time complexities, whose regular execution would require an extraordinary amount of time, while dynamic versions often have near linear complexity.

We still need to note that dynamic programming is only applicable to a specific class of problems, and that, in principle, it uses considerably more memory to store intermediate values that are usually not relevant (or even accessible) to the end user. But, given the increasing availability of memory, in most cases this tradeoff in memory for time is more than acceptable.

A drastic example of this is the situation with the Fibonacci numbers, where for the price of a simple map of numbers, totaling less than 1KB, we get a speed-up of over several million times. And not only that, the dynamic version can be used to generate a larger range of numbers, because the time needed to generate numbers over 50 with the naive algorithm, even on today's modern hardware, becomes significant.

References

[1] J. W. Hunt, M. D. McIlroy, “An Algorithm for Differential File Comparison”, Computer Science, 1975

[2] Hasan Laiq, Al-Ars Zaid, “Performance Improvement of the Smith-Waterman Algorithm”, 2007

[3] Sedgewick, Robert “Algorithms In C: Fundamentals, Data Structures, Sorting, Searching”, Pearson Education, 1998.

[4] Sedgewick, Robert, "Implementing Quicksort programs", Communications of the ACM, Volume 21, Issue 10, 1978

[5] Ajtai, M.;Komlós, J.;Szemerédi, E. “An O(n log n) sorting network”, Proceedings of the fifteenth annual ACM symposium on Theory of computing, 1983

[6] Huang, B. C.; Langston, M. A.."Fast Stable Merging and Sorting in Constant Extra Space", 1992

Two-phase Classification of Colorectal Cancer Stages

Frosina Stojanovska, Viktorija Velinovska, Monika Simjanoska and Ana Madevska Bogdanova
Faculty of Computer Science and Engineering
Ss. Cyril and Methodius University
Skopje, Republic of Macedonia
stojanovska.frose@gmail.com, velinovska.viktorija@gmail.com,
monika.simjanoska@finki.ukim.mk, ana.madevska.bogdanova@finki.ukim.mk

***Abstract*—Staging of colorectal cancer is one of the essential factors required to identify the patient's true therapy for recovery. Despite the various clinical colorectal cancer staging methods, this problem remains critical for personalized stage determination. In this paper, we study the problem of colorectal cancer stage determination using gene expression data obtained from DNA microarrays. The goal is to construct a supervised machine learning classification model that will be able to detect the stage of colorectal cancer, that is, the model should be able to separate the stages utilizing 11 biomarkers as features.**

Dataset resampling and analyzing the errors between the real and the predicted class during validation phase, led to the creation of two-phase classification model, dividing the main problem of determining the stage of the colorectal cancer into sub-problems. In the first phase of classification, it is necessary to create a classification model that will successfully divide the data between two groups obtained by joining stage I and IV as one sub-group, and stage II and III as the second sub-group. Once an instance of the data set is classified into one of the combined classes, according to this class, the second level classification reveals the true cancer stage of the instance. Random Forest is the machine learning algorithm that performed best in all the experiments, compared to KNN, SVM, Naive Bayes and MLP.

***Keywords*—gene expression, colorectal cancer, stage detection, machine learning, ensemble methods**

I. Introduction

Colorectal cancer (CRC) is malignant cancer located in the colon and/or rectum. According to the statistics from the World Health Organization (WHO) [1] provided in 2017, the cancer is one of the leading sources of death worldwide, and colorectal cancer is the third most common type of cancer that occurs in men and women.

After the cancer is diagnosed, it is essential to find the level of cancer expansion in the affected body part. This is the process of cancer staging, which helps the doctors to choose the most appropriate treatment for the recovery of the patient. There are four stages of colorectal cancer spread in the AJCC TNM (Tumor size, Lymph Nodes affected, Metastases) staging system [2], starting from stage I (1) to stage IV (4), and additionally stage 0 representing a very early phase of cancer. With this order, an earlier stage refers to a lower degree of cancer. Histopathology is used in clinical practice for discovering the cancer stage, with analysis of the local tumor invasion and the presence of CRC metastases in lymph nodes. Histologic staging has difficulty detecting the cancer stage in individuals, so there is a need for more sensitive and better methods [3].

Cancer is a disease caused by several genetic and epigenetic alterations [4]. These genetic changes can lead to an unusual growth of the cells that are transforming into cancer cells. Cancer research includes solutions from bioinformatics, for instance, diagnostic protocols or pattern discovery in cancer by analyzing biological data, especially of the omics data [5]. The progression of omics data analysis with bioinformatics technologies involves the integration of huge amount of data, including genomics, transcriptomics and proteomics data from many different sources. The multi-omics analysis is continuously more popular in biomedical research, and as a result authors in [6] had built a freely available platform LinkedOmics for analysis and comparison of cancer multi-omics data within and across multiple cancer types.

Machine learning methods are rising as a solution to many problems in distinct domains. These techniques are utilized to model the progression and treatment of cancerous conditions [7]. Machine learning, with its supervised, semi-supervised, unsupervised, or even reinforcement learning methods, has the ability to give an interpretation, or, a possible solution of many biological problems [8]–[10]. The authors in [11] give an overview of various machine learning models applied to cancer prognosis and prediction.

Some implemented machine learning algorithms rely on gene biomarkers as features to build the models. Biomarkers have a key role in cancer disease discovery, treatment selection, drug discovery, and personalized medicine [12]. Although there are plenty of studies that report biomarkers as significant related to some disease, there are still very few of them validated of proven and robust clinical utility [13].

Lately, many researchers study and analyze the gene expression profile data associated with CRC. The authors in [14] inferred a colon cancer gene regulatory network and studied its functional and structural meaning using gene expression data. The goal in this direction is to make a comparative analysis of this kind of networks of more than one cancer networks [14]. The research in [15] introduces a study for finding the potential key candidate genes and pathways in CRC from the differentially expressed genes (DEGs).

In this paper, we investigate and present a solution to the

problem of detecting the CRC stage employing supervised machine learning models built with gene expression data in infected cells. The rest of the paper is organized as follows. Section II presents the dataset and the methods used to obtain this dataset. Additionally, it gives details of the machine learning methods used to build the model of stage classification. The details of the experiments and the results are given in Section III. Finally, the last section, Section IV, recapitulates the main findings and offers suggestions for future work.

II. Materials and Methods

DNA microarrays are used to study the extent to which certain genes are active in cells and tissues. Two widely used types of DNA microarrays are Affymetrix and Illumina chips [16]. More detailed information about this technology is described in [17].

A. Colorectal Cancer Dataset

The dataset used in this paper consists of 657 instances with features comprised of gene expression from 11 genes, selected to be the biomarkers associated with CRC, and 1 feature for the CRC stage. The distribution of the four CRC stages is given as:

- *Stage I* - gene expressions from 137 patients.
- *Stage II* - gene expressions from 257 patients.
- *Stage III* - gene expressions from 182 patients.
- *Stage IV* - gene expressions from 81 patients.

The dataset is constructed by merging several CRC datasets from Gene Expression Omnibus database [18] (whose identifiers are provided in [19]). The 11 biomarkers extracted by the analysis presented in [19] showed influence in colorectal cancer determination. These biomarkers are used in this paper as features to investigate their importance as gene biomarkers in colorectal cancer stage determination.

The selected biomarkers genes are:

- *CDH3* - Cadherin 3 (or P-cadherin) is a protein-coding gene that encodes a classical cadherin of the cadherin superfamily. This gene is located on the chromosome 16 and is associated with specific hereditary genetic disorders and several cancers including CRC [20].
- *CHGA* - Chromogranin A is the gene that encodes a protein that is part of the granin family of neuroendocrine secretory proteins. CHGA is used as an indicator of neuroendocrine tumors including carcinoids [21], [22].
- *DHRS9* - Dehydrogenase/reductase 9 is the official name of this gene that encodes a protein which has an oxidoreductase activity toward hydroxysteroids and is a member of short-chain dehydrogenases/reductases (SDR) family. Paper [23] provides evidence for association of the decreased expression of DHRS9 with disease progression and poor outcome of CRC patients.
- *GUCA2A* - GUCA2A or guanylate cyclase activator 2A is an endogenous activator of intestinal guanylate cyclase. The differential expression of this gene in CRC was associated with the tumor stage in [24].
- *GUCA2B* - Guanylate cyclase activator 2B encodes a preproprotein that is proteolytically processed to generate multiple protein products from the guanylin family. This gene was one of the six colorectal cancer related genes in [25].
- *HPGD* - Hydroxyprostaglandin dehydrogenase 15-(NAD) encodes nonmetalloenzyme alcohol dehydrogenase protein responsible for the metabolism of prostaglandins.
- *MMP3* - Matrix metallopeptidase 3, as a gene from the cluster of MMP genes, encodes protein from the matrix metalloproteinase (MMP) family. In [26] MMP3 is introduces as a prognostic factor of tumor progression in three common poor prognosis tumor types (pancreatic, pulmonary, and mammary carcinoma).
- *MMP7* - Matrix metallopeptidase 7 is another gene from the cluster of MMP genes. This gene is overexpressed in association with CRC liver metastases in paper [27].
- *PYY* - Peptide YY is the full name of the gene that encodes preprotein as one of the neuropeptide Y (NPY) family of peptides.
- *SCG2* - Secretogranin II encodes one type of neuroendocrine secretory proteins. This gene was overexpressed in advanced prostate cancer as shown in paper [28].
- *VIP* - Vasoactive intestinal peptide (or VIP) encodes a glucagon protein. This gene is expressed in several tissues, most abundant in pancreatic islets cells and nerve ganglion [29].

The biomarkers importance has been experimentally investigated by using the Random Forest method. Fig. 1 presents the importance of the biomarkers in descending order, showing the gene MMP7 to be the most important and the gene HPGD to be the least important biomarker for separation of the stages with the Random Forest model.

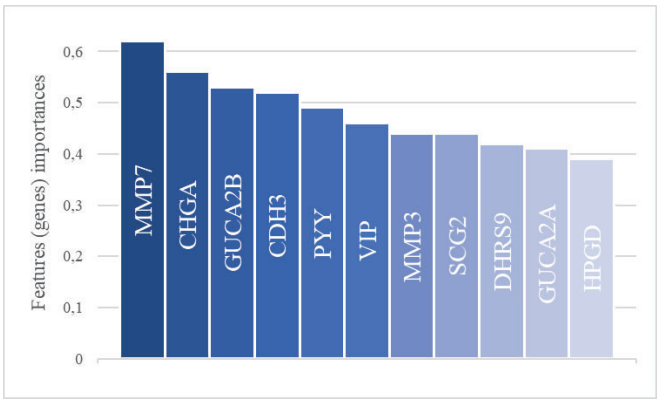


Fig. 1. Feature importance of the biomarkers according to Random Forest method.

B. The Methodology

Several supervised machine learning algorithms were built to classify the stage of expansion of the CRC. The models were supposed to find the separation of the CRC stages using the dataset explained in the previous Section II-A.

This section provides an overview of the selected and applied algorithms: Support Vector Machines, K Nearest Neighbors, Multilayer Perceptron, Naive Bayes and Random Forest.

1) SVM: Support vector machines (SVM) are the standard machine learning technique utilized for many problems. SVMs take the data as input and process it into a large dimensional space. Although SVM can be quite complex, considering the small dimension of our dataset used for training, instance and feature size, this was not a problem in our case. The SVM classifier is important with our approach not only because it promises a good performance, as shown with many other similar implementations, but also it is a model that can capture the multivariate statistical properties of our data.

We need a model that distinguishes four different CRC stages, although SVM works with a binary class. Consequently, we applied SVM with a pairwise classification (one vs one). The proper kernel function holds the ability to model the high-dimensional associations from the data. We have selected the following options for the kernel function: polynomial kernel, Pearson VII function-based universal kernel (PUK), and radial basis function kernel (RBF kernel). PUK kernel function, shown in (1), had the best performance from all kernel functions, so the results in Section III refer to SVM model with PUK kernel function.

$$K(x_i, x_j) = \frac{1}{[1 + (2\sqrt{\| x_i - x_j \|^2}\sqrt{2^{1/\omega} - 1/\sigma})^2]^{\omega}} \quad (1)$$

2) KNN: We used the IBk algorithm to implement the KNN classification. This algorithm actually represents the KNN algorithm, where IB refers to instance-based (the other name under which the nearest neighbors are known), while k allows us to specify the number of closest neighbors. KNN as a lazy approach works without creating a model and classifies a new data point with the data itself. We can notice that this method is much simpler than SVM. KNN is not limited to linearity, so it can capture even nonlinear relations between the features. This factor made this algorithm relevant to our problem regarding that we did not know the type of interaction between the genes.

We use the Euclidean distance as a measure of calculating the closeness of data points, having in mind that all the features are actually real numbers. To determine the optimal value for the parameter k, we considered a space of values ranging from 3 to 21. The optimal k-value in most of the experiments was 15. Also, the best results required standardization of the attributes, that is, the gene expression of the 11 biomarkers. Apart from the distance measure and the number of closest neighbours, KNN has another parameter - vote weighting. Initially, the weights of every data point were equal. Setting the weight to be the inverse distance $(1/distance)$ the method remarkably improved its performance.

3) MLP: In the last few years, Multilayer Perceptron (MLP) has become one of the most promising methods in machine learning. This neural network consists of hidden layers with multiple perceptrons that enable the modelling of any function required for achieving the best separation of the data instances. With this property, MLP was one of the algorithms selected to model the function of the dataset with unknown feature relations.

We implemented the network with one hidden layer. Adding additional layers did not bring any gain, which is expected given the small size of the dataset. The backpropagation algorithm was used to change the weights of the neurons in the process of training. These weights were adjusted using the gradient descent and squared error loss function. In this study, we use the sigmoid function as the activation function.

4) Naive Bayes: The classification with the Naive Bayes model is based on the Bayesian theorem which uses independent assumptions between the predicates. This classifier is easy to build, so it is also suitable for large datasets. This algorithm is a simple technique for constructing a classifier, where the model in this study is built with probabilities obtained through the features of gene expressions and CRC stages in the dataset. We do not have information about the independence of the gene biomarkers. However, despite the naive design and obviously too simple assumptions, this classifier has proven to work well in very complex real situations, overcoming other much more complex classifiers, so it is part of our experiments.

5) Random Forest: Random Forest is an ensemble learning procedure called Bootstrap Aggregation, or, Bagging, adopted for classification, regression and similar problems [30]. This method has excellent performance in classification tasks, equivalent to standard methods as SVMs. Random Forest has promising features including the ability for classification of both two-class and multi-class problems of more than two classes. Also, it is able to measure the feature (gene) importance. Another advantage is that the parameter fine-tuning is simple, with a selection of small numbers of parameters including the number of input features, the number of trees in each forest, as well as the minimum size of the leaf nodes.

The class determination, that is, the classification, is obtained by the mean of classes received by all trees. With this, Random Forest tries to fix the overfitting that trees do with the training datasets. This appeared as a method that would help to find the perfect decision boundaries between CRC stages. This algorithm works efficiently with both large and small datasets. It can handle a huge number of input features without removing some of them.

Generated ”forests” can be stored for the next use of other datasets. Also, it is able to calculate closeness between pairs of cases that can be used for clustering, finding outliers, or to give interesting views of the data (by scaling), which can be used to visualize the correlations in our dataset.

III. Experiments and Results

To build and evaluate the described models in the previous section, we used the Weka software tool [31], as well as the web tool ArrayMinning described in [32]. We performed the experiments according to the complexity of the method, starting from Naive Bayes, up to MLP and SVM.

All the classification models were tested using cross-validation with k folds, where for the parameter k, we assigned

a value of 10. The results obtained with the use of all the classifiers are shown in Table I. This validation technique is used in all of the experiments in this section.

The performance of the techniques has been measured by using *correctly classified (CC) instances*, *Area Under the Curve (AUC)*, *Kappa statistics (KS)* and *Mean Absolute Error (MAE)*. We use CC to show the overall accuracy of the classification of the method, and AUC refers to weighted average of the area under the Receiver Operating Characteristic (ROC) curve of every class. KS denotes the reliability of the method, i.e. measures how the improvement of the predictor is relative to a random predictor (1 means perfect predictor, 0 means the predictor is no better than a random one). The KS metrics is given as

$$KS = \frac{p_a - p_r}{1 - p_r} \tag{2}$$

where p_a is the success rate of the actual predictor and p_r is the success rate of a random predictor. MAE is the sum of the absolute differences between predictions and actual values. It is commonly used in regression models, and for the classification is defined as

$$MAE = \frac{\sum_{i=1}^{n} \sum_{j=1}^{k} \mid a_j - p_j \mid}{n} \tag{3}$$

where n is the number of instances in the test set, k refers to the number of classes, a_j is the actual class value (1 if the particular instance class is j and 0 otherwise), and p_j is the predicted probability of the model for the instance to be classified as class j. We compared the performance of the methods using the CC and AUC metrics and used the other metrics as a profound observation of the precision and separation ability of the models.

A. One-phase Classification

The performance of the models shown in Table I is not satisfactory, i.e., the models are not capable of separating the stages of cancer. Therefore, we tried to pre-process the dataset with different methods before training the models. Of all the pre-processing techniques that were performed, the only thing that made a significant improvement was the sampling of the dataset. This method modified the dataset while leaving the same number of instances. In fact, samples were randomly selected from the ”old” dataset in order to build the ”new” dataset, preserving the initial distribution of the classes. When selecting a sample of the dataset, it can be re-selected as a sample in the next iterations - the process called sampling with replacement. The last iteration is actually the iteration with which the dataset has the same number of samples as in the start before pre-processing.

Table II holds the results of the classification models after performing sampling with replacement. From the results in this table, Random Forest can be distinguished as the best model for classification, which in the previous attempt, shown in Table I, performs slightly less than SVM and KNN. KNN

TABLE I
CLASSIFICATION RESULTS

Metric	SVM	KNN (k=15)	Multilayer Perceptron	Naive Bayes	Random Forest
CC	**44.44%**	**44.44%**	41.25%	42.77%	42.77%
AUC	0.594	0.646	0.624	0.654	**0.658**
KS	0.141	**0.180**	0.114	0.178	0.156
MAE	0.327	0.323	**0.313**	**0.313**	0.320

is very close to the performance of Random Forest. Next are the models of SVM and MLP, while Naive Bayes model is the weakest model in this combination of classification of the given data set.

TABLE II
CLASSIFICATION RESULTS AFTER DATASET SAMPLING

Metric	SVM	KNN (k=15)	Multilayer Perceptron	Naive Bayes	Random Forest
CC	57.08%	73.67%	55.40%	43.99%	**75.95%**
AUC	0.699	0.921	0.735	0.670	**0.926**
KS	0.352	0.624	0.370	0.206	**0.656**
MAE	0.311	**0.160**	0.249	0.310	0.187

Fig. 2 shows the difference in the performance of the models, i.e. the improvement of the accuracy or the percentage of correctly classified instances with the original dataset before sampling, and the dataset after sampling.

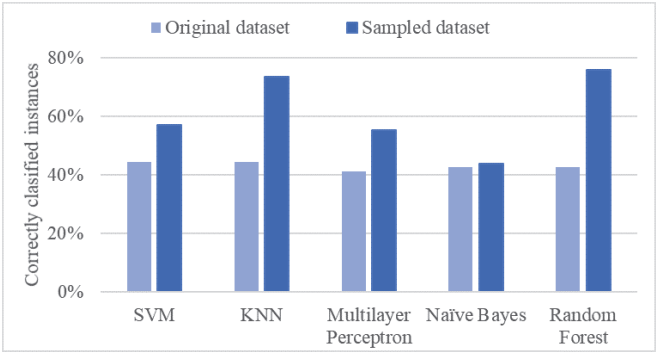


Fig. 2. Difference in the accuracy of the models before and after dataset sampling.

B. Two-phase Classification

Analyzing the errors between the real and predicted classes during testing, we observed an association between the first and fourth stage. With this property, we decided to create a two-phase classification model, where the main problem of determining the stage of the cancer is divided into solving two sub-problems. Therefore, we created two sub-groups of the cancer stages from the resampled dataset, where the first sub-group combines the first and the fourth stage and the second sub-group is a compound of the second and third

cancer stage. Fig. 3 provides a visual representation of the two-phase classification model of CRC stages.

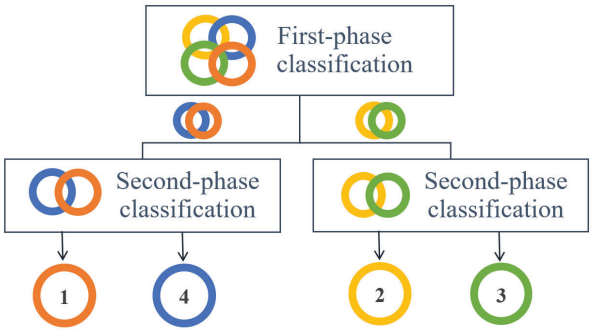


Fig. 3. Visual representation of the two-phase classification model - The first phase is the detection of the sub-groups, and the second phase separates the sub-groups into the real (actual) cancer stages. Cancer stage I is the orange circle, stage II is represented by the yellow circle, stage III is the green one, and the blue circle is stage IV.

The first stage of the classification model considered the separation of the instances from the sub-groups. This stage uses the same machine learning models as in the previous problem. The results are presented in Table III. Random Forest, obtained the highest accuracy of 87%, that is, this algorithm can classify the instances of two sub-groups, with reasonably high correctness. The other algorithms are not nearly satisfactory as Random Forest, especially Naive Bayes.

TABLE III
FIRST-STAGE CLASSIFICATION RESULTS IN THE TWO-PHASE CLASSIFICATION MODEL

Metric	SVM	KNN (k=15)	Multilayer Perceptron	Naive Bayes	Random Forest
CC	76.10%	82.57%	75.34%	68.49%	**87.21%**
AUC	0.663	0.935	0.766	0.699	**0.946**
KS	0.377	0.615	0.459	0.268	**0.700**
MAE	0.239	**0.186**	0.278	0.371	0.211

After the first-stage classification determines the aggregate class, i.e. the sub-group of the instance, depending on the identified sub-group, the exact stage of cancer should be determined. Hence, the next sub-problem required finding two separate classifiers. The first classification model knows how to divide the first sub-group into the first or fourth stage of cancer. Respectively, the second model splits the second subgroup into the second and third stage of cancer. Again, the models were built with the same machine learning techniques.

The results of the classification model that distinguish the first and fourth CRC stage are given in Table IV. This classification divides the first sub-group. Random Forest again shows the best performance with an accuracy of 85.32%. Table V presents the results of the other classification model in the second-stage that separates the second sub-group into the second and third CRC stage. As before, Random Forest is again dominating in the process of separation of the classes,

TABLE IV
SECOND-STAGE CLASSIFICATION RESULTS FOR THE FIRST SUB-GROUP

Metric	SVM	KNN (k=11)	Multilayer Perceptron	Naive Bayes	Random Forest
CC	77.06%	83.03%	79.36%	62.84%	**85.32%**
AUC	0.707	0.946	0.758	0.666	**0.956**
KS	0.472	0.628	0.538	0.248	**0.678**
MAE	0.223	0.180	0.115	0.200	**0.106**

with an accuracy of 83.37%. In the second-stage classification, KNN, MLP and SVM have notable outcomes, whereas Naive Bayes is not appropriate for resolving this difficulty.

TABLE V
SECOND-STAGE CLASSIFICATION RESULTS FOR THE SECOND SUB-GROUP

Metric	SVM	KNN (k=9)	Multilayer Perceptron	Naive Bayes	Random Forest
CC	73.80%	79.73%	78.36%	60.59%	**83.37%**
AUC	0.716	0.916	0.746	0.645	**0.923**
KS	0.442	0.578	0.553	0.137	**0.653**
MAE	0.226	0.210	0.220	0.212	**0.128**

Fig. 4 illustrates the two-phase classification model, with accuracy performance of each particular classification model for every sub-task.

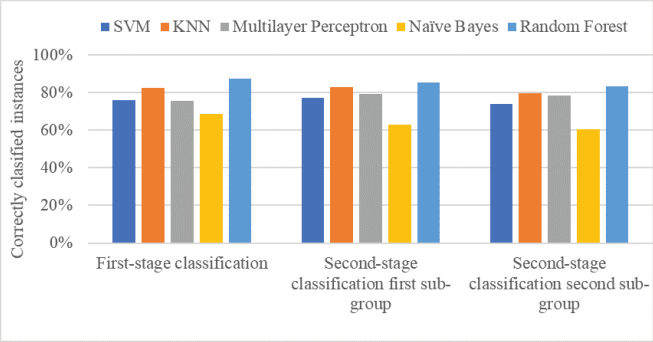


Fig. 4. Performance of the models in the individual tasks in the two-phase classification model.

IV. CONCLUSIONS

By using machine learning techniques, we classify four stages of colorectal cancer in patients, using the gene expression of 11 biomarkers obtained by DNA microarray technology. Recent research has shown that changes in gene expression are associated with different types of cancer.

The choice of best machine learning algorithm to be applied, is based on the nature of the problem and the data set that is used. We used several methods for building the classification model: KNN, SVM, MLP, Naive Bayes and Random Forest. With the initial set, we did not obtain good results, however, when a dataset resampling is applied, the classification significantly improved. The model with Random Forest stands out

as the best classifier model with an accuracy of 76%, along with KNN with 74% accuracy, which was not as satisfactory as we expected.

Considering the unexpected association between the first and the fourth stage, a two-phase classification model was created. In the first phase, the model divides the data between two sub-groups obtained by joining the first and fourth stage as one sub-group, and the second and third stage as the second sub-group. As the best classifier for this case, Random Forest stands out with an accuracy of 87%. The second phase contains two classifiers to divide each individual subgroup to obtain the right cancer stage of the instance. Random Forest again shows the best performance - for the classification of the first and fourth stage the accuracy is 85%, while for the classification of the second and third stage of cancer the accuracy is 83%.

Given the results, we can conclude that the ensemble machine learning methods, represented by Random Forest, along with slightly worse KNN, provide better modelling of the CRC biomarkers gene expressions. The importance of the developed two-phase classification of gene expression for other cancers or other biomarkers remains to be revealed.

The future work will include a deeper analysis of the problem and the CRC data. One option to consider is transfer learning which has the potential of combining previously gained knowledge and solve new related issues. It is mostly implemented with Deep Learning architectures, however, Random Forests are also capable of model transferring. This can be very helpful with very small and limited datasets, as in our case.

REFERENCES

[1] (2017) World health organization. [Online]. Available: http://www.who.int/en/

[2] S. B. Edge and C. C. Compton, "The american joint committee on cancer: the 7th edition of the ajcc cancer staging manual and the future of tnm," *Annals of surgical oncology*, vol. 17, no. 6, pp. 1471–1474, 2010.

[3] S. A. Bustin and S. Dorudi, "Gene expression profiling for molecular staging and prognosis prediction in colorectal cancer," *Expert review of molecular diagnostics*, vol. 4, no. 5, pp. 599–607, 2004.

[4] A. Balmain, J. Gray, and B. Ponder, "The genetics and genomics of cancer," *Nature genetics*, vol. 33, p. 238, 2003.

[5] B. Stransky and P. Galante, "Application of bioinformatics in cancer research," in *An Omics Perspective on Cancer Research*. Springer, 2010, pp. 211–233.

[6] S. V. Vasaikar, P. Straub, J. Wang, and B. Zhang, "Linkedomics: analyzing multi-omics data within and across 32 cancer types," *Nucleic acids research*, vol. 46, no. D1, pp. D956–D963, 2017.

[7] K. Kourou, T. P. Exarchos, K. P. Exarchos, M. V. Karamouzis, and D. I. Fotiadis, "Machine learning applications in cancer prognosis and prediction," *Computational and structural biotechnology journal*, vol. 13, pp. 8–17, 2015.

[8] P. Baldi and S. Brunak, *Bioinformatics: the machine learning approach*. MIT press, 2001.

[9] I. Inza, B. Calvo, R. Armañanzas, E. Bengoetxea, P. Larrañaga, and J. A. Lozano, "Machine learning: an indispensable tool in bioinformatics," in *Bioinformatics methods in clinical research*. Springer, 2010, pp. 25–48.

[10] P. Larranaga, B. Calvo, R. Santana, C. Bielza, J. Galdiano, I. Inza, J. A. Lozano, R. Armananzas, G. Santafé, A. Pérez *et al.*, "Machine learning in bioinformatics," *Briefings in bioinformatics*, pp. 86–112, 2006.

[11] J. A. Cruz and D. S. Wishart, "Applications of machine learning in cancer prediction and prognosis," *Cancer informatics*, vol. 2, p. 117693510600200030, 2006.

[12] P. Carrigan and T. Krahn, "Impact of biomarkers on personalized medicine," in *New Approaches to Drug Discovery*. Springer, 2015, pp. 285–311.

[13] G. Novelli, C. Ciccacci, P. Borgiani, M. P. Amati, and E. Abadie, "Genetic tests and genomic biomarkers: regulation, qualification and validation," *Clinical cases in mineral and bone metabolism*, vol. 5, no. 2, p. 149, 2008.

[14] F. Emmert-Streib, R. de Matos Simoes, G. Glazko, S. McDade, B. Haibe-Kains, A. Holzinger, M. Dehmer, and F. C. Campbell, "Functional and genetic analysis of the colon cancer network," *BMC bioinformatics*, vol. 15, no. 6, p. S6, 2014.

[15] Y. Guo, Y. Bao, M. Ma, and W. Yang, "Identification of key candidate genes and pathways in colorectal cancer by integrated bioinformatical analysis," *International journal of molecular sciences*, vol. 18, no. 4, p. 722, 2017.

[16] M. Barnes, J. Freudenberg, S. Thompson, B. Aronow, and P. Pavlidis, "Experimental comparison and cross-validation of the affymetrix and illumina gene expression analysis platforms," *Nucleic acids research*, vol. 33, no. 18, pp. 5914–5923, 2005.

[17] V. Trevino, F. Falciani, and H. A. Barrera-Saldaña, "Dna microarrays: a powerful genomic tool for biomedical and clinical research," *Molecular Medicine*, vol. 13, no. 9-10, p. 527, 2007.

[18] (2013) Gene expression omnibus. [Online]. Available: https://www.ncbi.nlm.nih.gov/geo/

[19] M. Simjanoska, A. M. Bogdanova, and S. Panov, "Gene ontology analysis of colorectal cancer biomarkers probed with affymetrix and illumina microarrays." in *IJCCI*, 2013, pp. 396–406.

[20] A. F. Vieira and J. Paredes, "P-cadherin and the journey to cancer metastasis," *Molecular cancer*, vol. 14, no. 1, p. 178, 2015.

[21] K. A. Mirkin, C. S. Hollenbeak, and J. Wong, "Impact of chromogranin a, differentiation, and mitoses in nonfunctional pancreatic neuroendocrine tumors 2 cm," *journal of surgical research*, vol. 211, pp. 206–214, 2017.

[22] W. Rogowski, E. Wachuła, A. Lewczuk, A. Kolesińska-Ćwikła, E. Iżycka-Świeszewska, V. Sulżyc-Bielicka, and J. B. Ćwikła, "Baseline chromogranin a and its dynamics are prognostic markers in gastroenteropancreatic neuroendocrine tumors," *Future Oncology*, vol. 13, no. 12, pp. 1069–1079, 2017.

[23] L. Hu, H.-Y. Chen, T. Han, G.-Z. Yang, D. Feng, C.-Y. Qi, H. Gong, Y.-X. Zhai, Q.-P. Cai, and C.-F. Gao, "Downregulation of dhrs9 expression in colorectal cancer tissues and its prognostic significance," *Tumor Biology*, vol. 37, no. 1, pp. 837–845, 2016.

[24] Y. Chen, Y. Zhu, H. Feng, Y. Liu, J. Qian, Y. Fan, and D. Li, "Differential expression of guanylin in colorectal cancer," *Zhonghua wei chang wai ke za zhi= Chinese journal of gastrointestinal surgery*, vol. 12, no. 5, pp. 515–517, 2009.

[25] B.-Q. Li, T. Huang, L. Liu, Y.-D. Cai, and K.-C. Chou, "Identification of colorectal cancer related genes with mrmr and shortest path in protein-protein interaction network," *PloS one*, vol. 7, no. 4, p. e33393, 2012.

[26] C. Mehner, E. Miller, A. Nassar, W. R. Bamlet, E. S. Radisky, and D. C. Radisky, "Tumor cell expression of mmp3 as a prognostic factor for poor survival in pancreatic, pulmonary, and mammary carcinoma," *Genes & cancer*, vol. 6, no. 11-12, p. 480, 2015.

[27] Z.-S. Zeng, W.-P. Shu, A. M. Cohen, and J. G. Guillem, "Matrix metalloproteinase-7 expression in colorectal cancer liver metastases: evidence for involvement of mmp-7 activation in human cancer metastases," *Clinical Cancer Research*, vol. 8, no. 1, pp. 144–148, 2002.

[28] M. Courel, F.-Z. El Yamani, D. Alexandre, H. El Fatemi, C. Delestre, M. Montero-Hadjadje, F. Tazi, A. Amarti, R. Magoul, N. Chartrel *et al.*, "Secretogranin ii is overexpressed in advanced prostate cancer and promotes the neuroendocrine differentiation of prostate cancer cells," *European Journal of Cancer*, vol. 50, no. 17, pp. 3039–3049, 2014.

[29] I. Gozes, M. Bodner, Y. Shani, and M. Fridkin, "Structure and expression of the vasoactive intestinal peptide (vip) gene in a human tumor," *Peptides*, vol. 7, pp. 1–6, 1986.

[30] Z.-H. Zhou, *Ensemble methods: foundations and algorithms*. CRC press, 2012.

[31] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The weka data mining software: an update," *ACM SIGKDD explorations newsletter*, vol. 11, no. 1, pp. 10–18, 2009.

[32] E. Glaab, J. M. Garibaldi, and N. Krasnogor, "Arraymining: a modular web-application for microarray analysis combining ensemble and consensus methods with cross-study normalization," *BMC bioinformatics*, vol. 10, no. 1, p. 358, 2009.

Functional Magnetic Resonance Imaging, Data Acquisition, Processing, and Applications: A pocket guide

Aleksandra Petrova[1], Ilinka Ivanoska[2], Kire Trivodaliev[3], Slobodan Kalajdiski[4]
Faculty of Computer Science and Engineering
Ss. Cyril and Methodius University
Skopje, Macedonia
aleksandra.petrova@students.finki.ukim.mk[1], ilinka.ivanoska@finki.ukim.mk[2], kire.trivodaliev@finki.ukim.mk[3], slobodan.kalajdziski@finki.ukim.mk[4]

***Abstract*— The human brain is considered to be one of the most complex things in our lives. As an object of research, the brain has been a challenge among the scientists for centuries. Recent solutions have led for scientists to be able study the brain on live subjects, by gathering data in a form of visual recordings of the structure and function of the nerve system. Functional neuro-images such as functional magnetic resonance imaging (fMRI), able us to detect activities in certain areas of the brain, which may differ from a normal state, thus allowing doctors to diagnose and track the effectiveness of diseases. Our aim is to provide an analysis of fMRI data acquisitions, preprocessing, processing and visualizations. We also propose possible software tools that might be suited for these types of analysis. The study is concluded with some potential applications.**

Keywords*— *functional magnetic resonance imaging, preprocessing, processing, analysis

I. Introduction

Neuro imaging includes usage of different techniques of direct or indirect scanning of the structure and function of the neuro system and is a relatively new discipline in medicine, neuro-sciences and psychology [3]. The neuro imaging taken into consideration for this paper can be divided into two big categories: Diffusion Tensor Imaging (DTI) [1] which focuses on the structure of the nervous system and Functional Magnetic Resonance Imaging (fMRI) [2] which focuses on the functional dependency of the components of the nervous system. The technique of making scans with a functional magnetic resonance on the brain is an unobtrusive way to access the function of the brain using the changes in the MRI signal associated with the functional activity of the brain. The most used method of acquiring fMRI scans is based on the changes in the Blood Oxygenation Level Dependent (BOLD) [4]-[6] signal, which occur because of the hemodynamic and metabolic consequences from the neuron answers.

As fMRI gets more and more of a widespread tool among scientists for studying the human brain in both healthy conditions and in disease, it also at the same time requires a lot of comprehension on many factors that influence the knowledge extraction from the workflow. These factors are compounded of preprocessing steps and analysis in several software packages. Having to work with the different steps and software tools can be quite complex which is why our aim is to provide you with this pocket guide with basic information and main points to consider when working on fMRI studies. Inside this paper we propose some guidelines on how you can approach these steps, and some points on how to acquire and which software tools you can use for analysis and statistical analysis.

II. Data Acquisition

fMRI scans are obtained in medical clinics, as their main purpose is to conduct scans on patients. There are different types of scanners used which give different types of neuro images. Due to the different response of Hemoglobin to magnetic fields, the fMRI scans need to have a stronger magnetic field (higher than 1.5T) so that MR signal is stronger as a result of blood being highly oxygenated. Besides the images having a magnetic field some other properties that can be measured in the neuro image are: Echo Planar Imaging (EPI) sequence with two parameters Time Repetition (TR) divided by Time Echo (TE), Flip Angle (FA), Field of View (FOV), dimensions of the Voxel and number of time points. All of these parameters can vary depending on the scanner.

Regarding data acquisition from sources [7], most of the data you can find is from clinics of patients which have a specific disease or patients on which a certain task has been performed on. This type of data is not always open, however there are some repositories in which you can find fMRI dataset scans [9]. Such type of a repositories are OpenfMRI [8], Adni [36], Cobre ect. The data is organized in a format know as Brain Imaging Data Structure (BIDS). The format includes data types as: Subject-level variables, Longitudinal and multi-session studies, Structural anatomical imaging data, Resting state and task-based fMRI data etc. The format includes data types as: Subject-level variables, Longitudinal and multi-session studies, Structural anatomical imaging data, Resting state and task-based fMRI data etc. These types are later taken into account when processing the data, depending on the form of studies, different data types are required in order to obtain the more accurate results.

The datasets found are raw unprocessed files in the Neuroimaging Informatics Technology Initiative (NIfTI) [10] format which is the most common format for neuro images. Besides the NIfTI format images can also be found in DICOM format which contains the same information as NIfTI such as TR, resolution, image orientation and others plus an additional header.

III. Software Tools

Before we get into the steps that go into preprocessing, we give a brief explanation of the software tools selected for this survey. The tools given here have as a main purpose to provide analysis and statistical analysis to the data that is being processed, according to the studies published that use these tools, which are very similar to the studies we propose in this paper. The tools we propose can be found in Table I along with their official websites.

TABLE I. Software tools for analysis and statistical analysis

fMRI software/toolbox	Website
AFNI	http://afni.nimh.nih.gov/afni
BrainVoyager	http://www.brainvoyager.com/
BROCCOLI	https://github.com/wanderine/BROCCOLI/
CONN	https://www.nitrc.org/projects/conn/
DPARSF	http://rfmri.org/DPARSF
ANTsR	http://stnava.github.io/fMRIANTs
Freesurfer	http://freesurfer.net/
FSL	http://fsl.fmrib.ox.ac.uk/fsl/fslwiki/
Nipy	http://nipy.org/
SPM	http://www.fil.ion.ucl.ac.uk/spm/

AFNI
The software [12] is a powerful tool for visualization, transformation and combination of the three dimensional functional neuroimaging results. The software can overlay neural activation maps onto higher resolution anatomical scans. Slices in each cardinal plane can be viewed simultaneously.

BrainVoyager
The software [13] is designed to allow analyses that would exploit optimally the high-resolution information available in fMRI data, to integrate volume-based analysis and cortex-based analysis with the option to visualize topographic activation data on flattened cortex representations and to combine hypothesis testing with data-driven analysis including interactive visualization tools so that data can be easily explored.

BROCCOLI
This software [14] can be used for parallel analysis of fMRI data on a large variety of hardware configurations. BROCCOLI (running on a GPU) can perform non-linear spatial normalization to a 1 mm3 brain template in 4–6 s, and run a second level permutation test with 10,000 permutations in about a minute. These non-parametric tests are generally more robust than their parametric counterparts, and can also enable more sophisticated analyses by estimating complicated null distributions. Additionally, BROCCOLI includes support for Bayesian first-level fMRI analysis.

CONN
The Conn software tool [15] offers a common framework to define and perform a large suite of connectivity analyses, including bivariate/semi partial correlations, bivariate/multivariate regression, seed-to-voxel connectivity, ROI-ROI connectivity, novel voxel-to-voxel connectivity, and graph theoretical measures for both resting state and task fMRI data.

DPARSF
DPARSF is a user-friendly software tool [16] for "pipeline" data analysis of resting-state fMRI. It can help the users to save time for data processing and reduce errors in cumbersome setting of parameters. DPARSF can also create a report for excluding subjects with excessive head motion and generate a set of pictures for easily checking the effect of normalization.

ANTsR
The ANTS framework [37] provides open-source functionality for deformable normalization with large deformations. Small deformation mappings and segmentation tools are also available. The software tool enables diffeomorphic normalization with a variety of transformation models, optimal template construction, multiple types of diffeomorphisms, multivariate similarity metrics, diffusion tensor processing and warping, image segmentation with and without priors and measurement of cortical thickness from probabilistic segmentations.

Freesurfer
This software [17] is a suite of tools for the analysis of neuroimaging data that provides an array of algorithms to quantify the functional, connectional and structural properties of the human brain. With an aim to automatically create models of most macroscopically visible structures in the human brain given any reasonable T1-weighted input image.

FSL
FSL (the FMRIB Software Library) [38] is a comprehensive library of analysis tools for functional, structural and diffusion MRI brain imaging data.

Nipy
This software [39] provides options to visualize the data in a 3D graphical display as well as a library of image registration and segmentation routines. The idea behind NIPY is to become the standard analysis library in neuroimaging in the medium term, providing the ability to call routines in other packages that are more familiar to researchers.

SPM

SPM is a software tool [18] which enables the integration of probabilistic cytoarchitectonic maps and results of functional imaging studies. The tool includes the functionality for the construction of summary maps combining probability of several cortical areas by finding the most probable assignment of each voxel to one of these areas. Its main feature is to provide several measures defining the degree of correspondence between architectonic areas and functional foci.

IV. Preprocessing

Preprocessing is one of the key steps to providing reliable results in the study. There are many datasets that can be found preprocessed, however they do not always have the quality required for the study we want to conduct.

Before preprocessing the data, it is common to start by checking the image format. In case of having a different format than the one required by the software tool, there are file format converters such as MRIcro [40], dcm2nii [41], MRIConvert [42], NiBabel [43], and some software tools have their own converters. Afterwards the first step is importing the data, checking if all of the subjects have an image as well if they are sorted correctly. This can be performed using the viewers in the software tools.

The reason why preprocessing has many steps necessary in because the fMRI images which are three-dimensional image volumes are typically collected one two-dimensional image slice at a time. These images when later put together are susceptible to individual slice artifacts due to timing errors and ration frequency spikes. To correct these artifact there are multiple steps that go into the preprocessing, most of them play a crucial role to the processing that comes later on.

A. Slice Timing

Once the subjects have been ensured, the preprocessing steps [11] can be implemented. One of these steps is slice timing [20], which aims to adjust the time-course of voxel data in each slice to account for these differences by interpolating the information in each slice to match the timing of a reference slice. There are two experimental approaches that integrate slice timing corrections with slice registration: one of them is mapping of single fMRI slices to a reference MRI collected in the same scanning session [32], a method known as map-slice-to-volume (MSV) approach, and the method of constrained slice motion between sequential volumes [33].

B. Motion Correlation and Outlier Detection

Physical head motion cannot be completely eliminated while scanning, but it may be reduced with using some immobilization when placing the subject in the scanner and with training in an MRI simulator [34]. In many cases the scans are not fully free from these motions, this is why we have motion correction [21] that is used to realign each volume to a reference volume (mean image, first, or last volume) using a rigid body transformation (x, y, and z rotations and translations). The threshold for motion correction is smaller voxels to limit the gradient change per voxel, but in this case it is required to increase the scan time and reduce the signal to noise/voxel. As a part of motion correction is motion outlier detection so that any possible outlier identified can be regressed out in order for it not to interfere with the analysis.

C. Spatial Transformation and Smoothing

The brain size, shape, orientation, and gyral anatomy vary largely across subjects. For these purposes another group of preprocessing steps is spatial transformation. These transformations are performed so that the images from the individual's native space are aligned with those acquired from a different modality or subject or into a common standard space. There can be two types of transformations, linear applied uniformly along an axis and usually represented as affine matrices, and non-linear, defined locally usually defined by warp or distortion maps. Following the spatial transformation as a step in preprocessing is spatial smoothing [22], a step in which during which data points are averaged with their neighbors, suppressing high frequency signal while enhancing low frequency ones, and results in the blurring of sharp edges. The smoothing method is usually consisted of convolving the fMRI signal with a Gaussian function of a specific width. These trasformations are consired to be one of the biggest impacts when it comes to preprocessing and it should definitelly be considered in the pipeline [35].

D. Temporal Filtering

The final step that can be included in the preprocessing is temporal filtering [23], with the objective to remove the effects of confounding signals with known or expected frequencies. This step can be split into low-pass filtering known as temporal smoothing for control and estimation of temrporal autocorrelations, and high-pass filtering for control low-frequency noise. The noise we get in the majority of the scans occurs at low frequencies and those are usually detected within the gray matter, whereas the high frequency noise is detected in the white matter and is found within physiological reactions such as aliased cardiac and respiratory signals.

The given steps are not all of the steps available for preprocessing, however these are the steps that are of importance to consider before proceeding with statistical analysis. Not all of the software tools offer every preprocessing step as they are designated for specific analysis. In Table II we give a summary of the steps given in each software tool, that we found of importance to analysis and statistical analysis. Some of the tools offer either pipelines that have been created precisely for the usage of that software tool or they offer the option to insert a predefined pipeline universally used, such as The Athena (offered in AFNI and FSL), The Burner (offered in SPM), The NIAK, The CIVET etc. These pipelines can be found online alongside with the steps that they offer and why they have been chosen. We also provide information regarding pipeline options in Table II.

TABLE II. PREPROCESSING STEPS OFFERED IN DIFFERENT SOFTWARE TOOLS

fMRI software/toolbox	Preprocessing steps						Pipeline options	
	Slice Timing	*Motion Correlation*	*Motion outlier detection*	*Spatial Transformation (Normalization)*	*Spatial Smoothing*	*Temporal Filtering*	*Provides pipelines*	*Can import pipelines*
AFNI	YES	YES	YES	YES	YES	YES	NO	YES
BrainVoyager	YES	YES	NO	YES	YES	YES	YES	NO
BROCCOLI	YES	YES	NO	YES	YES	YES	YES	No data found
CONN	YES	NO	NO	YES	NO	YES	YES	YES
DPARSF	YES	YES	YES	YES	YES	YES	NO	NO
ANTsR	NO	YES	NO	YES	YES	NO	NO	NO
Freesurfer	YES	YES	NO	YES	YES	NO	YES	No data found
FSL	YES	YES	YES	YES	YES	YES	YES	YES
Nipy	YES	YES	YES	YES	YES	YES	NO	YES
SPM	YES	YES	NO	YES	YES	YES	YES	YES

V. PROCESSING

When it comes to processing the data, the analysis taken differ depending on the study made. Most of the studies are designed around two types of states that can be conducted on subjects, resting state [24] and task-based [25]. The resting state represents a state where the subjects being scanned are not performing a task. In the resting state subjects are in a state where they have to keep their eyes closed, not to think about anything in particular and not falling asleep. Under these conditions the patterns found in the images are stable. Where as in the task-based state, subjects are presented to some stimuli as a series of times which change from one condition to another. The order of the conditions is also important, as the aim is for them to be counter-balances. The images found within this state vary more, as they display hemodynamic responses.

As a part of the software tools we find options to visualize the subjects we are studying. They can be visualized through the given images viewed in all of the planes, surfaces of the images, graphs that display the connectivity of the brain, surfaces displaying the network in the brain, surfaces displaying the activated fragment of the brain etc. It is common to have the option to display the brain as a graph, depending on the type of connectivity of the brain that is being researched upon.

Additionally to the visualization, as results during the processing phase we often get statistics regarding on the properties we are examining. These statistics are necessary to form the results to the hypothesis given for the study. The statistics are often predefined and provided from software tools [19]. The tools given here have as an aim to deliver statistics on the data given. Some are oriented for resting states, some for task-based and some for both. In Table III are given the options for each of the tools for both types of analysis.

TABLE III. TYPES OF STATISTICS IN DIFFERENT SOFTWARE TOOLS ACCORDING TO THE TYPE OF STATE

fMRI software/toolbox	Resting state	Task-based state
AFNI	YES	YES
BrainVoyager	YES	YES
BROCCOLI	YES	YES
CONN	YES	YES
DPARSF	YES	NO
ANTsR	NO	YES
Freesurfer	NO	YES
FSL	YES	YES
Nipy	NO	YES
SPM	YES	YES

VI. APPLICATION

fMRI studies have been emerging in the recent years as it has led to multiple understandings in the field of cognitive science. There are many studies that occur in diverse aspect of the cognitive science such as: languages [26], motor functions [27], sensors, memory [28], decision-making etc. fMRI studies can be also used to detect activities in certain areas of the brain, which may differ from a normal state, thus facilitate the diagnosis and evaluation of treatment effectiveness for diseases affecting the brain. Mental disorders are one of the prominent areas in this line of research. These types of research could potentially be used in the future to give prognostic and diagnostic information simply by looking at a scan [29]-[31].

The software tools presented in this paper are focused on providing statistical analysis for the studies conducted, however they can these analyses can serve for studies in different fields. The Table IV given bellow provides the number of citations per paper that represents the given software tool and the number of cited papers that use the software tool and focus on one of the areas given. The areas selected are the ones that we found most common among the studies conducted with the chosen software tools.

TABLE IV. NUMBER OF CITATIONS OF THE PAPERS FOUND FOR THE SOFTWARE TOOLS AND THE MOST COMMON FIELD THEY ARE USED IN

fMRI software/toolbox	Cited	Main area of focus		
		Statistical	Medical	Technical
AFNI	6517	5420	4560	1960
BrainVoyager	74	59	54	42
BROCCOLI	41	34	35	10
CONN	716	639	521	195
DPARSF	1407	1220	1050	204
ANTsR	186	145	160	73
Freesurfer	1374	1140	1040	443
FSL	2638	2200	1960	915
Nipy	304	235	214	121
SPM	2639	2370	1680	844

Out of all of the tools given here we can observe that the most popular ones are AFNI, FSL, SPM. These tools are quite used as they have been developed earlier than some of the others provided here. Another positive side is that they do not require specific hardware to be used on. As previously shown in this paper these tools also include a lot of preprocessing steps, alongside with options to process and visualize the data. Some of the other tools given in this table are based upon the popular ones, such as CONN that uses SPM integrated.

As seen in the table above many of the software tools are being used for medical purposes, these studies include better understanding of how the brain works such as studies on neuronal dynamics, anatomical structure, regions of interest etc. Another set of medical studies is connected with emotions, morality, empathy. Last but not least another type of study that is conducted with medical purposes is meta-analysis. All of these studies show the emerge of medical discoveries in the area of fMRI. These studies have been successfully conducted with the usage of such tools as we have provided in this paper. Using medical problems in studies of fMRI has become recently very popular, which is why we provide you with some numbers in Table V of few of the types of problems that can be found among studies and have used the tools listed. The numbers given are based on the citations of papers that focus on medical problems which have cited the tool used for their study.

TABLE V. NUMBER OF CITATIONS OF THE PAPERS FOUND FOR THE SOFTWARE TOOLS THAT TACKLE A MEDICAL PROBLEM

fMRI software/toolbox	Medical problems		
	Mental Disorders	Emotions	Meta-analysis
AFNI	2550	985	2570
BrainVoyager	19	7	31
BROCCOLI	4	0	13
CONN	376	131	438
DPARSF	785	251	723
ANTsR	35	11	47
Freesurfer	491	103	554
FSL	838	230	995
Nipy	12	1	12
SPM	961	551	1450

CONCLUSION

Functional MRI is currently one of the most projecting fields in the world of research, assuring to have a bright future as we combine it more with computer science. Although now we can use the knowledge of computer science to obtain better results for our studies, having a comprehension for all of which fMRI is consistent of is crucial too. With fMRI growing in neuroscience, we get to work with improved images, higher sensitivity which plays a key role in choosing our models and processing our data. Our aim was to provide some guidelines and references to the tools most often used nowadays when it comes to making statistical analysis. We give our take on the steps that we consider that should be taken into account when working with fMRI, from the data acquisition, preprocessing to processing together with future ideas on where they can be used.

REFERENCES

[1] Soares, Jose, et al. "A hitchhiker's guide to diffusion tensor imaging." Frontiers in neuroscience 7 (2013): 31.

[2] Soares, José M., et al. "A hitchhiker's guide to functional magnetic resonance imaging." Frontiers in neuroscience 10 (2016): 515.

[3] Keiriz, Johnson JG, et al. "Exploring the Human Connectome Topology in Group Studies." arXiv preprint arXiv:1706.10297(2017).

[4] Aguirre, G. K., Zarahn, E., and D'Esposito, M. (1998). The variability of human, BOLD hemodynamic responses. Neuroimage 8, 360–369. doi: 10.1006/nimg.1998.0369

[5] Cardenas, D. P., Muir, E. R., Huang, S., Boley, A., Lodge, D., and Duong, T. Q. (2015). Functional MRI during hyperbaric oxygen: Effects of oxygen on neurovascular coupling and BOLD fMRI signals. Neuroimage 119, 382–389. doi: 10.1016/j.neuroimage.2015.06.082

[6] Faro, S. H., and Mohamed, F. B. (2010). BOLD fMRI: A Guide to Functional Imaging for Neuroscientists. New York, NY: Springer

[7] Turner, Robert, et al. "Functional magnetic resonance imaging of the human brain: data acquisition and analysis." Experimental Brain Research 123.1-2 (1998): 5-12.

[8] Poldrack, Russell A., and Krzysztof J. Gorgolewski. "OpenfMRI: open sharing of task fMRI data." NeuroImage 144 (2017): 259-261.

[9] Poldrack, Russell A., et al. "Toward open sharing of task-based fMRI data: the OpenfMRI project." Frontiers in neuroinformatics 7 (2013): 12.

[10] Poldrack, Russell A., Jeanette A. Mumford, and Thomas E. Nichols. Handbook of functional MRI data analysis. Cambridge University Press, 2011.

[11] Strother, Stephen C. "Evaluating fMRI preprocessing pipelines." IEEE Engineering in Medicine and Biology Magazine 25.2 (2006): 27-41.

[12] Cox, Robert W. "AFNI: software for analysis and visualization of functional magnetic resonance neuroimages." Computers and Biomedical research 29.3 (1996): 162-173.

[13] Goebel, Rainer. "BrainVoyager—past, present, future." Neuroimage 62.2 (2012): 748-756.

[14] Eklund, Anders, et al. "BROCCOLI: Software for fast fMRI analysis on many-core CPUs and GPUs." Frontiers in neuroinformatics 8 (2014): 24.

[15] Whitfield-Gabrieli, Susan, and Alfonso Nieto-Castanon. "Conn: a functional connectivity toolbox for correlated and anticorrelated brain networks." Brain connectivity 2.3 (2012): 125-141.

[16] Yan, Chaogan, and Yufeng Zang. "DPARSF: a MATLAB toolbox for" pipeline" data analysis of resting-state fMRI." Frontiers in systems neuroscience 4 (2010): 13.

[17] Fischl, Bruce. "FreeSurfer." Neuroimage 62.2 (2012): 774-781.

[18] Eickhoff, Simon B., et al. "A new SPM toolbox for combining probabilistic cytoarchitectonic maps and functional imaging data." Neuroimage 25.4 (2005): 1325-1335.

[19] Behroozi, Mehdi, and Mohammad Reza Daliri. "Software tools for the analysis of functional magnetic resonance imaging." Basic and Clinical Neuroscience 3.5 (2012): 71-83.

[20] Sladky, Ronald, et al. "Slice-timing effects and their correction in functional MRI." Neuroimage 58.2 (2011): 588-594.

[21] Johnstone, Tom, et al. "Motion correction and the use of motion covariates in multiple-subject fMRI analysis." Human brain mapping 27.10 (2006): 779-788.

[22] Mikl, Michal, et al. "Effects of spatial smoothing on fMRI group inferences." Magnetic resonance imaging 26.4 (2008): 490-503.

[23] Woolrich, Mark W., et al. "Temporal autocorrelation in univariate linear modeling of FMRI data." Neuroimage 14.6 (2001): 1370-1386.

[24] Van Den Heuvel, Martijn P., and Hilleke E. Hulshoff Pol. "Exploring the brain network: a review on resting-state fMRI functional connectivity." European neuropsychopharmacology20.8 (2010): 519-534.

[25] Hermundstad, Ann M., et al. "Structural foundations of resting-state and task-based functional connectivity in the human brain." Proceedings of the National Academy of Sciences110.15 (2013): 6169-6174.

[26] Spreer, J., et al. "Determination of hemisphere dominance for language: comparison of frontal and temporal fMRI activation with intracarotid amytal testing." Neuroradiology 44.6 (2002): 467-474.

[27] Mattay, Venkata S., et al. "Neurophysiological correlates of age-related changes in human motor function." Neurology58.4 (2002): 630-635.

[28] Strandberg, Maria, et al. "fMRI memory assessment in healthy subjects: a new approach to view lateralization data at an individual level." Brain imaging and behavior 5.1 (2011): 1-11.

[29] Machulda, Mary Margaret, et al. "Comparison of memory fMRI response among normal, MCI, and Alzheimer's patients." Neurology 61.4 (2003): 500-506.

[30] Koshino, Hideya, et al. "Functional connectivity in an fMRI working memory task in high-functioning autism." Neuroimage24.3 (2005): 810-821.

[31] Cortese, Samuele, et al. "Toward systems neuroscience of ADHD: a meta-analysis of 55 fMRI studies." American Journal of Psychiatry 169.10 (2012): 1038-1055.

[32] Kim, Boklye, et al. "Motion correction in fMRI via registration of individual slices into an anatomical volume." (1999).

[33] Bannister, Peter R., J. Michael Brady, and Mark Jenkinson. "TIGER–a new model for spatio-temporal realignment of fMRI data." Computer Vision and Mathematical Methods in Medical and Biomedical Image Analysis. Springer, Berlin, Heidelberg, 2004. 292-303.

[34] Huettel, Scott A., Allen W. Song, and Gregory McCarthy. Functional magnetic resonance imaging. Vol. 1. Sunderland: Sinauer Associates, 2004.

[35] LaConte, Stephen, et al. "The evaluation of preprocessing choices in single-subject BOLD fMRI using NPAIRS performance metrics." NeuroImage 18.1 (2003): 10-27.

[36] Jack, Clifford R., et al. "The Alzheimer's disease neuroimaging initiative (ADNI): MRI methods." Journal of magnetic resonance imaging 27.4 (2008): 685-691.

[37] Avants, Brian B., Nick Tustison, and Gang Song. "Advanced normalization tools (ANTS)." Insight j 2 (2009): 1-35.

[38] Jenkinson, Mark, et al. "Fsl." Neuroimage 62.2 (2012): 782-790.

[39] Millman, K. Jarrod, and Matthew Brett. "Analysis of functional magnetic resonance imaging in Python." Computing in Science & Engineering 9.3 (2007).

[40] Rorden, Chris, and M. Brett. "MRIcro." Availabile online at: http://www. sph. sc. edu/comd/rorden/mricro. html (2005).

[41] Rorden, C. "Dcm2nii DICOM 2 NIFTI conversion." (2007).

[42] Smith, J. "MRIConvert. The Lewis Center for Neuroimaging, University of Oregon." (2011).

[43] Brett, Matthew, et al. "nibabel: 2.1. 0." Zenodo (2016).

TurtleBot: Navigation and Image Processing

Marija Todosovska
Faculty of Computer Science and Engineering
Ss. Cyril and Methodius University
Skopje, Macedonia
mtodosovska@gmail.com

Simon Hermann
Institute of Physics
Technical University Berlin
Berlin, Germany
simon.Hermann@posteo.de

Dajana Stojchevska
Faculty of Computer Science and Engineering
Ss. Cyril and Methodius University
Skopje, Macedonia
dajana.sk@hotmail.com

***Abstract*—The problem of this project was to create a solution which would allow a TurtleBot to explore a bordered space under specific circumstances. A TurtleBot is a low-cost, personal robot kit with open-source software. The problem was set as a bordered field in which there were a number of faces, rings and cylinders. The robot would explore the field, recognize the faces in it, find the rings and cylinders, and pick up the rings. It would talk with the people in order to infer information and perform specific tasks.**

We used ROS and OpenCV for the solution to this problem. The robot scanned the field using the Kinect sensor and stored a map of it. We used the canny edge detector, and the k nearest neighbors algorithm provided by OpenCV, and vector transformations in order to approach objects on the field. We used the ViCos Lab face-detector for face detections. For the cylinder detection we used the Point cloud library and the RANSAC algorithm. We implemented the speech recognition using the Google recognition application. The general logic of our implementation can be split into two main parts: the search for faces, rings and cylinders, and the conversation with the people. We performed the integration of all these parts during runtime by calling specific scripts in a specific order.

The ring and cylinder detection was implemented well and worked correctly. All the faces were detected and recognized correctly, and the logic of the solution worked well and made the correct conclusions. Altogether, the solution performed well to the task, with the minor exception of not always picking up all rings.

Keywords—machine vision, intelligent navigation, robot navigation, face detection, face recognition

I. Introduction

A TurtleBot is a low-cost, personal robot kit with open-source software[1]. A standard TurtleBot has a mobile base, 3D sensor, a mounting hardware kit, and a laptop computer. The TurtleBot we used was built on the base of an iRobot Roomba. Changes were made to the standard base, by removing the vacuum module, and replacing it with a removable battery. As well as this, our model used the Microsoft Kinect sensor. The TurtleBot uses ROS[2] [1], and can be programmed using the TurtleBot SDK, it offers solutions in both C++ and Python. The ROS wiki[3] offers a specific section[4] that contains the TurtleBot SDK, as well as extensive tutorials on programming the TurtleBot both in C++ and Python.

[1]https://www.turtlebot.com/ (last visited: 17.04.2018)
[2]http://www.ros.org/ (last visited: 17.04.2018)
[3]http://wiki.ros.org/ (last visited: 17.04.2018)
[4]http://wiki.ros.org/Robots/TurtleBot (last visited: 17.04.2018)

The problem was set in the following way. There was a bordered field in which the robot could move. In it there were 12 people (six women and six men). There were four different colored cylinders, representing towers, and four different colored rings, hanging on special hooks mounted on the walls. The robot was supposed to find out which of the rings was magical, and to which tower it should be taken by asking the people questions. It was then supposed to collect the magical ring, for this purpose it was equipped with a lance, and bring it to the correct tower. The rules were as follows. One of the people in the field knew which the magical ring was, and another one knew which the correct tower was. Women could answer any type of question, but half of them always lied. Men could only answer yes or no questions, but they always told the truth. The robot could ask women only one question at a time, and men two questions at a time. The robot was supposed to explore the field and create a map of it, detect and recognize the faces, find and collect the magical ring, detect the cylinders and discover which one was the correct one, and finally bring the magical ring to the correct cylinder.

Both the TurtleBot and the Kinect sensor have been used for problems similar to this one in the past [2] [3].

II. Methods

A. The Map

The map that we used all throughout the problem was created at the beginning by moving the robot around the field manually allowing for it to scan the field using the Kinect sensor. This was done a couple of times in order to ensure the most accurate result. The scanned map is shown on Fig. 1. On the basis of this map we later created the map in Fig. 5.

B. Ring Detection

The robot would start its performance by detecting the rings on the field. It was our conclusion that it would be easier if we detected and collected every ring at the start, so later the robot would not have to go back in order to pick up the magical ring. We used OpenCV[5] [4] for this purpose, we subscribe to the topic $/camera/rgb/image\_raw$ in order to get all the pictures that the robot 'sees' in a form of $geometry\_msgs :: PoseStamped$. With the help of OpenCV's $cv\_bridge :: CvImagePtr$ we get a pointer

[5]https://opencv.org/ (last visited: 17.04.2018)

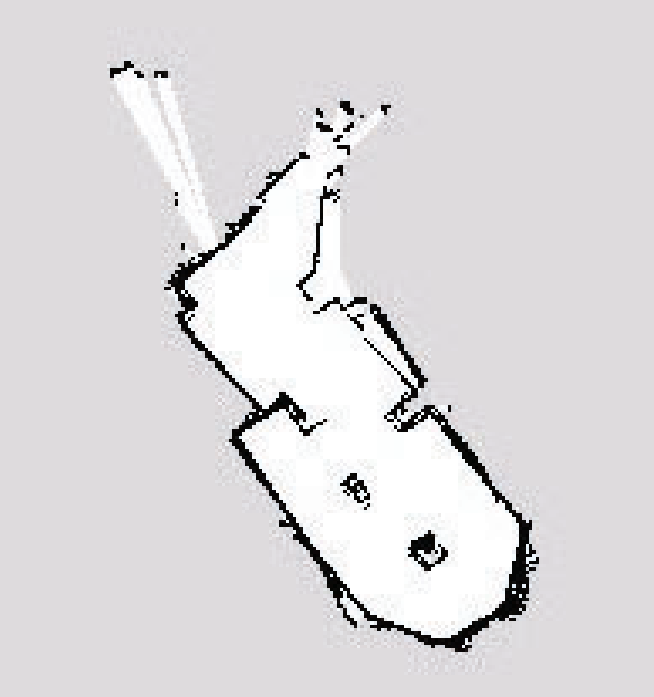


Fig. 1. The map obtained after scanning the field, this map was used for navigation. We can see that the lines are blurred, this is because during the scanning process the robot corrects its map as it gets new data.

to the image that comes from the robot's sensors. Then, we convert it to gray-scale and set the appropriate parameters to it (blur, contrast, brightness). We do this in order to have a clearer picture so that the algorithm can distinguish the rings (circles or ellipses) in the environment better and avoid false detections. The algorithm/transformation applied in order to find the circles is the $HoughCircles$ function from OpenCV. From this we get the pixels on the image where the center of the circle is detected. Here a problem arises when we need the exact location of the center on our map. Transforming the relative positions to coordinates in the map frame is not simple because it tends to detect them somewhere outside the map (because the center is not a point on the ring itself). It is worth mentioning the failure of the localizer, here. It was provided by the ViCos ROS Lab[6] and maps the RGB image with the depth information of the Kinect sensor, which often failed to give us the right positions and often returned zeros. Another problem we faced, often, was that the data of the camera was too old, which again resulted in a failure of locating the detections. Regarding mapping the rings, as well as faces on the map, we filter them out by height.

C. Approaching Objects on the Map

In the following section we describe how we calculate the transformation of the relative positions to coordinates given in the map frame. We used this calculation to approach detected faces, detected cylinders and to calculate an initial position for picking up a ring. The strategy for approaching objects can be split into to cases: when the object is located on a wall, and when the object is located on a box.

[6]https://github.com/vicoslab/vicos_ros (last visited: 17.04.2018)

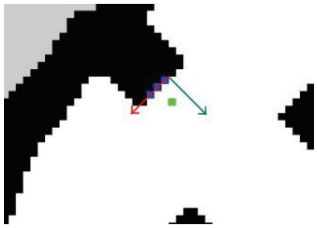
D. Approaching Objects Located on a Wall

Initially, we get a point in the map frame close to a wall. Using the Canny edge detector, provided by OpenCV, from the map, which the TurtleBot uses to navigate, we obtain an array of pixels along the walls. Using a k nearest neighbor algorithm, also provided by OpenCV, we can find the closest points on the wall to the detected object. To account for imperfections within the map, e.g. walls not being perfectly straight, we consider the three closest points of the wall to the detected point and use the function $fitLine$ provided by OpenCV to get a 2D vector pointing along the wall. Generally it would have been possible to consider larger amounts of points along the wall, in order to have a better estimation of the wall orientation, but in cases of objects close to corners this would increase the risk of considering too many points belonging to a part of the wall with a different orientation than the one needed. After obtaining the orientation of the wall we still need to verify in which of the two possible directions the vector is pointing. This is done by turning the vector clockwise by 90 degrees which results in a normal vector to the wall's orientation. By adding the normal vector to one of the points on the wall and subtracting it once, we can see which of two resulting points is within the map and which is on the wall or outside the map. Therefore, we know if the normal vector is pointing inside the map or outside the map. Knowing the direction in which the normal vector is pointing tells us the direction in which the initial vector is pointing. By normalizing both vectors we can calculate any arbitrary position towards the given object with respect to the orientation of the wall. An illustrated example of the computation can be seen in Fig. 2a. In the cases of the faces and cylinders, the calculated goals are simply the initial position plus the normal vector pointing inside the map times approximately 15 cm. In the case of approaching the faces we can add an additional improvement to finding the orientation of the closest wall. We found out that sometimes the localization of the faces went wrong and the obtained position of the face was inside the wall rather then on the wall itself. In cases of straight walls this presented no problem for the calculation of the wall's orientation.

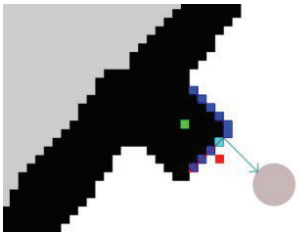
E. Approaching Objects Located on a Box

Contrary to the previous case, if the face was positioned on a box, there were 4 different orientated walls that were close by. This could easily result in the wrong part of the box being considered as the closest wall and therefore returned as a wrong orientation. To prevent such results we estimated that a face recognition would not be possible with an angle bigger then 45 degrees, when the normal of the face and the orientation of the robot were compared. In this case, we can assume that the wall closest to the robot, of all the walls close to the face, has to be the wall on which the face is situated. Therefore, we modified the calculation by first computing a

bigger set of closest points to the estimated position of the face, e.g. 10 points. After this, we sort these points by the distance to the robot's current position. The closest of these points to the robot's current position is assumed to be on the right part of the wall. Taking the two closest points to the point on the wall, we compute the orientation of the wall again and proceed as before. The sort function of the C++ $std :: vector$ class is used with an implementation of Euclidean distance for the given coordinates for sorting the points given by the k nearest neighbor computation. An illustrated example can be seen in Fig. 2b. The same assumption could unfortunately not be made for ring detections. As the rings have a 90 degree orientation towards the closest wall, the normal of the rings would point along this wall's orientation rather then being normal to it. Therefore the closest wall to the robot would definitely result in a false wall orientation.



(a) Appropriate accurate localization of the object and the resulting orientation vectors.



(b) Badly located object (in this case a face) where the orientation can still be obtained by comparing the distance of the k nearest neighbors with the robot's position. There is a red dot representing the actual position of the object and a cyan dot, which was chosen to be the closest to the robot and therefore used for the calculation of the direction vectors.

Fig. 2. Two examples of calculating the orientation of the wall (red arrow) and the normal of the wall (cyan arrow). The green dot represents the localization of the object. The blue dots are the k nearest neighbors and the gray circle represents the robot.

F. Picking up Rings

We faced two major issues when picking up the rings. First, due to the implementation of the detection of rings, using $HoughCircles$ given by OpenCV, we only detected rings if the angle between the robot's orientation and the normal of the circle was small. Second, within the localization of the detected rings on the map frame we lost many detections, as was mentioned in Section II-B. Furthermore the localization of the rings was sometimes partly incorrect and therefore either within a wall or too far away from a wall. To account for all these uncertainties the following algorithm was implemented.

G. Picking up Rings: The Algorithm

1) Calculate an initial position of the robot using the first located position of the ring. Following the scheme presented in the previous section the closest point on the wall $\vec{P}_{Wall}$ is calculated as well as the normal of the wall $\vec{n}$, facing inside the map, and the orientation of the wall $\vec{v}$ which is the normal vector rotated by 90 degrees clockwise. All the direction vectors are normalized. The initial position $\vec{P}_{init}$ is then calculated by the following formula: $\vec{P}_{init} = \vec{P}_{Wall} + 25cm \cdot \vec{n} - 50cm \cdot \vec{v}$. Using this formula we made sure that we would always pick up the ring from the right side, given that $\vec{n}$ and $\vec{v}$ are calculated correctly. Furthermore the robot pose was orientated along vector $\vec{v}$.
2) Try to re-detect the ring without localizing the detections on the map frame, to ensure that we do not loose any detections:
   - If the ring is re-detected, align the robot so that the x-coordinate of the ring center is within a certain interval by turning left or right, depending on whether the x-coordinate is larger or smaller then the given interval. After this, check if the next detection is within the interval and turn accordingly. Repeat this until the robot is aligned within the given interval. As the detections had small fluctuations within positioning the center of the circle we only considered averages of 10 detections.
   - If there are no detections within 5 seconds, move back for a short distance and wait again. Do this up to two times. If there are still no detections, take the initial pose of the robot for the approach of the ring.
3) Move forward to pick up the ring and backwards afterwards to check if the ring is still there.
4) Again, wait up to 5 seconds for detections of the ring. If there are none move back up to 2 times and wait each time again for 5 seconds. If there are any detections at all start the whole process again beginning at step 2. Else assume that the ring is picked up.

H. Picking up Rings: The Setup

The prefactor of $\vec{n}$ was chosen to be 25 cm as the rings were approximately 15 cm away from the wall and the lance of the robot approximately 10 cm away of the center of the robot. The prefactor of $\vec{n}$ was chosen empirically, as the distance of 50 cm was sufficiently far, to have a high probability of the re-detection while not being too far, to create problems with imperfections in the odometry. If we managed to detect the ring after reaching the initial position we had a high success rate. If the ring was not picked up within the first try we would normally also re-detect the ring when checking if the pick up was not successful and we could pick the ring up in the second or third try. Despite this, in the second and third try the success rate of pickups became worse due to the alignment of the robot getting less accurate while moving back because we used the coordinates within the camera frame. Furthermore, the calculation of the initial position of the robot failed often, as the initial detection of the ring was located too far from its actual position and the wrong wall was chosen to calculate the initial position. If the ring was located close to the borders of the map, these wrongly calculated initial positions were often

put outside the map. In such a case the program would return a failure to the logic, indicating the ring was not picked up.

I. Face Detection and Recognition

We used the ViCos Lab face-detector for the face detection, which uses the DLIB[7] frontal face detector and returns a rectangle approximately marking the spot where the detected face should be. The resulting images, with sizes mostly between 80x80 pixels and 120x120 pixels, are reduced to 80x80 pixels sized images. If a detection results with a position of the face such that a 80x80 rectangle could not be fitted, the detection is omitted. The resulting 80x80 images were converted into a vector by putting the 80 pixels of the first row of the image in a vector and appending the next 80 pixels of the second row to the same vector and so forth. For each pixel the integer values of the RGB channels are stored, so that it represents the pixel in the i-th row and the j-th column with representing the different color channels, the first 4 vector entries are: $v_1 = p_{1,1}(R), v_2 = p_{1,1}(G), v_3 = p_{1,1}(B), v_4 = p_{1,2}(R)$ Using a set of 190 images for each of the twelve people we got 22800 vectors of size 19200. Using Point cloud library (PCL)[8] [5] we create a subspace u_1 which maximizes the variance of the projected input vectors. In a second step using Linear Discriminant Analysis (LDA) we create a subspace $\vec{w}_1$ which maximizes the distances between different classes. The eigenvectors of both subspaces are stored, as well as the mean of the input vectors creating the PCL subspaces $\vec{v}_{mean}$ and the mean of each class within the LDA subspace $\vec{m}_i$.

A new image can be recognized by reducing the image to the 80x80 pixel area in which the face is supposed to be and creating a vector $\vec{v}$ as described before. The resulting vector is first centered by subtracting the mean input vector $\vec{v}_c = \vec{v} - \vec{v}_{mean}$ and then projected onto the PCL subspace: $a_i = \vec{v}_c \vec{u}_i^T$

The obtained vector is then projected on the LDA subspace: $a_i^{LDA} = \vec{a}\vec{w}_i^T$. Comparing the resulting vector $\vec{a}_i^{LDA}$ with the means $\vec{m}_i$ of the different classes, we find the closest class and assume that the input image belongs to this class. As we reduce the images to the part where the face is supposed to be, the recognition is depended on the positioning of the rectangle. This positioning gets less accurate with angle towards the normal of the face increasing or if the robot has a high angular velocity, e.g. is turning. A comparison between badly and well located faces is shown in Fig. 3.

To prevent wrong recognitions due to high angle towards the face normal, each face gets approached before a command is sent to recognize it. Furthermore, a set of 10 recognitions is considered and the face recognized the most times is chosen to be the correct recognition. However, the recognition from a constant position were in 90% of cases all equivalent. Therefore a robust calculation of the position to approach faces was critical for a reliable recognition of faces. A second issue was the changing illumination at different hours and days. An example of different illuminations can be seen in Fig. 4. Using images of different days and different times could somehow account for different illuminations, even though taking new pictures at the given time always increased the overall performance of the recognition. In the end we managed to get, with a constant set of pictures (which we did not change for different times or days), a recognition rate around 70%.

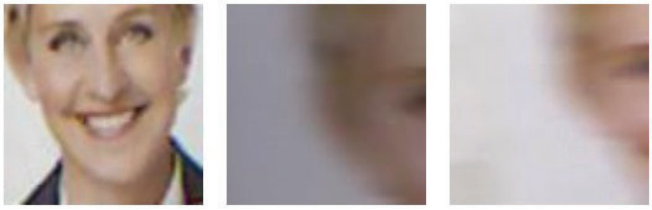


Fig. 3. Different detections of the same face. While the picture on the left is easy to recognize, the recognition probability reduces towards the right, were the actual face is less visible within the rectangle.

Fig. 4. Similar pictures of the same face taken at different days and times, which results in different illuminations of the faces.

J. Cylinder Detection

The cylinder detection was implemented using the point cloud provided by the Kinect sensor of the robot and the PCL library. The point cloud given by the Kinect sensor is first reduced by a voxelgrid algorithm to points of size $1cm^3$. In the next step the point cloud is segmented looking for planar models by using the RANSAC algorithm. Found planar models, such as walls and the ground, are extracted from the point cloud. In the resulting point cloud we segment again using the RANSAC algorithm, looking for cylindric objects. The resulting points are scanned and sorted by their RGB values. If we have found enough points with a certain color, while there are not too many points with a second color we consider to have detected a cylinder. If there are points which do not have values which we consider to be red, green, blue or yellow, they are omitted. We got the best result by setting the lower threshold for the number of points with the correct color to be 10 points. If we detect a bigger amount of points with a second color we omit the detections. The threshold here was 5 points. The appropriate threshold for deciding whether a point was considered red, green, blue or yellow or if it has a different color and would therefore be omitted proved to be crucial for the amount of found points. The best results we obtained with the intervals presented in Table I.

The position of the cylinder was obtained by averaging all the x, y and z-coordinates of the obtained point on the

[7] http://dlib.net/ (last visited: 17.04.2018)
[8] http://pointclouds.org/ (last visited: 17.04.2018)

cylinder. Even though this would not result in the center of the cylinder this approximation was accurate enough to approach the cylinder.

TABLE I
THRESHOLD FOR THE RGB CHANNELS FOR A POINT WHEN DISTINGUISHING WHETHER IT IS RED, GREEN, BLUE OR YELLOW.

	R	***G***	***B***
blue	< 105	< 105	> 110
red	> 110	< 105	< 105
green	> 110	< 105	< 105
yellow	> 150	> 150	< 105

K. Speech Recognition

For the speech recognition we used the Google recognition application[9]. The jsk_common_msgs ROS package is used in order to connect the master to the device where the application is installed. The messages are then sent to the $Tablet/voice$ topic in the form of $VoiceMessage$, arrays of strings that have been heard. After this, we check if any of the words that have been recognized is a word that is expected in the given stage of the program. If it is not, the robot simply asks for a repeat.

L. The General Logic

The general logic of our implementation can be split into two main parts: the search for faces, rings and cylinders, and the interrogation of the people in order to determine the magical ring and the magical tower.

M. Search for Objects

The search of the map was implemented in such a way that would allow for the robot to go around the map in sufficiently many steps, and sufficiently close to the walls so as to recognize faces, rings and cylinders. First, the map is segmented into a grid of squares roughly the size of the robot (the size was chosen so that we can use these squares as goals, and the robot can safely go there). The color of each square is determined from the colors of the pixels of which it consists. Gray for unknown, black for unavailable, and white for available. If all the pixels are white, then the square itself is white. If there is even one black or gray pixel then the square becomes black or gray. A drawing of the map, and how the grid would look is shown on Fig. 5. When exploring the map, the robot would have to be in positions that would allow for it to detect faces, rings and cylinders. In order to achieve this, the exploration strategy would have to cover straight walls, as well as corners, and pillars. We analyzed the grid shown above, and came to the conclusion that there is a relatively finite number of shapes that appear in corners, straight walls, sole standing boxes, and pillars. The 10 shapes we found and used are shown on Fig. 6. After finding the shapes, we analyzed the shapes on the map in order to determine where we would like the robot to stop and spin in regard with walls and corners. Having done

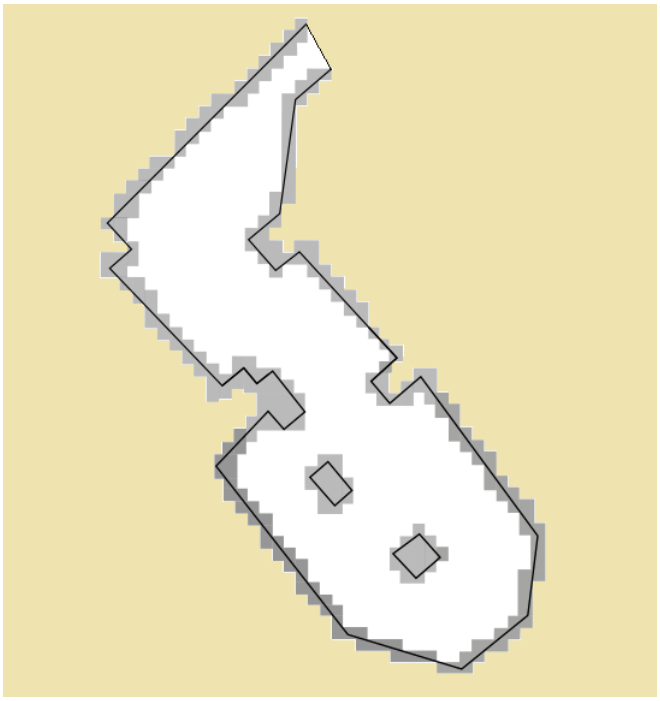


Fig. 5. The map shown after the segmenting and coloring has been done. The inside of the map is white, representing the available space, the gray squares would normally be black, but are shown as gray in order to allow for the actual walls to be visible underneath. Whereas the beige represents what would normally be the gray unknown area, outside of the field.

this, for every shape we determined on which square, in regard to the shape, it would spin (the robot would spin in order to achieve greater coverage, as it normally looks in one direction when moving). This strategy produced a large number of goals, which should ensure that the robot moves in such a way that it would find all the items on the field. However, many of these goals were very close to each other, mainly because multiple shapes were close to each other. In order to avoid this, we determined a least distance which should be between two goals. This however, would allow for a situation in which a less critical goal (a goal originating from a straight wall next to a corner) is chosen over a more critical goal (the goal originating from the corner itself which would allow for the corner to be visited). In order to avoid this, we grouped the goals into three categories of importance. The chosen goal would always have the category with highest importance of all the goals in the given radius. This process creates a list of goals the robot would have to reach. These goals use the coordinate frame of the map, and need to be transformed. In order to do this we used tf[10]. We transform the message into a $MoveBaseGoal$, and use a $MoveBaseClient$ to send it. We gain access to the map by subscribing to the ”map” topic. We use the matrix that is published to create a new map that gives us the representation of the real field as a matrix that could be analyzed as a grid.

[9]https://cloud.google.com/speech/ (last visited: 17.04.2018)

[10]http://wiki.ros.org/tf (last visited: 17.04.2018)

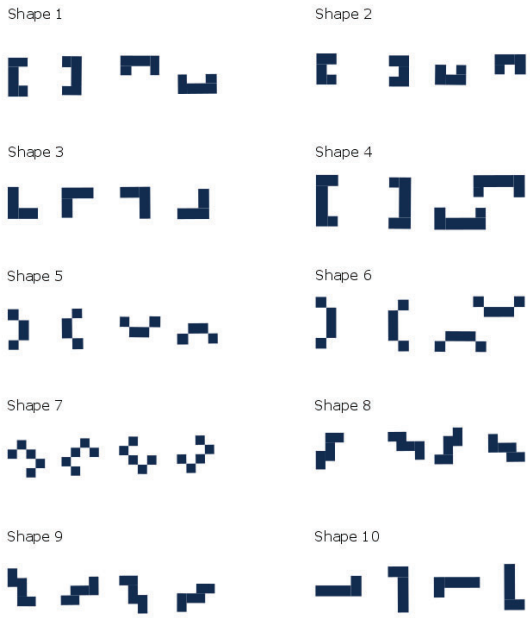


Fig. 6. The shapes that are important for the exploration strategy in order to be able to detect all objects in the field.

N. Interrogation

The interrogation begins by finding a male person on the field. Once he has been found, the robot will approach him, and ask him, by listing the women one by one, whether they are lying, until one that is telling the truth has been found. Once this is completed, the robot goes to the woman and asks her who knows where the magical ring should be taken, and which the magical ring is. Because there is a limit on the number of questions that can be asked at once, the robot will go to another person on the field once the limit is reached, say "Hello [person's name]", and come back to the current position if more questions need to be asked. Once we have found which the people are who know which the magical ring and the magical tower are we have 3 cases:

- both are female
- one is female, and one is male
- both are male

In the first case, we do not know whether any of the women are lying. In order to determine this, we go to a man on the field and ask whether each of them is lying. After this we begin the routine for finding out the ring and tower by either asking directly (if they are not lying), or by asking yes and no questions (if they are). In the second case, we first visit the man, determine which ring/tower is magical, by asking yes and no questions, and then ask him if the woman is lying. We can then go to her and determine the other magical item. In the last case, we just ask the men yes and no questions, until we have determined which the magical items are. Once we have done that, we can just go to the magical tower, as we have already picked up the rings at the beginning.

III. Performance Analysis

The results of the solution are as follows. The approach to the detected rings was not successful, and as a result not all rings were picked up. However, the ring and cylinder detection was implemented well and worked correctly. All the faces were detected and recognized correctly, and the logic of the solution worked well and found the correct ring and cylinder. Altogether, the solution performed well to the task, with the minor exception of not always picking up all rings, and thus sometimes failing to pick up the needed ring.

IV. Conclusion

In conclusion, the TurtleBot as a whole, alongside OpenCV and ROS allows for complicated tasks to be executed. It allows a great control over image processing and analysis, as well as navigation. Our solution utilized the hardware it provided along with the open-source software as best we knew how and produced sustainable results. Despite the fact that the environment we used was highly predictable and controlled, we can see many everyday situations in which the TurtleBot, or a robot very much like it, and a solution not very much unlike ours could be used in a real world scenario. All in all, the TurtleBot equipped with open-source software is a cheap solution for creating not only simple, but complex tasks as well, concerning image processing, navigation, and other spheres of artificial intelligence.

References

[1] M. Quigley, K. Conley, B. Gerkey, J. Faust, T. Foote, J. Leibs, R. Wheeler, and A. Y. Ng, "Ros: an open-source robot operating system," in *ICRA workshop on open source software*, vol. 3, no. 3.2. Kobe, Japan, 2009, p. 5.

[2] S. Boucher, "Obstacle detection and avoidance using turtlebot platform and xbox kinect," *Department of Computer Science, Rochester Institute of Technology*, vol. 56, 2012.

[3] M. Jalobeanu, G. Shirakyan, G. Parent, H. Kikkeri, B. Peasley, and A. Feniello, "Reliable kinect-based navigation in large indoor environments," in *Robotics and Automation (ICRA), 2015 IEEE International Conference on*. IEEE, 2015, pp. 495–502.

[4] I. Culjak, D. Abram, T. Pribanic, H. Dzapo, and M. Cifrek, "A brief introduction to opencv," in *MIPRO, 2012 proceedings of the 35th international convention*. IEEE, 2012, pp. 1725–1730.

[5] R. B. Rusu and S. Cousins, "3d is here: Point cloud library (pcl)," in *Robotics and automation (ICRA), 2011 IEEE International Conference on*. IEEE, 2011, pp. 1–4.

Real Time Remote Monitoring of Vital Parameters in Emergency Situations

Ivana Kozolovska, Bojana Koteska, Monika Simjanoska and Ana Madevska Bogdanova
Ss. Cyril and Methodius University
Faculty of Computer Science and Engineering
Rugjer Boskovikj 16, 1000 Skopje, Macedonia
Email: kozolovska.ivana@students.finki.ukim.mk
{bojana.koteska, monika.simjanoska, ana.madevska.bogdanova}@finki.ukim.mk

***Abstract*—Advances in telemedicine have resulted in the creation of medical systems that can wirelessly monitor the vital parameters of patients. These medical systems have significant role in saving people's life and reducing the death-rate, especially in cases of massive disasters, where there are large number of patients, limited resources, and insufficient information about their health state. This paper presents software solution for remote monitoring of vital parameters in real time, by using multiple reliable biosensors. The solution provides monitoring of several vital parameters: ECG, heart rate, respiratory rate, blood pressure and oxygen saturation. Collected data from the biosensors are stored locally on a mobile device and on the remote server which can be used for real-time monitoring of the patient state in the hospital. The solution is tested in the General Hospital in Celje, Slovenia.**

***Index Terms*—vital parameters, emergency, biosensors, telemedicine**

I. Introduction

Emergency situations pose immediate risk to people's health, life or environment. Regardless of the scope and the scale of the emergency scenario, the reaction of the medics are in the beginning the same. For example, the reaction of the medic in a case of an incident where a man was injured in a car accident or in a case when hundreds of people are injured in a train accident is basically the same. The first situation can be considered as a **small scale**, and the second as a **large scale** emergency scenario [1].

In the era of modern emergency medicine, a lot of wearable biosensors are available on the market. These sensors measure one or more vital parameters and some of them are able to transmit the vital data to nearby portable devices via Bluetooth or Wi-Fi. These biosensors are usually light weighted, long lasting and use low power protocols for data transmission. The low-cost biosensors have no or low memory capacity and their streaming frequency is usually from 100 to 1000 Hz. Wearable biosensors are very useful in emergency situations since they can provide crucial information about the patient's vital medical state and priority queue for further patient processing.

In the case of a large scale emergency scenario, the fast reaction of the first responders is crucial to save people lives. First responders are responsible for attaching the wearable biosensors to injured people. The data emitting starts immediately after the biosensor is attached to the patient. When hundreds of people are injured, the data from many injured patients can be collected on one first responder's portable device, for ex. smartphone. If Internet connection is available, selected data can be transferred to remote hospital server. Once they are transferred, the data become available to the authorized medical staff in the hospital responsible for further patient treatment.

Similarly, in small scale emergency scenario, a wearable biosensor is attached to the patient. For example, if car accident happens, during the transport to the nearby hospital, the medic in the reanimobile and also the medic from hospital can monitor the vital parameters for one or several injured persons.

The popularity of the biosensors and importance of enabling the noninvasive diagnosis of vital functions of the human body in emergency situations is presented in [2], where authors review the latest developments in body-worn wireless health-monitoring systems and their current challenges and limitations and to discuss future trends for such worn devices for these applications. The features and application of wearable biosensors in medical care are also explained in [3]. An ubiquitous emergency medical service system (UEMS) that consists of a ubiquitous tele-diagnosis interface providing ubiquitous accessibility of patients biosignals in remote areas where the ambulance cannot arrive directly is elaborated in [4].

In this paper, we propose a wireless system that enables real time remote monitoring of vital parameters in emergency situations. The goal of the proposed system is to provide continuous vital data transfer during the patient transport in the reanimobile. Authorized medics in the hospital can monitor the vital parameters of the patient in real time before the patient is transferred to the hospital which is useful for the further patient treatment. The system integrates three different biosensors to satisfy the requirements for measuring the needed vital parameters when performing first traige (heart rate, respiratory rate, blood pressure and oxygen saturation) [5]. It also provides options for checking the dynamics of these four parameters (history), marking the patient injured body parts, calculating the Glasgow Comma Scale value and storing data about the given medications during the transport. The system is tested

in the emergency department of the General Hospital in Celje, Slovenia. In order not to violate the patient's ethic rights, prior consent was required for testing the system. The reliability is confirmed with comparison of the data measured with the sensors and the data measured with the standard operating machines in the hospital.

II. Technical Solution

In the provided solution three different sensors are included in order to achieve real time remote monitoring of the needed vital parameters:

- **Zephyr BioHarness 3.0 sensor** [6] (shown in Fig. 1) - collects different vital parameters: ECG (electrical activity all over the heart), Heart Rate (number of heartbeats per minute), Respiratory Rate (number of breaths per minute), Temperature (skin temperature), Posture (body position), Activity Level (acceleration), Subject Status as well as the battery level of the device. The data are streamed at a frequency of 250 Hz.

Fig. 1. Zephyr BioHarness 3.0 sensor

- **MyTech Wrist Cuff Blood Pressure Monitor sensor** [7] (shown in Fig. 2)- measures the systolic and diastolic blood pressure (in mmHg) and pulse. The data are sent using the Bluetooth protocol.

Fig. 2. MyTech Wrist Cuff Blood Pressure Monitor sensor

- **Nonin Onyx ®9560 Saturated Blood Oxygen** [8] (shown in Fig. 3) - measures the oxygen saturation (SpO2) in range from 0 % to 100 % and pulse in range 18 to 321 beats per minute (BPM). The operating frequency is from 2.4 to 2.4835 GHz and it uses Bluetooth v2.

The data from these three sensors are transferred to the mobile device of the paramedic in the reanimobile. Received data are stored locally on the device as .csv files. If Internet connection is available, the data is sent to the remote SQL database hosted on a Windows server machine by using the android ksoap library and SOAP web services developed in C# which provide the communication between the android device and the SQL server. This is crucial action for remote monitoring of the vital parameters as the medical team in the hospital can have insight in the patients health state during his transport. As to be completed the presented scenario, a mobile device or a tablet is needed to be included on the side of the medical team in the hospital. The communication between the sensors and tablet and the transfer of the measured vital data via Bluetooth and via SOAP services is shown in Fig. 4.

Fig. 3. Nonin Onyx ®9560 Saturated Blood Oxygen sensor

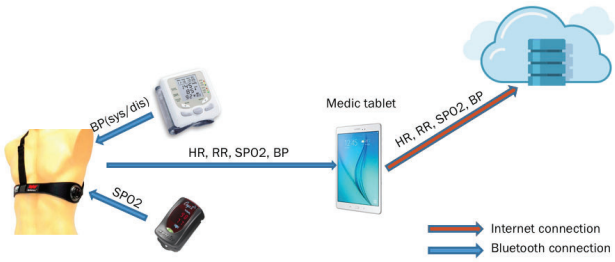


Fig. 4. Transfer of the measured vital data via Bluetooth and via SOAP services.

III. Software Solution

The Android application is intended to be used by both, the medical team in the ambulance vehicle and the medical team in the hospital. The user interfaces differs depending on the given role. Medical team interfaces in the ambulance vehicle possess more functionality unlike interfaces intended for the team in the hospital.

The solution is developed under Android platform and supports Android devices with operating system Android 4.4 (and higher). The data gathering from the Bioharness sensor is performed by using the Bioharness 3 SDK for Android platform.

Our solution enables simultaneous monitoring of four parameters: heart rate, respiratory rate, blood pressure and oxygen saturation. Fig. 5 presents a screen of the proposed solution. If the connection with the Zephyr BioHarness 3 is successful, the heart rate and respiratory rate will appear in the next screen. The MyTech blood pressure sensor must be put on the wrist on the patient. In order to get data for MyTech Blood pressure sensor the medical team in the reanimobile should manually turn on the sensor, measure blood pressure of the patient and press the Send button on the sensor. After this, the medic have to put the Nonin sensor on patients finger. The data from the Nonin Oxygen saturation sensor are streamed continuously. During the transport, the obtained measurements for these four parameters are transferred to the remote server in

Fig. 5. Real time monitoring of the vital parameters.

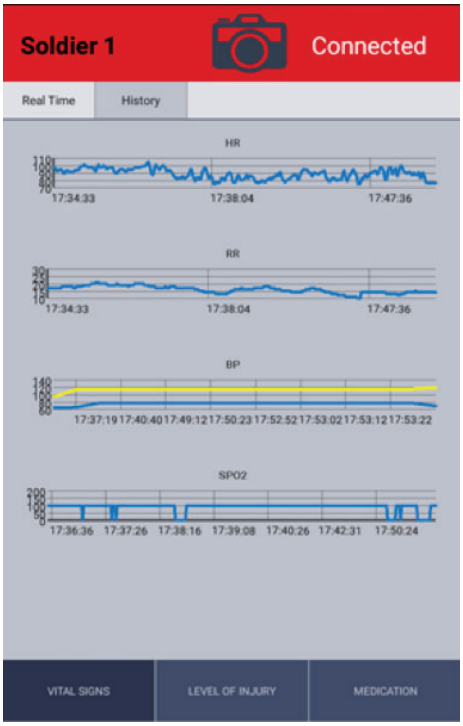


Fig. 6. History of the vital parameters.

hospital and the medical staff can monitor the vital parameters of the patient.

With tapping the history tab the user can see graphics of the biodata history for all four parameters. These graphs are also scrollable and zoomable which provides easy detection of potential abnormalities. (Fig. 6).

Other functionality, like capturing images of the wound, calculating the level of injury, marking the injured body parts, updating the list of the given medications, are also provided in order to capture a more realistic picture of the patient's health.

IV. Conclusion

In this paper we have presented a software system for real time vital parameters monitoring by using three commercially available biosensors. The application provides the ability to monitor the patient's heart rate, respiratory rate, blood pressure and oxygen saturation. In a case of an emergency scenario, during the transport in the reanimobile, the vital data can be sent to the remote hospital server. Another advantage is that the solution is wireless. This is very beneficial for the doctors in terms of the space available in the reanimobiles . There is also an opportunity to check the history (dynamics of the vital parameters) and also to zoom and scroll the signal for making deeper visual analysis by the doctors.

The database records created from the patitent's data are useful for further statistical analysis. The application is developed according to the doctors demands in the General Hospital in Celje, Slovenia where it is tested and confirmed to be reliable.

Acknowledgment

This research is supported by SIARS, NATO multi-year project NATO.EAP.SFPP 984753.

References

[1] M. Voss and K. Wagner, "Learning from (small) disasters," Natural hazards, vol. 55, no. 3, pp. 657–669, 2010.

[2] P. J. Soh, G. A. Vandenbosch, M. Mercuri, and D. M.-P. Schreurs, "Wearable wireless health monitoring: Current developments, challenges, and future trends," IEEE Microwave Magazine, vol. 16, no. 4, pp. 55–70, 2015.

[3] S. Ajami and F. Teimouri, "Features and application of wearable biosensors in medical care," Journal of research in medical sciences: the official journal of Isfahan University of Medical Sciences, vol. 20, no. 12, p. 1208, 2015.

[4] T.-H. Tan, M. Gochoo, Y.-F. Chen, J.-J. Hu, J. Y. Chiang, C.-S. Chang, M.-H. Lee, Y.-N. Hsu, and J.-C. Hsu, "Ubiquitous emergency medical service system based on wireless biosensors, traffic information, and wireless communication technologies: development and evaluation," Sensors, vol. 17, no. 1, p. 202, 2017.

[5] K. Mochizuki, R. Shintani, K. Mori, T. Sato, O. Sakaguchi, K. Takeshige, K. Nitta, and H. Imamura, "Importance of respiratory rate for the prediction of clinical deterioration after emergency department discharge: a single-center, case–control study," Acute medicine & surgery, vol. 4, no. 2, pp. 172–178, 2017.

[6] Zephyr Technology, "Zephyr BioHarness 3.0 User Manual," accessed: 2018-03-16. [Online]. Available: https://www.zephyranywhere.com/media/download/bioharness3-user-manual.pdf

[7] MyTech, "MyTech Wrist Cuff Blood Pressure Monitor sensor," accessed: 2018-03-16. [Online]. Available: https://usermanual.wiki/Mytech-Technology/HPL-108-1346666.pdf

[8] Nonin Medical, Inc., "Nonin Onyx ®II Model 9560," accessed: 2018-03-16. [Online]. Available: http://www.nonin.com/documents/6470_ENG.pdf

AQI Measuring Station With Alexa Integration

Dimitri Dojcinovski, Andrej Ilievski, Vesna Kirandziska, Nevena Ackovska - FCSE Skopje

***Abstract*—Air pollution in recent years is alarmingly part of our everyday. This paper presents our attempt at making a state of the art indoor and outdoor air quality measurement station using cloud storage and integrated voice assistant. We have used a Raspberry Pi 3, NodeMCU microcontroller and three different types of sensors.**

***Index Terms*—Air pollution, dust particles, air quality station, particulate matter, alexa skill**

I. Introduction

The existing fact of PM particles in the air, are a serious threat to our health. The newest researches came to a conclusion that the smallest PM particles (PM1), the particles with a diameter of $<1\,\mu m$, are the most dangerous of them all. Our human body has no protection against these very small particles. They enter our bodies through the respiratory system, we inhale them, and a significant part go deep into our lungs and continue out the blood stream. At worst, PM1 contribute to deadly diseases such as hearth attacks and lung cancer. These very small particles can reach the lungs and pass through the cell membranes of the alveoli (tiny sacs in our lungs where oxygen and carbon dioxide are exchanged) and continue out into the blood stream. PM10 ($<10\,\mu m$) and PM2.5($<2.5\,\mu m$), at worst, cause decreased lung function, which is still a life hazard. Besides PM, the concentration of CO_2 in our living space is harmful. So, with that in mind, to function properly, our body needs clean air, we should clean air in our living space. But, nowadays clean air is getting it if your lucky, and ventilating is a risky move, if the air is not cleaned, there is a risk that indoor air will contain a very large quantity of harmful particulates which will find their way into peoples respiratory tracts and circulation systems. In that case, it is highly recommended using an air purifier. For now, thats the permanent best solution to maintain healthy lung function and lower the risk of air pollution diseases.

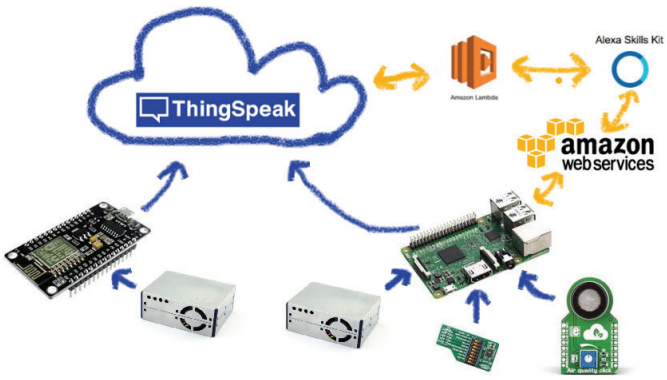


Fig. 1. The workflow of the whole station.

On Fig. 1 you can see the workflow of the whole project. The station is composed of: Raspberry pi 3, one microSD card, microphone, Bluetooth speaker, two PM sensors, wi-fi module and a temperature and humidity sensor. The sensors measure the desired particles, then upload the data on the cloud, and the Alexa assistant presents the data to us when we ask her.

II. Related Work

There have been similar projects lately, the first one by the macedonian IT company Netcetera [2], who spread a dozen similar devices to our outdoor station across the city of Skopje, and the Macedonian Telekom [1], who only announced putting devices with particle sensors on their base stations. Our project differentiates from the aforementioned projects firstly, by having an indoor station in addition to the outdoor one, secondly, by integrating the Amazon voice assistant "Alexa", thus making our product more appealing to the public because it supports the basic functionalities that Amazon already offers and continuously expands, in addition to our voice app, and finally our particle sensors detect the smallest measurable particles - PM1, which are not measured as of writing this paper by neither the governmental measuring stations, nor the previous projects. These PM1 particles by numerous recent researches [3] [4] are shown to be the deadliest and most detrimental to humans' health. In the following section we discuss the indoor station and an overview of each of the parts of the setup, after that we review the outdoor station and finally the results from running the indoor station a couple of days.

III. Indoor Setup

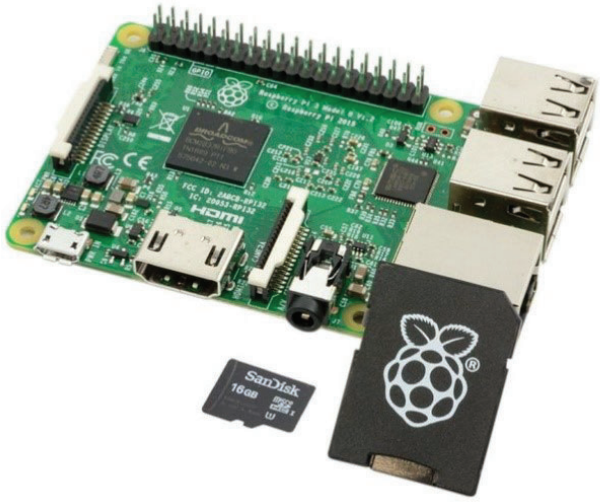


Fig. 2. Raspberry Pi3 Model-b.

A. Setting up the Raspberry and Alexa

First, we set the base of our project, we installed Raspbian OS [5] on the raspberry pi, one of many official OSs that the tiny computer can operate on. Its a Unix-like OS and comes preinstalled with python, Java, Mathematica etc. On the raspberry we set the Alexa Voice Service(AVS), which requires the user to log in with an amazon developer account [6]. For us, the alexa skills, that whole concept, was a new thing. The skills enable users to create a more personalized experience. Then with an invocation command, the alexa activates the right skill [7] to give us the right response. For and example, if we wanted to know how is the air pollution, we say: alexa ask air quality how is the pollution. The word in the blue is an invocation word to let alexa know we are speaking to her, the word in red is the custom invocation word, that we crafted for our project , that activates the custom alexa skill, and the words in green is part of the Interaction model. The interaction model is made of intents, upon which a fitting lambda function needs to be activated and give the required information. Lambda is part of the amazon web services(AWS), which lets you run code without provisioning or managing servers. It lets you program in many languages, including python, in which this project is coded on [8]. From the lambda function, the information is sent back to the AVS and finally the user can hear the information.

Fig. 3. The particulate matter sensor PMS5003.

B. Operation of the particulate matter sensor

For measuring the particulate matter in the air we used the Plantower PMS5003 sensor [9]. The sensor uses laser scattering to radiate suspending particles in the air, then collects scattering light to obtain the curve of scattering light change with time. The microprocessor calculates equivalent particle diameter and the number of particles with different diameter per unit volume [10]. In our project, the sensor measured on every 1 minute. 1 minute is the optimal time, because it takes the sensor minimum 30 seconds to heat up, and another 15 seconds to do the measuring, then do the average of the measuring and put the results on the cloud. For our project we estimated with this configuration, the sensor on average uploaded on every 53 seconds. The sensor is hooked up to the raspberry pi via the serial line (pin TXD and RXD of the Raspberry). The sensors power is 5v, and the logic is 3.3v. In the python code, we used the serial library to read the information from the data pin and save it in an array. After that, the data form the array is formatted in a frame. On every frame we have a time stamp, so we know when it measured, and its appended on a list of measurements. After 15 seconds, the array has 15 elements, and we do an average on the array and that data is sent to the cloud.

C. Carbon dioxide concentration level sensor

Fig. 4. Mikroelektronika CO_2 sensor.

The device shown in the fig. 4, represents a type of Mikroelektronika clickBoard, with an oscillator and a sensor. This board is designed to integrate easily with any Microcontroller that supports a mikroBus connection, but since we didnt have an additional Raspberry Pi Click shield, it was necessary first, to study the data sheet of the entire device and separately, the sensor itself - MQ-135. Having no electronics background, we needed to learn that the oscillator controls the sensitivity of the sensor. The sensor works based on the conductivity of the sensitive material made out of Tin dioxide (SnO2), which has a different resistance depending on the quality of the air. For more precise operation of the sensor, there is additionally an electrode that heats the air. The electrode and the sensor altogether should be switched on for at least 24 hours before making correct measurements. The conductive material is sensitive to NH3, NOx, alcohol, gasoline, smoke and CO_2. It is 6 times cheaper than a specialized CO_2 sensor, because it wasnt made as a CO_2 sensor, but our case for being a decent sensor is that the other gases are much less prevalent in the air.

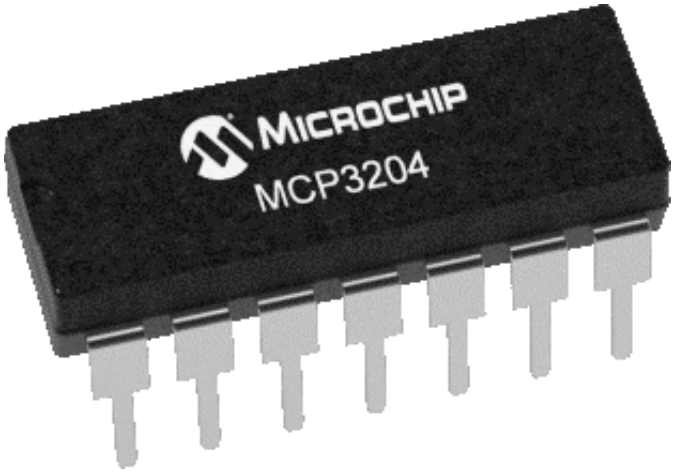


Fig. 5. Microchip ADC MCP3204.

D. Analog to Digital converter

With the raspberry lacking any analog pins, the board sends an analog signal from 0-1023 to our Analog to Digital Converter (MCP 3204). This converter has a 12-bit resolution, communicates with the SPI serial interface and operates at a voltage of 2.7 to 5 V. There are 4 analog input pins and one output pin, which displays the corresponding charges for each input pin depending on the voltage that it receives. We found a library [11] to read the values of the converter that we communicate through the SPI protocol. Using the values from the library, we convert them back to analog, so we can use another very detailed library [12] we found, written in C/C++ for Arduino, but we translated it to python to work on our Raspberry Pi 3 . Analog values are processed according to the sensor documentation and as a product we obtain a measure of the CO_2 level in the air, measured in ppm - parts per million. But before we get to it, we had to determine the reference resistance of the sensor in atmospheric air where the CO_2 levels are known. This calibration is needed because each sensor is specific. Additionally, we need to include data from the temperature and humidity sensor, because the resistance of the Tin dioxide, as noted in the documentation, is dependent on these factors. Finally, when we make the calibration, we get the desired values of the CO_2 level in the air. The function of this sensor is to read the amount of CO_2 in the air, as well as the general air quality. It is instantly sensitive to an increased amount of alcohol, gasoline, gas, and other harmful gases in the air.

E. Temperature and humidity sensor

The SHT1X microelectronics boar was far simpler to integrate. It uses an I^2C serial interface, where one pin is used to transfer data and the other one to control the clock. We used the PIC32-ARM setup so that the data is transmitted via pin P3, while the clock pin is P2. The board itself has a 14-bit resolution ADC, which means that we receive digital signals that can be used directly by the Raspberry, thus we do not need

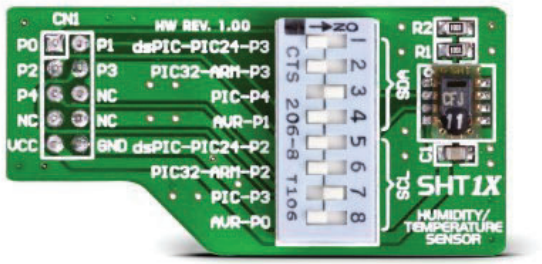


Fig. 6. Mikroelectronika SHT1X Board.

to send them through our ADC and significantly reduces our work. We used a library [13] to obtain the values we needed to re-adjust the values of the CO_2 sensor, which is dependent on temperature and humidity.

IV. Outdoor setup

NodeMCU is the name for the platform for IoT that covers the firmware running on NODEMCU v.3 devices (ESP8226 WI-FI SOC) and the devices itself. Our task was to first install that firmware on the device so that we can program it. For that we used the Arduino IDE. The connection with the sensor was made as shown on the picture below, but for that we realized that the plug of the PM sensor was not compatible with the device, so first we had to split the wires of the plug and connect each separate wire with on of the female jumper wires we had. For the code we used a library that was written to work on the Arduino, but it is also supported by our NodeMCU. The Arduino code is a C/C++ code structured in such a way that one function is run only at the start called setup(), and the other function called loop() runs repeatedly until the device is powered off or reset. In the setup function we connect the device to the WI-FI. In the loop function we power on the sensor, wait until the sensor preheats, read the data, upload it on the Thingspeak server via HTTP request, and switch off the sensor for 30 seconds. There was an option to put the NodeMCU device into deep sleep in this period, so that we could save additional energy consumption, but as noted in the documentation of the code the function didnt work properly and the device would simply stop working.

V. Thingspeak Cloud platform

ThingSpeak is a cloud platform specifically designed for IoT devices, where data is collected and then used based on what you want to do with that information. We have created 3 tables for our project. One table with data for indoor PM, temperature, humidity and CO2, and another table for outdoor PM. To upload the data we used the HTTP Request method POST. In python this is enabled with the http library, where you must connect to the server, in our case ThingSpeak uses a port 80 that requires TCP socket, which is obtained from our environment. To read the data, we used the urlib2 library that

overturns the data from ThingSpeak to JSON format. And then converting from JSON to a plain string, the Lambda custom function, processes the PM data and in the end, it outputs an AQI which is the overall evaluation of the air pollution.

VI. Results

The graphs represent an analysis of the indoor air when the station was running for a few days, and the results were as expected. The level was dropping along the day. The little bumps during the night were made because, we opened the window to exchange some air. The big increase of air particles, resulted when we tried to vacuum clean the same room the indoor station was in which resulted in increased dust particles movement in the air. This proves the accuracy and reactiveness of the sensors. The outdoor station serves the purpose of a more frequent, reliable and accurate information about the future consumer.

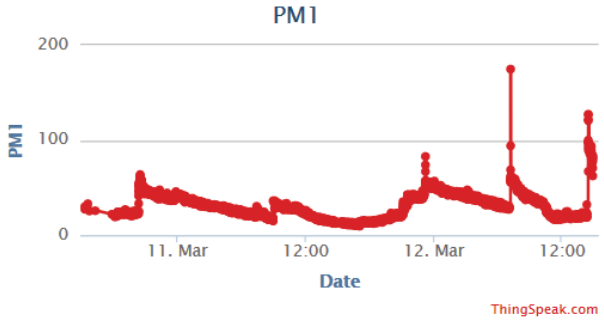


Fig. 7. PM1 size particles graph.

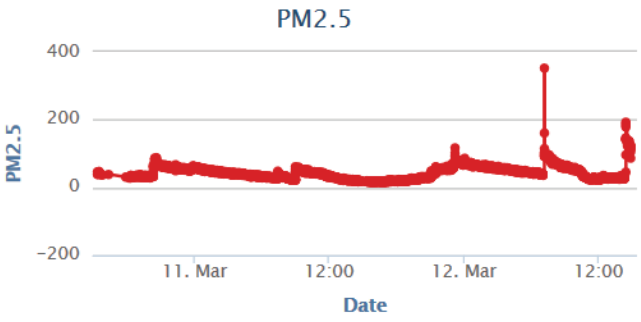


Fig. 8. PM2.5 size particles graph.

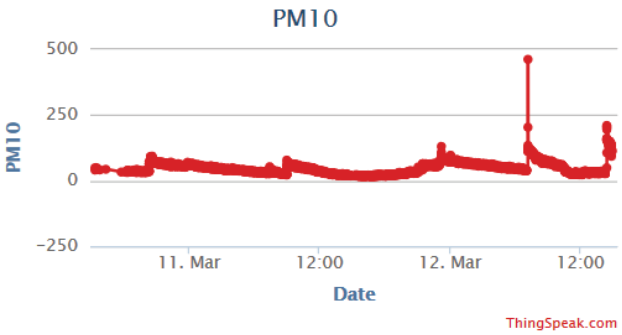


Fig. 9. PM10 size particles graph.

VII. Conclusion

All in all, we succeeded in executing our idea to be better informed about the air we breathe not only outdoors but also indoors where we spent most of our time in the winter when the outdoor pollution is high and we managed to do this while future proofing our stations, by preparing them for a voice-first future. As the technology advances, we can see a future where these sensors are integrated in the personal assistants, like the Amazon Alexa devices, which it is becoming a part of our daily life and its a must accessory of our living space. Developing a web and mobile application can further improve the access to information to all of us. By having multiple such stations across one city and by measuring the wind speed and direction and knowing the precise location and elevation of the measuring station we can help build a 3D map of a city, to showcase how the pollution is roaming around us, how outdoor pollution affects indoor pollution, and furthermore we can use machine learning to predict the pollution.

Acknowledgement

The authors would like to thank the Faculty of Computer Science and Engineering - Skopje for partially financing this work.

References

[1] telekom.mk. (2018). Makedonski Telekom. [online] Available at: https://www.telekom.mk/ns-newsarticle-svetski-kongres-na-mobilni-tehnologii-vo-barselona—makedonski-telekom-najavi-prototip-na-smart-maski-za-zastita-od-zagaduvane.nspx [Accessed 31 Mar. 2018].

[2] Skopjepulse.mk. (2018). Skopje Pulse. [online] Available at: https://www.skopjepulse.mk/ [Accessed 29 March. 2018].

[3] Zwozdziak A, Swka I, Willak-Janc E, Zwozdziak J, Kwieciska K, Baliska-Mikiewicz W. Influence of PM1 and PM2.5 on lung function parameters in healthy schoolchildrena panel study. Environmental Science and Pollution Research International. 2016;23(23):23892-23901. doi:10.1007/s11356-016-7605-1.

[4] Gongbo Chen, Shanshan Li, Yongming Zhang, Wenyi Zhang, Daowei Li, Xuemei Wei, Yong He, Michelle L Bell, Gail Williams, Guy B Marks, Bin Jalaludin, Michael J Abramson, Yuming Guo, Effects of ambient PM1 air pollution on daily emergency hospital visits in China: an epidemiological study, The Lancet Planetary Health,Volume 1, Issue 6,2017,Pages e221-e229,ISSN 2542-5196, https://doi.org/10.1016/S2542-5196(17)30100-6.

[5] Raspberrypi.org. (2018). Installing operating system images - Raspberry Pi Documentation. [online] Available at: https://www.raspberrypi.org/documentation/installation/installing-images/README.md [Accessed 7 Jan. 2018].

[6] GitHub. (2018). alexa/alexa-avs-sample-app. [online] Available at: https://github.com/alexa/alexa-avs-sample-app/wiki/Raspberry-Pi [Accessed 7 Jan. 2018].
[7] Hacker Noon. (2017). My first Alexa custom skill Hacker Noon. [online] Available at: https://hackernoon.com/my-first-alexa-custom-skill-6a198d385c84 [Accessed 7 Jan. 2018].
[8] Docs.aws.amazon.com. (2018). What Is AWS Lambda? - AWS Lambda. [online] Available at: https://docs.aws.amazon.com/lambda/latest/dg/welcome.html [Accessed 24 Jan. 2018].
[9] AQICN.org. (2018). The Plantower PMS5003 and PMS7003 Air Quality Sensor experiment. [online] aqicn.org. Available at: http://aqicn.org/sensor/pms5003-7003/ [Accessed 04 Feb. 2018].
[10] Rigacci.org. (2018). [online] Available at: https://www.rigacci.org/wiki/lib/exe/fetch.php/doc/appunti/-hardware/raspberrypi/plantower-pms5003-manual_v2-3.pdf [Accessed 05 Feb. 2018].
[11] Webiopi.trouch.com. (2018). The Raspberry Pi Internet of Things Toolkit - Now in two flavors. [online] Available at: http://webiopi.trouch.com/ [Accessed 28 Jan. 2018].
[12] GitHub. (2018). ViliusKraujutis/MQ135. [online] Available at: https://github.com/ViliusKraujutis/MQ135 [Accessed 14 Feb. 2018].
[13] GitHub. (2018). drohm/pi-sht1x. [online] Available at: https://github.com/drohm/pi-sht1x [Accessed 15 Jan. 2018].

Exploring different heuristics for WSN localization based on trilateration

Andrej Jovanov, Robert Stoimenov, Biljana Risteska Stojkoska
Faculty of Computer Science and Engineering
Saints Cyril and Methodius University,
1000 Skopje, Macedonia
{andrej.jovanov@students., robert.stoimenov@students., biljana.stojkoska@}finki.ukim.mk

***Abstract* –Localization remains a challenging problem in wireless sensor networks community. There are many algorithms for localization proposed in the literature, ranging from very basic to very complex. Among them, trilateration is one of the oldest and simplest approach, that can be used alone, or as part of more complex pipelines. In this paper we explore two different heuristics that adopt trilateration as a base technique. Our heuristics use distance information together with knowledge about the most appropriate anchors in terms of quality. We show that by improving trilateration with different heuristics, localization error can be significantly reduced, while maintaining low computational cost.**

Keyword- localization, wireless sensor network, trilateration, heuristics

I. INTRODUCTION

Localization is still attractive problem in Wireless Sensor Network (WSN) community. Although it is one of the oldest problem defined in WSN, common solutions are very rare, especially for indoor environment. With the flourish of WSN and its seamless integration with Internet of things (IoT) [1], the localization problem has gain even more interest, not only among researcher, but also among entrepreneurs [2].

Most of the algorithms for localization are based on distance measurements. Traditional ranging techniques for distance measurements are based on received signal strength indicator (RSSI), which is proved to be very unreliable method for distance estimation [3]. Therefore, most of the mathematical techniques for localization fail to maintain the same performances under the field conditions. For example, multidimensional scaling technique which is based on very exact mathematical background, although provides very small localization error in simulations, it gives high localization error when evaluated in real environment [4].

In the near future, with the technological advances, we expect new sophisticated ranging techniques that will be embedded in the wireless devices [5]. Having more accurate distance estimation, even basic localization techniques can perform small localization error, maintaining low complexity at the same time.

Therefore, in this paper we investigate very basic technique for localization based on trilateration [6]. To improve the performance of trilateration, we applied iterative approach, which is additionally refined. For refinement, we developed two different heuristic, implement them and evaluate the performances in order to choose the better one. The results from our simulations show that basic techniques for localization, refined with appropriate heuristics, can provide very acceptable localization error.

The rest of this paper is organized as follows. The mathematical and geometrical background of trilateration is given in Section II. In Section III we describe our iterative trilateration together with the two different heuristics for refinement. Simulation settings and results are presented in Section IV. Finally, the paper is concluded in Section V.

II. TRILATERATION

In this section we will provide a brief mathematical background of trilateration as a technique for localization in wireless sensor network.

Let's assume that a sensor device with unknown location is within the communication range of at least three other sensor devices with apriori known locations, also known as anchor devices. If the three anchors have coordinates O1(x1, y1), O2(x2, y2) and O3(x3, y3) respectively, the unknown device A(x, y) lays in the intersections of the three circles with centers in O_1, O_2 and O_3 (Fig.1). Using the distance equation assuming that r_i are the distances from the unknown point $A(x, y)$ to the anchor points O_1, O_2, O_3 respectively, we can obtain the coordinates of the unknown:

$$(x - x_1)^2 + (y - y_1)^2 = r_1^2$$

$$(x - x_2)^2 + (y - y_2)^2 = r_2^2$$

$$(x - x_3)^2 + (y - y_3)^2 = r_3^2 \quad (1)$$

After expanding out the squares in each equation, subtracting one from another and solving the system of two equations in two unknowns as given in (1), (2), (3), (4), (5):

$$x^2 - 2x_1x + {x_1}^2 + y^2 - 2y_1y + {y_1}^2 = r_1^2$$

$$x^2 - 2x_2x + {x_2}^2 + y^2 - 2y_2y + {y_2}^2 = r_2^2$$

$$x^2 - 2x_3x + {x_3}^2 + y^2 - 2y_3y + {y_3}^2 = r_3^2 \quad (2)$$

$$(-2x_1 + 2x_2)x + (-2y_1 + 2y_2)y = r_1^2 - r_2^2 - x_1^2 + x_2^2 - y_1^2 + y_2^2$$

$$(-2x_2 + 2x_3)x + (-2y_2 + 2y_3)y = r_2^2 - r_3^2 - x_2^2 + x_3^2 - y_2^2 + y_3^2 \quad (3)$$

To represent the constants in (3), using A, B, C, D, E and F (4), we get (5) and (6).

$$A = (-2x_1 + 2x_2)$$

$$B = (-2y_1 + 2y_2)$$

$$C = r_1^2 - r_2^2 - x_1^2 + x_2^2 - y_1^2 + y_2^2$$

$$D = (-2x_2 + 2x_3)$$

$$E = (-2y_2 + 2y_3)$$

$$F = r_2^2 - r_3^2 - x_2^2 + x_3^2 - y_2^2 + y_3^2 \quad (4)$$

$$Ax + By = C$$

$$Dx + Ey = F \quad (5)$$

$$x = \frac{CE - FB}{EA - BD}$$

$$y = \frac{CD - AF}{BD - AE} \quad (6)$$

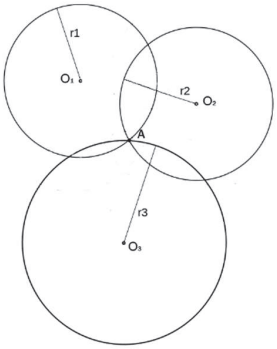


Figure 1.Geometrical interpretation of trilateration

The trilateration localization scheme is a distributed approach, since every unknown sensor device should use all available data from its surrounding to obtain its own location. The algorithm steps are described as follows:

1. Anchors advertise their location to other sensor devices from the surrounding.

2. Unknown nodes discover all anchors from the surrounding and select three closest anchors.

3. Unknown nodes apply equations (1) – (6) to calculate their coordinates.

III. ITERATIVE TRILATERATION

The main drawback of trilateration is its inability to localize all sensor devices in the network, i.e. only devices that has three anchors in its close proximity would be localized. Even if there is sufficient number of anchors, which are not evenly distributed around the environment, the trilateration will not achieve the desired results. Therefore, trilateration is usually applied as iterative approach. In each consecutive iteration, the sensor devices being previously localized become new anchors. The steps in iterative trilateration are as follows:

1. Anchors advertise their location to other sensor devices from the surrounding.

2. Unknown nodes discover all anchors from the surrounding and select three closest anchors.

3. Unknown nodes apply equations (1) – (6) to calculate their coordinates.

4. Unknown nodes become new anchors. Start step 1 until all nodes become anchors.

In this paper we apply the iterative trilateration approach. Additionally, we apply a refinement using two different heuristics (H1 and H2), based on different approaches for choosing the most appropriate three surrounding anchors that will be used to localize the unknown device. Since the unknown device will have many anchors in its surrounding, some of them will be closer, but will have smaller reliability. In order to qualitatively distinguish the anchors, they are assigned a weight, that stands for unreliability. Namely, as new anchors are created in each consecutive iteration, they are supposed to have embedded some localization error.

In the first heuristic H1, the unknown node chooses the closest anchors in order to localize itself. In the second heuristic H2, it uses the most reliable anchors from its surrounding. If we denote the weight of the anchor nodes with $W(A_i)$ respectively, then the weight of the unknown node, after becoming new anchor, will be the sum of the weights of the anchors used for its localization, increased by 1 (7).

$$W(U_n) = W(A_1) + W(A_2) + W(A_3) + 1 \quad (7)$$

The step for each heuristic are described as follows.

H1:

1. Anchors advertise their location to other sensor devices from the surrounding.

2. Unknown nodes discover all anchors from the surrounding and select the five closest anchors.

3. Choose the best three anchors eliminating the possibility for anchors collinearity.

4. Unknown nodes apply equations (1) – (6) to calculate their coordinates.

5. Became a new anchor and go to 1.

H2:

1. Anchors advertise their location to other sensor devices from the surrounding.

2. Unknown nodes discover all anchors from the surrounding and select the five anchors with the smallest weight.

3. Choose the best three anchors eliminating the possibility for anchors collinearity.

4. Unknown nodes apply equations (1) – (7) to calculate their coordinates.

5. Became a new anchor and go to 1.

Since trilateration works only for three non-collinear points, we aim to reduce the possibility to choose three colinear anchors. Therefore, for both heuristics, in Step 2 we select 5 anchors. Then, in step 3, we select the best three anchors, eliminating those anchor combinations that are collinear or near collinear.

IV Simulation Results

To investigate and to compare the performances of proposed heuristics for iterative trilateration, we ran simulations. The parameters of the simulation settings were as follows:

1. Fixed number of sensor devices (N=100) deployed randomly with uniform distribution inside a square area (100r x 100r) where r is a unit length distance.

2. Different communication range R, from 35r to 70r, with step 10r.

3. Different range error Er, modeled as uniform distribution from 5%R to 50%R with step 10%R or 5%R for different situations.

4. Different anchor fraction A, ranging from 10%N to 50%N with step 10%N.

The results for each scenario represent an average over 30 trials.

In this section, we investigate iterative trilateration accuracy in open 2D space. We address accuracy as average localization error (*ALE*), which is the average distance between the estimated positions and the real positions of all sensor devices, normalized to the devices communication range. *ALE* is computed using (8).

$$ALE = \frac{\sum distance(pos_i^{calculated} - pos_i^{real})}{(N-A)R} 100\% \qquad (8)$$

The first aim of our simulation was to discover which heuristic is more appropriate for solving localization problem in WSN. Fig 2, Fig 3 and Fig 4 show that heuristic H1 perform better and gives smaller localization error compared with heurist H2 for all simulation configurations.

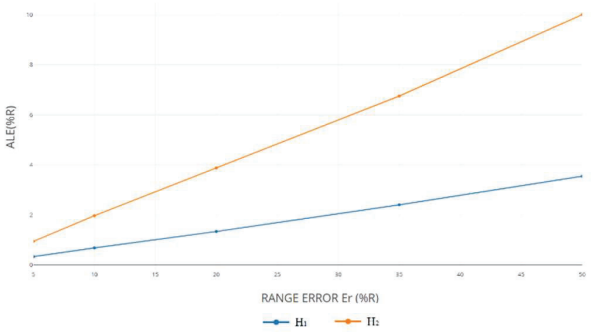


Figure 2: Evaluation of iterative trilateration, with variable range error Er, communication range R = 40r and anchor fraction A = 35.

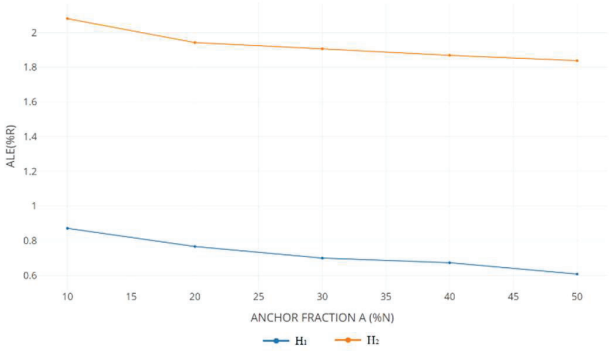


Figure 3: Evaluation of iterative trilateration, with variable anchor fraction A, communication range R = 40r and range error Er = 10%R.

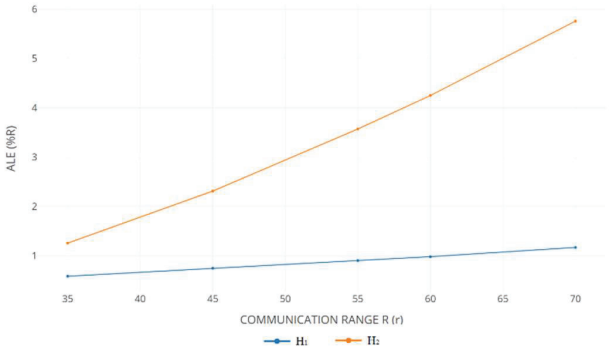


Figure 4: Evaluation of iterative trilateration, with variable communication range R, range error Er=10%R and anchor fraction A=35.

As expected, using more anchor nodes gives slightly smaller localization error. Number of anchors affects the results when the communication range is high (Fig 5). For small communication range (Fig. 6), there is no evident localization

accuracy improvement, especially for small to medium anchor fraction.

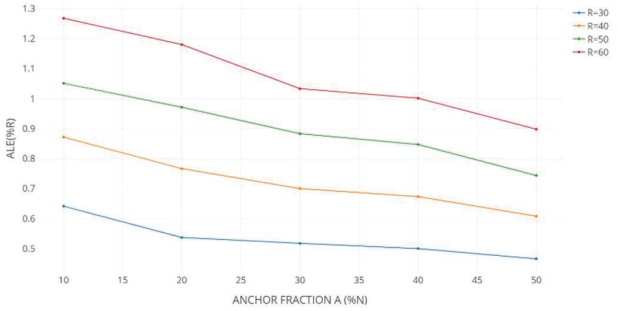


Figure 5: Evaluation of iterative trilateration, with variable anchor fraction *A*, communication range *R* and range error *Er*=10%*R*.

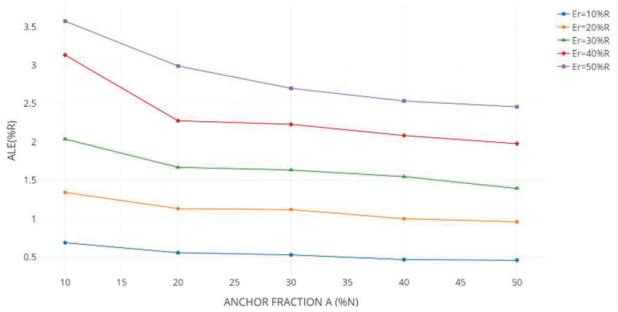


Figure 6: Evaluation of iterative trilateration, with variable anchor fraction *A*, range error *Er* and communication range *R*=30*r*.

Range error E_r has great impact on localization error. As E_r increases, H1 deteriorates (Fig 7 and Fig. 8).

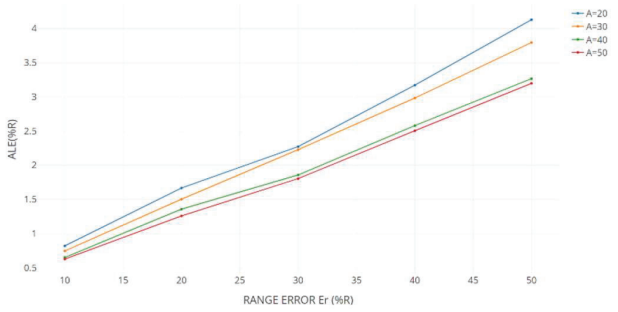


Figure 7: Evaluation of iterative trilateration, with variable range error *Er*, anchor fraction *A* and communication range *R*=40*r*.

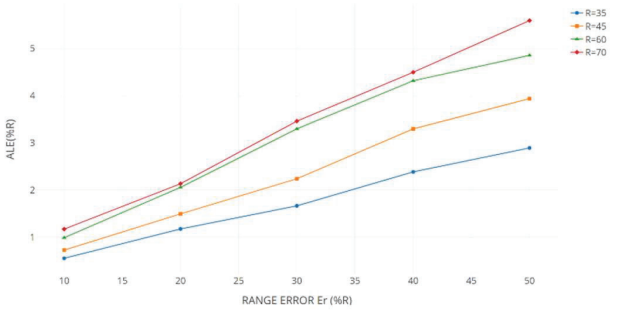


Figure 8: Evaluation of iterative trilateration, with variable range error *Er*, anchor fraction *A*= 40%N and communication range *R*.

Fig 9 shows evaluation of iterative H1 trilateration, for different communication range *R*, different range error *Er* and anchor fraction of 20 nodes. Fig 10 shows the results of H1, for different communication range R, different anchor fraction A and range error *Er*=10%*R*. As can be seen from the figures, in both cases, H1 performs smaller estimation error for smaller *R*. The reason is that *Er* increases with R, and *Er* greatly affects the localization error.

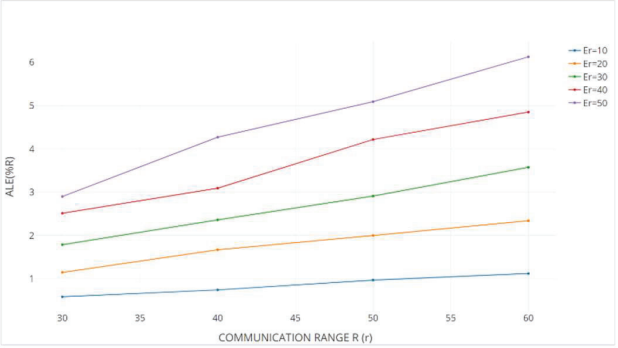


Figure 9: Relationship between *ALE* and *R* for different *Er*

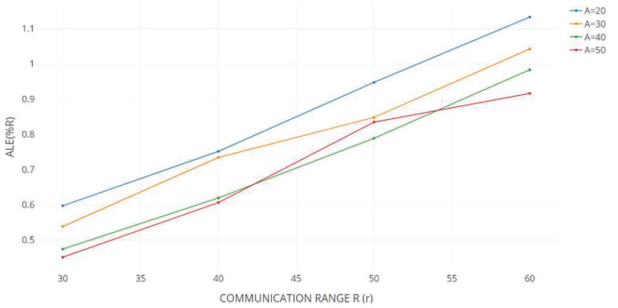


Figure 10: Relationship between *ALE* and *R* for different *A*

V. CONCLUSION

In this paper we explore two different heuristics that adopt trilateration as a base technique. Our heuristics use distance information together with knowledge about the most appropriate anchors in terms of quality. We show that by improving trilateration with different heuristics, localization error can be significantly reduced, while maintaining low computational cost.

ACKNOWLEDGEMENT

This research work was conducted as a student project for the undergraduate course "Sensor Systems" at the Faculty of Computer Science and Engineering, Saints Cyril and Methodius University, Skopje.

REFERENCES

[1] Shit, Rathin Chandra, Suraj Sharma, Deepak Puthal, and Albert Y. Zomaya. "Location of Things (LoT): A Review and Taxonomy of Sensors Localization in IoT Infrastructure." IEEE Communications Surveys & Tutorials (2018).

[2] Stojkoska, Biljana Risteska, Ivana Nizetic Kosovic, and Tomislav Jagušt. "A Survey of Indoor Localization Techniques for Smartphones." In th International Conference ICT Innovations 2016. 2016.

[3] Parameswaran, Ambili Thottam, Mohammad Iftekhar Husain, and Shambhu Upadhyaya. "Is RSSI a reliable parameter in sensor localization algorithms: An experimental study." In Field failure data analysis workshop (F2DA09), vol. 5. IEEE, 2009.

[4] Stojkoska, Biljana Risteska, Ivana Nižetić Kosović, and Tomislav Jagušt. "How much can we trust RSSI for the IoT indoor location-based services?." In Proceedings of the 25th International Conference on Software, Telecommunications and Computer Networks (SoftCOM,'17), Split, Croatia. 2017.

[5] Brena, Ramon F., Juan Pablo García-Vázquez, Carlos E. Galván-Tejada, David Muñoz-Rodriguez, Cesar Vargas-Rosales, and James Fangmeyer. "Evolution of indoor positioning technologies: A survey." Journal of Sensors 2017 (2017).

[6] Zhou, Zhong, Jun-Hong Cui, and Shengli Zhou. "Localization for large-scale underwater sensor networks," Networking 2007. Ad hoc and sensor networks, wireless networks, next generation internet, pp. 108-119, 2007.

Experiences of student's projects for the course Microprocessing systems

Andrej Jovanov, Jane Lameski, Vesna Kirandziska, Nevena Ackovska
Faculty of Computer Science and Engineering
Ss. Cyril and Methodius University, Skopje, Macedonia
{andrej.jovanov, jane.lameski}@students.finki.ukim.mk, {vesna.kirandziska, nevena.ackovska}@finki.ukim.mk

***Abstract*—Hardware projects for the computer science students, are more challenging, because our department is mostly focused on software and programming than it is on electronics, and hardware. For the course Microprocessing systems students should make projects using real hardware to achieve the highest grade. Having in mind that real projects are more interesting and motivating for the students, our Professor and Teaching Assistant, made a collaboration with a company that had a real hardware problem that had to be resolved. In this paper we will present the problem itself, our access to it, our experience from it and the results from this collaboration. We will stress out the benefits and difficulties of this kind of practical work for students. As a conclusion, at the end we will address the educational benefit of this kind of collaborations.**

Keywords—microprocessing systems, education, hardware, computer science students

I. INTRODUCTION

The earliest home security systems date back to the early 1900's. These systems were generally expensive and very hard to monitor. In the past 100 years as technology has changed, home security systems have also changed [9]. If we look at different home automation systems over time, they have always tried to provide efficient, convenient, and safe ways for home inhabitants to access their homes. Irrespective of the change in user expectations, advancement of technology, or change of time, the role of a home automation system has remained the same [8].

Nowadays, the crime rate is rapidly increasing day by day. Every day we can see in the news that someone killed his mother, a girl has been molested, a couple has been murdered, a family has been looted etc. This news has become a common scenario in our society [7]. The need for building an automation system for office or home is increasing day-by-day with numerous benefits. Main solutions for security in a distributed system are encryption, authentication and authorization. By encrypting a data using encryption a non-allowed user can't understand what the original meaning of that data is. Encryption provides confidentiality. To verify a claimed identity of a user or a host authentication is used. For controlling access in a service for an outsider or insider user authorization is used [10].

As an assignment for the subject, me and my colleague got a chance to work on a real commercial hardware problem for a company. The project was about authentication and authorization for every employee. Because the resources for task were limited (we used hardware components available in our faculty laboratory for intelligent systems), we had to come up with a simple and practical solution. Nowadays with the rapid expansion of technology, for this project there are many possible solutions, such as Face ID that Apple uses on their iPhone X [11], the iris scan [12], which method Samsung implemented on their new generation of Galaxies, but to implement these solutions, we would have needed extremely sophisticated equipment.

Our first task was to investigate the possible solution given the microcontrollers, sensor, actuators and others equipment available for the students' projects. Considering this, one solution we thought of was Bluetooth authorization weather using a smartphone, smartwatch or any gadget that has Bluetooth connection [1]. There were other possible solutions such as the RFID cards [2] which are widely used in many facilities, people also use passcode door lock systems [3] for their home entrance and considered as one of the most secure is the fingerprint-based (Biometric) solution [4], because the fingerprint of one person never matches the other [6].

During our first meeting with the clients, we discussed all the ideas previously mentioned. At first, we thought that the Bluetooth would be good solution for the problem because of few reasons. This authorization is very easy to use. Many people nowadays use smart gadgets, such as phones, watches, tablets, laptops etc. Usually most of the people use at least one of these devices daily, so it would not be a problem if one touches the display few times so that one can enter a building. This solution is much better then carrying keys in ones' pockets or forgetting the passcode [5]. The problem in this method is that these previously mentioned gadgets run on battery, not everyone owns one and not everyone feels comfortable taking them to work. To provide the best security to the lockers and to make the work easier, the solution is taking help of two different technologies viz. EMBEDDED SYSTEMS and BIOMETRICS. We, together with the clients, decided to make an implementation of a fingerprint-based solution for authorization.

In the next section we will elaborate on the hardware components for our project and next the two versions of our solution will be presented. In Section IV the results are presented and finally a conclusion is given.

This work was partially financed by the Faculty of Computer Science and Engineering at the "Ss. Cyril and Methodius" University, Skopje.

II. HARDWARE COMPONENTS

Given that fingerprint matching requires significant computation, we needed a module with a dedicated function for real-time computing. An open-source electronics platform based on easy-to-use hardware and software was the selling factor for the students going with Arduino. To be more specific Arduino Board – Arduino UNO and software Arduino IDE to program the board.

The board feature serial communications interfaces, sets of digital and analog I/O pins and USB. For programming the microcontrollers with the provided integrated development 1environment (IDE) we used C++ which is one of the supported programming languages. Fingerprint module that can be operated easily by our Arduino development board was Adafruit Fingerprint Sensor Module. The module consists of optical fingerprint sensor, high-speed DSP processor, FLASH chip for storing the fingerprints and high-performance fingerprint matching algorithm. The fingerprint module was provided together with already written Arduino library in C++ for the basic functionalities. The module performs series of functions like fingerprint enrollment, image processing, fingerprint matching, searching and template storage. When enrolling, user needs to enter the finger two times. The system will process the two-time finger images, generate a template of the finger based on processing results and store the template. When matching, user enters the finger through optical sensor and system will generate a template of the finger and compare it with templates of the finger library. Image acquiring time and the average searching time through the templates is less than 1 second. This module can store 120 templates given that the generated template from the acquired image is 512 bytes [14].

Accuracy of the module is presented with help of two important measurements, false acceptance rate and false recognition rate. The false acceptance rate, or FAR, is the measure of the likelihood that the biometric security system will incorrectly accept an access attempt by an unauthorized user. The false recognition rate, or FRR, is the measure of the likelihood that the biometric security system will incorrectly reject an access attempt by an authorized user [15]. Adafruit Fingerprint Sensor Module reports great results given that FAR is less than 0.001% and FRR is less than 0.1% [14].

Our Arduino UNO board lacks hardware serial and because of that we connected the fingerprint module trough #2 and #3 pins using software serial.

After successfully connecting we used the libraries for testing the module and start making changes. We have made two versions of the application trough that process and next we will discuss them in more details.

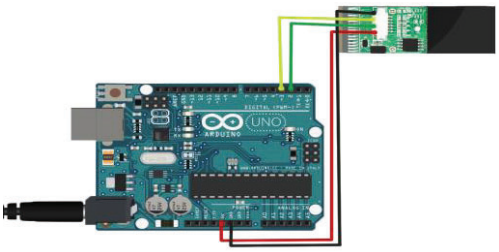


Fig. 1. Sketch of version one.

Fig. 2. Version one of the project.

III. THE PROJECT VERSIONS

A. *Version One*

The first version of our program was designed to be used only with the fingerprint module itself. The admin was added and only that person had the ability to change modes such as going to admin mode or going to searching and matching mode. It was simple program and our first version of the application in which we tested the fingerprint module.

For example, if we wanted to add a new employee, first the admin had to place the finger to verify that the admin is making the change, then in the next 5 seconds he had to place the same finger again to change the mode from ‘searching and matching’ to ‘adding a new fingerprint’. Placing the finger two times in 5 seconds would mean deleting the next fingerprint that will be read and so on.

We scheduled a second meeting with the client and presented the application we have made beside the lack of full functionality. Trough out the presentation we saw that this solution is difficult to use, and together with the clients and the teachers, talked about the other possibilities on how the user-admin can control the application. The decisions that were made on the meeting, on how to improve the application, resulted with an upgraded version.

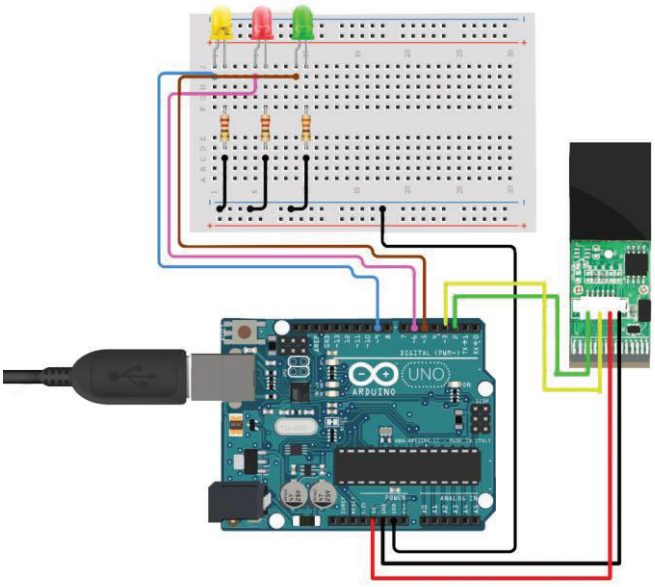


Fig. 3. Sketch of version two.

B. *Version Two*

We decided to start from scratch and enable the Arduino program to communicate with our custom user interface trough serial communication that will be installed only on the administrator's computer and directly connected to the system itself. This made the process of adding, deleting and changing modes a lot easier. We will discuss these functionalities in more detail later in this section.

In this version we improved the look of our prototype by adding indicators that explain the work of the system. We added light-emitting diodes in three different colors near the fingerprint sensor indicating if the user is successfully matched, denied or the system is processing the fingerprint.

Fig. 4. Version two of the project.

The "FingerprintSensor" program was created in Visual Studio using C# programming language. This program is meant to be used by the admin who will operate with the authentication and authorization solution that we have provided.

The program offers three different modes for the fingerprint module. By changing the mode to 'Add Employee' the administrator will enable the fingerprint sensor to scan for new fingerprint and assign unique id to that employee. The predefined mode is 'Scan' and in that mode the fingerprint module is scanning and searching for a match. The last mode is 'Delete' for deleting the stored fingerprint templates that match the scanned one. All the results from the fingerprint module are stored at the desired location. Adding the current time when the result is stored, gives the ability to monitor the employees and to know exactly when they entered or left the building.

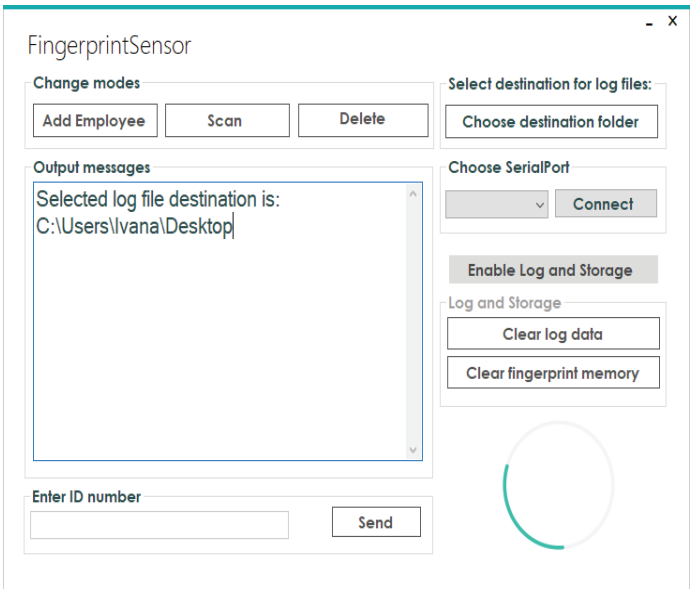


Fig. 5. "FingerprintSensor" program.

IV. RESULTS

During the final meeting, when we presented the final version of the project, some of the employees did not feel keen on testing, because of possible privacy violation. The comments that we heard from them were, "Is this safe?", "Is it possible for my fingerprint to be violated?", "Is it possible to steal the fingerprint database?" etc. We answered all the questions and after that some of the employees tested the project. The solution worked well, we had multiple employees get their fingerprints read using "Add Employee" mode and then their ID displayed on the "Scan" mode. At the end of the presentation, we deleted all the fingerprints, so that the employees will not need to worry.

The biggest problem of all, is the law for data safety, which leaves the question whether this project will be implemented at all. The only way to steal the fingerprints is to somehow approach the sensors memory, which is only possible if one steals the sensor, which is not an easy task and thus we think that the fingerprints are secure, inside the hardware. Even though the project is easy to use, to get a permission for its usage is quite difficult, because it must go

through complicated and extensive revision from the government.

At the end of the day, we think that the project is successful, and we hope that it will be used soon. As for the future, we strongly believe that the unfinished task of synchronizing our project with the already existing system in the company, will eventually happen.

Fig. 6. The final version of the project.

V. CONCLUSION

The greatest challenge that we faced during this project was getting to know how the hardware work, and how to connect them with microcontroller, in our case Arduino and the Version one of the project. We spent good amount of time trying to think of the solution that was asked from us, and then after implementing it, fixing the bugs that we encountered with was more than a challenge. After the second meeting we got relieved, because everything became much easier when we used software to control the work of the fingerprint sensor.

The approach of practical use of the knowledge in the field of Microprocessors, showed it self as more efficient and motivating for us. Most of the students do not have a chance to interact with hardware nor to work on a real "commercial" project during their studies. This approach on learning and working on real projects at the same time, gave students a chance to collaborate with clients and prepare themselves for the upcoming challenges waiting for them after graduation.

ACKNOWLEDGMENT

We formally thank the company doXteam [13] for allowing this project, their collaboration and great support.

REFERENCES

[1] Bluetooth Door Lock (Arduino), 2017. [Online]. Available: http://www.instructables.com/id/Bluetooth-Door-Lock-Arduino/

[2] A. Mukherjee, Security Access Using RFID Reader, 2016, [Online]. Available: https://create.arduino.cc/procjecthub/Aritro/security-access-using-rfid-reader-f7c746.

[3] Password Based Door Lock , 2016. [Online]. Available: https://create.arduino.cc/projecthub/rishabh411/password-based-door-lock-3df2e0

[4] Arduino Fingerprint Lock, 2013. [Online]. Available: http://www.instructables.com/id/Arduino-Fingerprint-Lock/

[5] Hae-Duck Joshua Jeon. Jiyoung Lim, Wooseok Hyun, Woojin Lee, "A Remote Lock System Using Bluetooth Communication", Innovative Mobile and Internet Services in Ubiquitous Computing, 2014, p. 441-446.

[6] A. Aditya Shankar, P.R.K. Sastry, A. L.Vishnu Ram, A.Vamsidhar, "Finger Print Based Door Locking System", International Journal Of Engineering And Computer Science, vol 4, num 3, 2015, pp. 10810-10814.

[7] Why is home security is important? , 2016. [Online]. Available: https://www.quora.com/Why-is-home-security-is-important

[8] Cyril Jose, A. and Malekian, R., "Smart Home Automation Security: A Literature Review". Smart Computing Review, 5, 2015, pp. 269-285.

[9] Ibrahim Geha, Kfoury Elie, and Ashraf Jaafar "SAFE HOME© An Advanced Home Security System", Department of Mechanical Engineering American University of Beirut Beirut, Lebanon, Volume 2, 2009 , pp. 234-239.

[10] Md. Ansarul Haque Halima Akhter, "Comparative analysis of authentication and authorization security in distributed system", International Journal of Research in Engineering and Technology, vol 3, num 11, 2014.

[11] About Face ID advanced technology, 2017. [Online]. Available: https://support.apple.com/en-us/HT208108

[12] K.Seetharaman and R.Ragupathy, "Iris Recognition for Personal Identification System", Procedia Engineering, vol. 38, 2012, pp. 1531-1546.

[13] doXteam. [Online]. Availble: http://www.doxteam.com/

[14] Hangzhou Zhian Technologies Co., Ltd, ZFM – 20 Series Fingerprint Identification Module (User Manual), 2008.

[15] Danny Thakker. Bayometric, [Online], Available: https://www.bayometric.com/false-acceptance-rate-far-false-recognition-rate-frr/, Accessed: 2018.

New Youth Initiative for Advanced STEM Education

Andrej Angelovski
Student at Faculty of Computer Science and Engineering
Skopje, Macedonia
andrej.angelovski@students.finki.ukim.mk

Kliment Serafimov
Student at Massachusetts Institute of Technology
Cambridge, Massachusetts, United States
kliment@mit.edu

Mile Jovanov
Faculty of Computer Science and Engineering
Skopje, Macedonia
mile.jovanov@finki.ukim.mk

***Abstract*— In Macedonia, the public high schools are much more focused on the breadth of a whole set of subjects, rather than the depth of a few. *School of The Future* is a youth initiative for talented primary and high school students. In this paper we will present the activities and results of *the School of The Future* project. The paper will primarily be a synopsis of the steps taken to increase local engagement in advanced STEM education in Macedonia. The goal is to demonstrate the extent to which motivated students with the support of their parents and teachers can organize events and classes for the betterment of the advanced high school STEM community.**

***Keywords*— Education; Talented youth; Informatics; Mathematics; Physics; Science; Olympiad;**

I. INTRODUCTION

The competitions in Science (i.e. informatics, mathematics, physics, chemistry…) are very important for the educational environment. They give opportunity to the talented and highly motivated students to express their knowledge on national and eventually at international level and hang out with their peers with the same or similar interests. Moreover, competitions not only engage talented students, they also promote the interest in science for all the other student that get involved on the basic levels of the competitions or that just hear the success stories of their school peers. They are an important piece in the pursuit to increase interest in a subject.

In Macedonia, the main work in this field of education is done by some of the faculties in the Macedonia's universities. For example, there is the work done by the Computer Society of Macedonia (CSM) that every year organizes a series of competitions to select the best high school programmers. Competitions like these are very motivating for students who largely self-study to prepare for them. The computer society also organizes yearly preparation summer camps. In the last ten years, besides other results, CSM also created a community of students studying advanced programing in high school. There was almost a 10-fold increase (from 48 students in 2009 to about 450 in 2016 [1]) in the turnout at the competitions in informatics in 8 years. All the competitions are entirely conducted through the system called MENDO [2].

The success of the introduction of the competitions, together with the improvement in the computer science curriculum, was significant in the increased enrolment of students at the computer science universities [1].

Additionally, in Macedonia there are several other science societies who organize competitions for selection for International Science Olympiads such as the Math, Physics, Astronomy, Mechanical Engineering, Chemistry, and Biology Societies. They function independently and don't provide yearlong preparation for competitions and rarely provide even summer and winter preparation camps.

However, in Macedonia there is lack of existence of an organized support in the on the topic of advanced STEM (Science, Technology, Engineering, Mathematics) education, as opposite of the experience in the neighboring countries. Organizations like the Mathematical Gymnasiums in Serbia and Bulgaria proved to be the key to the success of their teams at the international science Olympiads [3]. Macedonia lacks such gymnasium, or for that matter any yearlong preparation for Olympiad level STEM competition.

With this in mind, the School of the future initiative was initiated by one of the young students, an international competitor in informatics. This bronze medalist of International Olympiad in informatics, Kliment Serafimov, initiated a series of events that quickly increased in the number of hours of participant engagement. Roughly 300 people participated in the events and classes. The main idea was to push the boundaries of the sphere of social engagement around advanced high school stem education. First major events of the initiative were: *Networking 101: International Youth Networking Academy*. Consequently, the *School of The Future* was established.

This paper presents the steps of the School of the future initiative, with the mission to provide mentorship and resources to talented high school students in the STEM field. It additionally explains how of a group or dedicated young people with help of their mentors created a yearlong privately sponsored preparation program for Olympiad level Mathematics, that started in August 2016 and attracted 40 of the most talented high school and elementary school students in the country. The organization and results of the main events of the initiative like winter preparation academies for programing and the summer events called "Networking 101: International Youth Networking Academy", editions 2015 and 2016 are explained in detail.

II. THE PILOT PROJECT - NETWORKING 101: INTERNATIONAL YOUTH ACADEMY

Networking 101 – International Youth Networking Academy is a summer academy for talented youth from Macedonia and worldwide. It aims to bring high school students together and allow them to develop socially and academically, for the aim of improvement the local (and global) society. There were two editions of the academy, in 2015 and 2016.

Networking 101 in 2016 was the second in a row summer academy that took place in Hotel Molika in the Pelister National Park from 11th to 19th June 2016.

The academy was structured into 5 branches: 3 being the main seminar and 2 structured as preparation camps. It had almost 90 people involved as participants, speakers, workshop leaders or organizers. 37 students were part of the three main courses, while an additional 18 attended the International Olympics in Informatics (IOI) and International Olympics in Mathematics (IMO) preparation sessions. Along with the 3 main organizers, there were 10 staff members and 14 teachers or guest speakers.

Many of the participants came from Macedonia, though a total of 11 international students were present, representing countries such as Botswana, Mexico, Israel, Germany, Turkey, and Bulgaria. Workshop leaders came from Peru, Slovenia, and the United States of America.

The academy aimed to bring together students that are talented in different spheres. It sought to cover a variety of topics and was structured in a way that provided efficient advancement for particular fields of learning.

The success of Networking 101 both in 2015 and 2016 portrayed the need and the potential benefit of establishing an educational space where talented young students interested in a variety of fields can develop and learn from each other.

III. OPENING OF THE SCHOOL OF THE FUTURE

The founding day of the school was 24.08.2016, thus creating the geometrical progression 2, 4, 8, 16 symbolizing the growth it aims to achieve. The day was celebrated with a full day event where three venues were organized:

- Open classes – Informational Event
- Youth of The Future – a youth conference
- Emotions of The Future – an artistic venue

These three venues served to initialize and commemorate the establishment of the School, as well as to demonstrate what type of activities it will pursue during its working.

A. Open Classes

The open classes aimed to present the mission, vision, staff, and services of the School of The Future (SOFT), as well as to give an overview of the material of the courses offered.

For the first time ever, the Open classes happened from 10am to 3pm on the founding day. Around 60 people attended, most of whom were students and parents interested in what the SOTF offered.

The Open classes started with explanation of the mission, vision and the structure of the school. After that, there were presentations about each one of the following courses:

- Competitive Mathematics
- Competitive Informatics
- Competitive Physics
- Introduction to Java
- Introduction to video game design

B. Youth of The Future

For the first time ever, this conference was organized on the founding day and served as a venue for presentation and discussion of issues, achievements and future hopes concerning the youth in Education, STEM (Science, Technology, Engineering, Mathematics) and business.

The conference featured some of Macedonia's most accomplished students, Macedonian students studying at some of the world's best universities, as well as local business leaders. The conference featured seven panel discussions and talks from 12 speakers, respecting the following agenda:

- SOTF - Background, mission, vision and goals
- Networking 101: International Youth Networking Academy
- The MIT experiences
- Studying at Oxbridge
- Young talents
- The role of youth in the IT and Start Up spheres of interest

C. Emotions of The Future

"Emotions of The Future" was an entertainment event organized at the founding day, with the purpose to promote young rising talents in the entertainment, music industry as well as art and design. It featured a wide variety of artistic expression such as poetry, dance, music, and comedy.

IV. IMPORTANT ACTIVITIES OF SCHOOL OF THE FUTURE

The Macedonian national Mathematics and Science competitions for youth are a series of competitions created with the intention of motivating the nation's highest achieving students to explore the fields of Mathematics and the Sciences further. Every year, between 4 and 6 students represent Macedonia at international Mathematics and Science Olympiads. The students are selected based on the results of the national competition, implying that all the competitive students aim to excel in exploring these fields compete to receive the honor of representing Macedonia at the International Olympiads, where the world's best young mathematicians and scientists come to represent their countries. The competitions that the SOTF focus to prepare students for are:

- International Olympiad in Informatics – IOI
- International Mathematics Olympiad – IMO
- International Physics Olympiad – IphO

In the following subsections we will explain all the important activities that were conducted in order to accomplish the given goal.

A. Activities of the first semester

During the first semester of development, the School of The Future organized one short course and three semester long classes. Being socially aware of the economical situations in many families in the country, all students who attended the classes received need and merit-based scholarships to lessen the financial burden of attending the classes for their families. The merit-based scholarships were determined according to two scholarship examinations that the School carried out.

The students who attended the classes are some of the most accomplished students in Macedonia who have demonstrated remarkable ability in the STEM field, more specifically in the subjects:

- Mechanical Engineering
- Competitive Mathematics
- Competitive Informatics
- Competitive Physics

1) Introduction to Mechanical Engineering and Avionics

The first ever course of the school was Mechanical Engineering which took place from August 24th to September 6th, 2016. It was executed as a project-based course led by Lochie Ferrier, an Aerospace and Astronautical Engineering student at MIT. He guided the team composed of three young talents through the development and completion of several hardware and software project over the course of 2 weeks.

The students were chosen based on their previous experience with STEM projects in informatics and mechanical engineering and were awarded need-based scholarships. The team completed 6 projects in the fields of mechanics and software.

2) Competitive Mathematics, Informatics, and Physics

The competitive Mathematics, Informatics, and Physics classes of the School of The Future attracted some of Macedonia's most prominent young mathematicians, programmers, and physicists from 7th to 12th grade.

The classes were offered on different levels in correspondence with different ages and skills that the students had prior to enrolling in the course. The students attended on average 6 hours of lecture a week and were examined on two-month basis. In total 30 students completed one of the courses offered.

All the mentors of the three courses have been previous competitors with high achievements in the subjects that they were teaching.

B. Winter Academies

The Winter Academies of the School of The Future brought 42 students together with 10 mentors in three subjects: Mathematics, Informatics, and Physics. In the academy participated 4 generations of competitors: elementary school, high school, university students, and teachers who have been previous competitors. The Academy happened from the 12th January to the 20th January 2017.

C. Summer Camps

The Summer Camps happened from 13th to 18th of June, where intensive academic environment was created, thus the students were offered advanced lectures on all Olympiad topics on the highest level to focus on the challenges ahead of them.

D. Second and Third Semester

In the second semester, the Competitive Mathematics classes continued. The classes were held regularly, on average 6 hours per week, up until the Junior and Senior Macedonian Mathematics Olympiad, where the teams that represent Macedonia in the international Olympiads were selected.

In the third semester, the Competitive Mathematics classes continued. The classes were held regularly, on average 4 hours per week. The total number of participants was 42 students distributed in 5 different groups according to the level.

In addition, in the breaks between the lectures, we stimulated playing table tennis, billiard and board games, so the students can also develop socially and create friendship with their talented fellow students.

V. Background

Building the network of young excellent students and highly respected professionals, that will allow sustainability of the idea, is one of the main goals of this initiative. In this section we mention some of involved persons that gave their time and expertise in the development of the project.

People who initiated this initiative are young Macedonian students, most of them studying at some of the most prestigious universities in the world. The founder of the School of The Future is Kliment Serafimov, a student of Computer Science at the Massachusetts Institute of Technology. He has represented Macedonia at international informatics competitions for 5 years in a row. Driven by passion to connect people and share knowledge in 2014 and 2015 he co-founded and co-organized two international youth academies – Networking 101, pilot project for what will become SOTF.

In addition, all the mentors had several years of experience in teaching talented students and have themselves been competitors at national and international competitions. Some of them are: Stefan Lozanovski (Mathematics), Dimitar Trenevski (Mathematics), Marija Mihova (Mathematics and Informatics), Slagjan Stankovik (Mathematics), Kliment Serafimov (Informatics), Dejan Maksimovski (Phisics), Lochie Ferrier (Physics and Machine engineering).

It is worth to mention that most of the students have had significant achievements for their ages, and some of them have had extraordinary achievements. Marko Calasan is one of them, he is noted for being the youngest certified computer systems administrator at the age of eight and the youngest certified computers systems engineer at the age of nine. Another one is Orhan Bagashov who built his own 3D printer at the age of 13. Darijan Shekerov has built 15 android apps, until he was only 9th grade.

Additionally, among the students there are many state physics, chess, mathematics, and informatics champions as well as students who have shown extraordinary ability in passion for pursuing knowledge in the STEM field.

VI. Results

The following are results from the competitions in Mathematics and Informatics featuring the students that went through some of the courses in the School of The Future. The results are given in Table I (Informatics) and Table II (Mathematics).

Table I. International Olympics in Informatics [4]

Year	SOTF Participants	Medals
2016	4 (out of 4 in MKD team)	1 Bronze
2017	4 (out of 4 in MKD team)	1 Bronze

Table II. International Olympics in Mathematics [5]

Year	SOTF Participants	Medals
2016	4 (out of 6 in MKD team)	3 Honorable Mentions
2017	4 (out of 6 in MKD team)	1 Bronze 3 Honorable Mentions

Additionally, the preparation for physics competitions have shown positive trend in Macedonia, as more students engage in the challenge to prepare for competitive physics.

VII. Conclusion

The motivation of School of the future initiative was to initiate a movement for improving education through a progressive reform with a long-term vision, initially in Macedonia for the most gifted students, and later propagate the platform to the whole world and for all students.

The impact of the SOTF was is obvious if we take into consideration the results that the students from the school produced that are extraordinary. Also, a network of valuable members was developed, and this network of people is continuing the efforts and activities in the spirit of the main idea.

Since the model is now well established in Macedonia, we should introduce the idea of implementing the educational model internationally.

Institutions like this one in our neighboring countries have shown amazing results on international competitions.

In general, many of the leaders in STEM in Serbia have graduated from their top Mathematics gymnasiums. The results at international Olympiads of the Bulgarians, Croats, and Serbians are comparable to the results of the competitors from countries like UK, France, Germany, and other bigger and financially more powerful countries. This has been the result of concentrating talented youth in Mathematics gymnasiums which are competitive, rigorous, and advanced topics are being studied. In Macedonia there is no such elementary or high school institution, although some gymnasiums show certain traits for developing those kinds of activities.

VIII. Future Work

The SOTF aims to serve as this institution for Macedonia and improve on the existing models by installing more entrepreneurial values in the students through the courses.

Mentoring the best high school students in STEM in Macedonia provides the unique opportunity to serve as the primary channel through which students and professors from world renowned universities like MIT, Harvard, Princeton, Oxford, Cambridge, and other can contribute towards the development of education in the country.

Finally, this is possible by a multitude of opportunities that are offered through programs such as MISTI at MIT which focus to connect their institution with leading talent in the rest of the world; the connections through that program could link the Macedonian leading professionals with leading professionals from the world's most distinguished universities.

Acknowledgements

This paper was partially supported by Faculty of Computer Science and Engineering at "Ss Cyril and Methodius" in Skopje.

References

[1] M. Jovanov, N.Ackovska, E.Stankov, M.Mihova, and M.Gusev, "A decade of engineering computer engineers", 2017 IEEE Global Engineering Education Conference (EDUCON), Athens, Greece.

[2] M. Jovanov, B. Kostadinov, E. Stankov, M. Mihova, and M. Gushev, State competitions in informatics and the supporting online learning and contest management system with collaboration and personalization features MENDO, Olympiads in Informatics, 2013, Vol. 7, 42-54.

[3] A. D. Todorova, V. M. Erakovic, and I. Tonov, "Comparative Education of Gifted Students in Mathematics in Some of the South - Eastern Europe Countries", MASSEE International Congress on Mathematics, MICOM 2009, Ohrid, Macedonia.

[4] The official website for IOI, http://ioinformatics.org/index.shtml, information that is used is since 2016.

[5] The official website for IMO, http://www.imo-official.org/default.aspx, information that is used is since 2016.

Interactive Digital Environment For Learning Historical Events

Lidija Jovanovska
Faculty of Computer Science and Engineering
Skopje, Republic of Macedonia
lidija.jovanovska@students.finki.ukim.mk

Antonio Dimovski
Faculty of Computer Science and Engineering
Skopje, Republic of Macedonia
dimovski.antonio@students.finki.ukim.mk

Boban Joksimoski
Faculty of Computer Science and Engineering
Skopje, Republic of Macedonia
boban.joksimoski@finki.ukim.mk

***Abstract*—The purpose of this paper is to outline the process of building a 3D environment depicting historical events with the intent to be used for educational reasons, while the primary assignment was to create a virtual representation of one of the exhibition rooms in the Museum of The Macedonian Struggle in Skopje. By developing this kind of virtual environment, students, as well as enthusiasts, can further their knowledge in specific points in history and grow their interests accordingly. The making of the 3D room required several skills and tools, such as the design of the objects, the real-time rendering of the scene and the definition of the ruleset for the user interface. The human models, along with their clothes and accessories, were recreated based on previously acquired reference images, in Autodesk Maya, a 3D computer graphics software. In this instance, the models were made to represent two historical figures from Macedonian history; Miss Ellen Stone, an American missionary sent to the Balkans, and Jane Sandanski, the leader of the Macedonian Revolutionary Organization. Furthermore, to tell the story of the Miss Stone Affair of 1901 in a more engaging way, the team created an interactive animation that covers the summary of the event from the perspective of Miss Ellen Stone. Apart from the audio-visual context of the 3D room, the virtual scene is filled with facts and information regarding the event. A fraction of that information is displayed directly through the various interactive objects such as the books, magazines and pictures. On the other hand, additional information can be gained through pop ups that are activated by navigating closer to their designated objects. To achieve this level of interaction, the team had to use assets from the popular game development platform Unity. By going through an educational virtual experience of this kind, the user can delve deeper into the matter of interest and undergo the process of learning in a new and easily accessible way.**

***Index Terms*—learning environment, 3D animation, real-time interaction, multimedia environment, gamification**

I. Introduction

The standard teaching methods, used for centuries, are becoming obsolete. Today's youth, known as generations Y and Z as well as future generations generally find traditional methods difficult to learn. The youth is accustomed to digital devices, an always on-line experience and on-demand services that are created with the sole purpose of retaining attention. Thus, it is very hard to keep the youth interested in relevant subjects while practicing the standard teaching methods.

One of the ways to improve learning techniques is to embrace new technologies and create an educational experience that blurs the line between learning and entertainment. This means that teachers must devise ways of learning that expand knowledge and present it in a stimulating environment.

Computer gaming has become one of the most engaging and entertaining experiences. The idea of producing a symbiosis between games and teaching has been present for a long time and it is appropriately implemented in various fields. This paper describes the creation of an on-demand, web-based virtual 3D environment that recreates a realistic historical museum setting. The environment is designed for interactive use and adopts game techniques to enhance user interaction. Our goal is to create an enjoyable and engaging setting where users can learn by using well adopted game mechanics.

II. Background Work

The ideas realized throughout the development of the environment are inspired by several virtual restitution and usability enhancement projects:

- The virtual restitution and virtual life simulation of the church of Hagia Sophia,a highly complex and endangered heritage edifice located in Istanbul, Turkey. The project describes the achievement of a photo-realistic simulation of the selected space, based on accurate cultural and archeological data. The recreated model of the church could be further used as a passive walk-through or a freely wanderable 3D real-time environment. [1]
- In 2006, the Smithsonian American Art Museum (SAAM) launched the website Meet Me at Midnight, an interactive, online adventure for kids eight to ten years old that presents an art history mystery in cartoon style.
- Later, in 2008, SAAM created the world's first museum-based alternate reality game (ARG) titled Ghosts of a Chance. The scavenger huntlike game took place in the physical world (at SAAM and the Congressional Cemetery) and also in the virtual world (on the museums website and other sites).The game was a 2009 Official Webby Honoree in the environmental and experience marketing category. Aside from The Collective that targets young adults, the museum also has a website that targets younger children called Wacky Kids Website. The website is interactive with art activities to do at home or school and links and book titles at the Denver Public Library.

- The Dallas Museum of Art in Texas opened a Center for Creative Connections inside the museum, providing both analog and digital ways for visitors to explore the creative process by interacting with art. The center also houses the museums Tech Lab, where visitors are encouraged to "experiment with and use technology in unique ways to respond to works of art in the Museums collection." At the Tech Lab there are Web-based interactive programs on computer stations, and drop-in classes and workshops for adults and families about social media, gaming, video, sound design, stop-motion animation, 3-D design, and more [2].

III. Project Description And Goals

The original idea of the project was to create a realistic digital clone of a museum setting displayed at the Museum of the Macedonian Struggle for Statehood and Independence. The setting is comprised of the wax sculptures of Miss Ellen Stone and Jane Sandanski, as well as other authentic objects from that period, including Jane's Mannlicher rifle and various books, pictures and newspaper articles that describe the event (Fig. 1).

Fig. 1. The museum setting including the wax sculptures and picture in the background.

As the project progressed, it was decided that it can be further expanded as a web accessible learning tool that can find its use in a standard curriculum, using gamification as a method for teaching important historical events.

A. Historical Context

The scene depicts one of the most important events in the history of the Macedonian people's struggle for independence and affirmation. The protagonists of the story are Miss Ellen Stone - an American missionary working on the Balkans and her captor Jane (Yani) Sandanski - a veteran revolutionary. In the early 20th century Macedonia was a tremendously volatile part of the world. Having been under Ottoman reign for more than six centuries, Macedonians sought independence devotedly. In order to draw attention from the American leaders in August, 1901 the IMRO (Internal Macedonian Revolutionary Organization) directed the kidnapping of Miss Ellen Stone and her pregnant fellow missionary Katerina Tsilka. [3] Aside from its obvious melodramatic qualities, the Stone Affair is noteworthy for a number of reasons. The 66,000 Miss-Stonki, as the revolutionaries called the ransom money, helped finance the Macedonian uprising of 1903. As far as the United States was concerned, the Miss Stone Affair introduced America to one of the burdens of major power status as it represents "America's first modern hostage crisis". [4]

The highlights of the previously described event, told through the perspective of the Miss, are displayed in the scene through the use of day-to-day language and visual cues.

IV. Methodology

A. Reference Data Acquisition And Process Description

Over the course of several visits to the museum, the museum setting was digitized through the acquirement of a large number of high quality two-dimensional photos using semi-professional camera. All the photographic data is used in a later stage as base for the design and virtual restoration of the models and material textures.

In the early stages of the project, the pursuit of an automated approach using photogrammetric techniques was considered [5]. Images of Ellen Stone's figure were imported in the 3D reality capture software - Autodesk ReCap for the purpose of testing the quality of commercial photogrammetry tools [6]. The resulting models and textures obtained from ReCap led to discarding any photogrammetric solutions due to the following reasons:

- The inability to isolate the object in almost perfect lighting conditions during the photographing sessions.
- The complex topology of the mesh output which rendered the models as too complex and unusable without significant retopology work.
- The inability to further edit the mesh and impractical controls for animating the character.

Furthermore, a custom pipeline was developed, akin to standard game development, to achieve an environment that is highly flexible to manage. The steps that constitute the pipeline, in sequential order are:

1) 3D modelling and UV mapping
2) audio creation and sound engineering
3) texture creation and projection on the 3D models
4) developing character controls for speech animation
5) character animation
6) importing the 3D assets in game engine
7) creating first person based experience in the game engine
8) adding various interactive elements to the scene
9) publishing the project as WebGL based environment
10) designing and programming browser-based user interactions with the environment

B. 3D Modelling

The design of models of an organic nature is a time-demanding process that can implicate possible problems in the latter phases of the pipeline if not done properly. To evade the potential risks, the free, open-source humanoid generation software MakeHuman was used [7]. Human generation softwares generate complex, anatomically and topologically

correct human models using numerous parameters. MakeHuman produced satisfactory results that closely resembled the characters in the scene and needed very little touch-up work.

For the creation and manipulation of the rest of the polygonal 3D objects in the scene Autodesk Maya 2017 was chosen as it offers various tools for modelling, rigging and animation and it was best known for the creators. We used polygonal based modelling due to its ease of use for real-time rendering purposes. Non-organic models bound to the humanoid characters constituted of distinctive clothes and personal accessories that add to their genuineness and historical context. Other objects including books and posters were simple to create but relied heavily on textures to provide proper information.

One of the main challenges was the recreation of the hair. Maya's generic dynamic hair system called nHair was not a fitting solution nor other hair solutions (like Maya Paint Effects strokes or different plug-in based solutions). Hence, the hair was recreated using traditional polygonal modelling techniques and the quality was subsequently compensated with higher quality textures. For those textures to be properly distributed along the surfaces, they need to be projected along the UV map of the object. Part of those maps were automatically generated and the rest of them were created with suitable tools.

C. Audio Creation And Sound Engineering

The entirety of the audio creation and sound design process was outsorced to a professional production enterprise. A professional voice actress was hired for creating the voice of Miss Stone according to a predefined script. The resulting sound file was accordingly processed for use in the virtual environment.

D. 3D Texturing

This phase concerns the creation and projection of the UV textured maps. The process of generating the textures follows one of three paths:

- Sampling of the photographic data and post-processing in Adobe Photoshop
- Creating the textures with brushes and tools in Adobe Photoshop
- Using previously generated textures found on the Internet

Originally, the textures were exported in a 2K resolution. This produced high visual quality along with adding to the "heaviness" of the scene. Adjustments were conducted, leading to the downsizing of the less relevant textures to a 1K resolution. After the appropriate importing in Maya, the UV textured maps were applied on the models, tested and subsequently integrated in the scene. The objects of a reflective nature were assigned the material type Blinn along with the corresponding texture, while the matte surfaces were shaded with a Lambert material.

The model and texture of the main character, Miss Stone, can be seen on fig. 2.

Fig. 2. The 3D model of Miss Stone with the mesh topology (left) and the textured version of the model (right).

E. Developing Controls And Animation

For the achievement of higher realism a speech animation was rendered. By viewing it users can easily engage with the story being told. This detail is also useful for its intuitive interaction method.

The animated clip was brought to life with the use of a mesh deforming technique in Maya named Blend Shapes. Facial transformations were applied on Miss Stone's face, resulting in a pose for each of the following phonemes: V-F, T, S-Z, M-B-P, A, E, I, O, U. By utilizing the advantages of visemes, the number of blend shapes decreased and with that, the size of the scene. [8] Afterwards, audio synchronization tests were conducted and interpolation improvements were implemented through the animation editor. The animation was exported as an .fbx file and later imported along with the other assets previously acquired into the game development platform - Unity.

F. Environment Programming

The idea was to design the scene in a way in which students would be able to intuitively interact with it. Therefore the models were placed in a room and each model was given its own interaction. The keys 'W' , 'A' , 'S', 'D', were assigned for moving around the scene and the 'Space' key was assigned for starting Miss Stone's voice line. Every other object in the room had its own pop-up which was triggered when the user was near it and it displayed information about the specific object. In order to know when the user was near the object, colliders were added to every object. This way, students wouldn't have to worry about what could and couldn't be clicked in the room (3 . To achieve this game-like appearance the best option was to use Unity. Then, for the web-like parts e.g the pop-ups, Javascript event functions were used and integrated with Unity. The information/text about the objects was kept in the Javascript file so it could be changed easily when needed. The final project was exported as a WebGL project in order to be used online.

Fig. 3. Additional information that is displayed when an interactive element (collider) is triggered.

TABLE I
STATISTICS FOR THE 3D SCENE

Property	# of vertices	# of faces
Miss Stone	21533	28229
Jane Sandanski	38354	37000
Environment	5986	5576
Total	65625	70599

V. RESULTS AND FUTURE WORK

Using all of the mentioned assets and methods, a digital environment was created that can be accessed at the following url: `http://www.mmb.org.mk/elenston`.

A screenshot of the resulting environment is displayed in Fig. 4.

The resulting scene, exported in a WebGL format along with accompanying files (html files, scripts, images, textual information, etc.) is approximately 70MB, while the details for the 3D objects can be found on table I

Fig. 4. Screenshot of the resulting learning environment.

With the intent of being presented in an online real-time 3D environment, the scene's principal design was based on the balance between the fidelity of the reconstruction and the size of the models. Therefore, the optimization of the size of the models will result in lower loading times, while retaining a high frame-rate. Thus, the 3D environment manages to maintain a satisfying user experience, despite being smaller and faster than the default. Having every consideration in view, the final scene was designed with the use of 47 textures and 59 materials, every detail optimized for achieving best quality/size ratio.

The environment was presented to a small selection of elementary and secondary school History teachers. Due to the large interest, the museum organized extra days with the intent of training additional 200 teachers to apply the environment as a learning tool. The overall satisfaction of the trainees, as well as different teaching methods, have sparked interest for further development of the environment and creating more similar tools for different contexts and subjects.

VI. CONCLUSION

With the use of new information technologies and the essential support of photographic data a complete methodology for the restoration and virtual presentation has been described in this paper.

Virtual restitution of highly complex museum exhibitions requires accurate choices for each phase of the process. For that, special attention must be given when the models have to be prepared for real-time platforms.

The goals were to create an engaging environment that implements game-based mechanics to improve learning experiences. A myriad of tools were used to create a realistic real-time visual experience, in effort to grow the interest for the subject. The feedback was more than encouraging, and it opened new ways of narrating and visualizing important events from the rich Macedonian history.

Anyhow, the method of crafting such an environment we presented is not restricted to featuring historical events and can be furthermore applied to enhance the academic work of many, if not all, branches of education.

VII. ACKNOWLEDGMENTS

This work is partially financed by the Faculty of Computer Science and Engineering and the British Council of Macedonia.

REFERENCES

[1] A. Foni, G. Papagiannakis, and N. Magnenat-Thalmann, "Virtual hagia sophia: Restitution, visualization and virtual life simulation," 2002.
[2] S. S. Bautista, *Museums in the digital age: changing meanings of place, community, and culture*. Rowman & Littlefield, 2013.
[3] M. Cornis-Pope and J. Neubauer, *History of the Literary Cultures of East-Central Europe: Junctures and disjunctures in the 19th and 20th centuries. Volume IV: Types and stereotypes*. John Benjamins Publishing, 2010.
[4] R. B. Woods, "The Miss Stone Affair."
[5] M. K. Yves Egels, *Digital Photogrammetry*. CRC Press, 1 ed., 2001.
[6] A. ReCap, "Autodesk recap," *Accessed March*, vol. 30, 2014.
[7] M. Team, "Makehuman," *Online] http://makehuman. org, Accessed October*, vol. 12, 2010.
[8] T. Chen and R. R. Rao, "Audio-visual integration in multimodal communication," *Proceedings of the IEEE*, vol. 86, no. 5, pp. 837–852, 1998.

Smart Braille – Accessibility is the key

Nikola Tasevski
Faculty of Computer Science & Engineering
Skopje, Macedonia
t27nikola@gmail.com

Stefan Tasevski
Faculty of Computer Science & Engineering
Skopje, Macedonia
stefan.tase97@gmail.com

***Abstract* — The project is a development of a fully functional mobile application that primarily aims to provide faster communication for the blind and visually impaired people through their mobile phones. The functionalities of the application can be divided into two purpose sets. The first is writing on a mobile phone, and the second is reading a text with the help of it.**

Keywords — braille, accessibility, design, keyboard, app, API

I. INTRODUCTION

Writing on phones by the blind or visually impaired people is carried out in two ways, by using the standard system keyboard whose main disadvantage is the time it takes when searching for the characters or using the speech-to-text feature which for obvious reasons is not always possible and even worse, completely violates the privacy of the user [1],[2]. In this respect, our application offers a Braille alphabet based keyboard (Currently only English Braille alphabet is supported) that consists of six keys that occupy the whole screen. They correspond to the six points used to mark all the characters in Braille alphabet. Although now, instead of detecting the combinations with the tips of the fingers, they are tapped on the screen to write the characters. Through appropriate adjustments such as the different beeps for the keys, a variety of swipes across the screen (to delete a character, delete a whole word, enter a space, enter a new row etc.) simplicity and speed are provided in using the keyboard. By doing this, the blind and visually impaired people knowing the alphabet can write messages privately, quickly and without error. This is the first mobile application that implements a system keyboard based on the Braille alphabet that is completely designed for usage by the blind and visually impaired people. There are other similar applications but they are more for educational purposes, providing a way to learn the Braille alphabet, but not a way to use it for texting.

II. THE KEYBOARD

A. Layout

While designing the keyboard layout we were guided by the needs of our target group of users. That being the blind and visually impaired people, it is best that the keyboard takes up the full display of the phone. Because our keyboard is based on the Braille alphabet, it is only natural to have six keys equal in size, which correspond to the six dots used for marking the characters in the Braille alphabet as shown in Fig. 1. Although, it is designed to take up the full screen, dependent on the application in which it is used, top or bottom bars may appear on the screen.

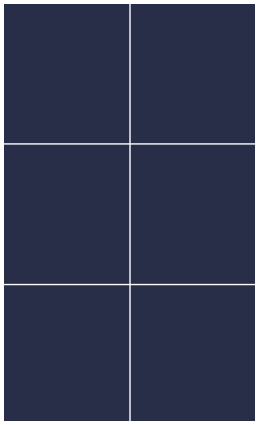


Fig. 1. The custom keyboard layout.

B. Functionallity

1) Typing: Normally the different combinations of the six dots are recognized by moving the fingertips across them [3], [4]. On the other hand, using our keyboard to type the different characters you have to tap the right combination for them on the corresponding keys. The activity diagram, which shows the flow of events when using the keyboard, is shown on Fig. 2. The problem that occurred is how to know when the user is finished tapping the combination he wants. Our first approach was using a timer, but that proved to not be very useful

because of the different writing speed of each individual. Additionally when TalkBack is activated the detection of taps on the screen may get slower depending on the phone used. Because of these two reasons the idea of using a timer was thrown out and we came up with a different approach. To finalize the tapping combination the user just needs to tap again the last key he pressed. The order of pressing the keys is not important only the combination they represent. The tapping combinations are inserted inside a database on the first run of the application and they are shown in two separate list views. One for letters, words and punctuation, and the other for numbers, since the Braille alphabet has the same combinations for numbers and some letters. The list views are accessible for the user if he needs to check or refresh his memory.

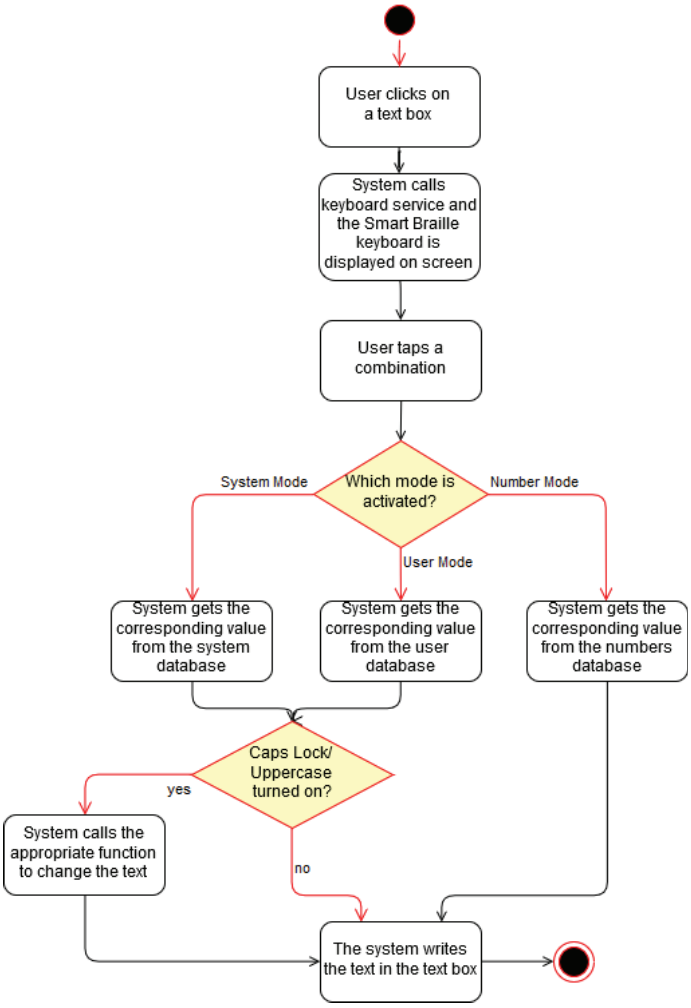


Fig. 2. Activity diagram for the keyboard implementation.

2) Movements: The typing of the different characters is explained in the section above, but for a keyboard to be fully functional there are more requirements. For the reasons explained before we reserved the whole screen only for tapping the combinations and implemented the additional functionalities using different movements across the screen. Those movements are not regular swipes for a couple of reasons. First of all, when TalkBack is activated for a swipe to be detected the user has to swipe with two fingers and for the sake of faster usage we decided our movements to be done with only one finger. The second reason, which came up while testing the application, is that swipes seldom are mistakenly detected for taps. The final reason is simply the lack of different swipes we need to implement all the necessities. With that being said, our implementation of the movements consist of pressing on one key, dragging to another and releasing the finger on it. But, for easier reading in the text that follows we will address this movements as swipes. Our keyboard supports ten different swipes listed with explanation below.

a. Space – Swipe right
b. Backspace – Swipe left
c. Uppercase* - Swipe up on the left side of the screen
d. Number Mode* - Swipe up on the right side of the screen
e. Caps Lock* - Swipe diagonally from bottom left to top right
f. Delete Word – Swipe diagonally from top right to bottom left
g. New Line – Swipe diagonally from top left to bottom right
h. User Mode – Swipe diagonally from bottom right to top left

   (Explained in section C)

i. Close Keyboard – Swipe down on the right side of the screen
j. Change Input Method – Swipe down on the left side of the screen

* Standard Braille rules with special characters also implemented.

To disable the enabled modes, the same swipe needs to be repeated or another mode to be enabled.

3) Additions for accessibility: The whole design of the keyboard is guided by the accessibility requirements [5]. Further than that we implemented another set of features which will provide smoother experience for the blind and visually impaired. In order to give the users a

way to know which key they are pressing and easily recognize if they mistyped the six keys have different tapping sounds. For simplicity, only two different system sounds are used. One for the odd and one for the even keys. That way every key has a different sound from its neighbors. On the most phones, TalkBack pronounces the character that is written or deleted with any keyboard including ours. Some other events, on the other hand, such as turning on the Uppercase, Number Mode etc. are not recognizes by the TalkBack. In that manner a simple solution came up – using toasts. With them a short message is shown and automatically read by the TalkBack.

C. *Customization*

To achieve a faster typing experience for our users an additional feature was added to our app - the keyboard customization. The idea is to give the users a way to add personal tapping combinations for words or phrases they most often use while typing. However all of the 64 Braille combinations are already in use so User Mode was implemented (the activation is mentioned earlier). When User Mode is activated all of the tapping combinations are free to be personalized by the user. For that we designed a special activity (shown in Fig. 3 with the activity for system combinations alongside) where the users are free to add custom values for the tapping combinations. The user can either write the combination himself or get a valid combination by clicking a button, which gives the user the first available tapping combination. The given valid combination starts with the first free combination that can be entered with the smallest amount of key presses. When the six combinations that can be entered with using one key press are full, it moves to the combinations that can be entered with two key presses and so on. These combinations are in a growing order so the smallest number will be picked. This way the users can quickly remember the combinations they entered. The combination is written with the numbers that correspond to the keys that are being tapped. The keys on the left column of the screen are labeled from 1 to 3 and on the right from 4 to 6. The pairs of combination and value are added to the database and can be removed at any time. When User Mode is activated all of the rules in the Braille Alphabet do not apply. In this activity the user combinations can be viewed within a list that functions well with TalkBack. The list is a scroll view, so our users have to use two fingers while swiping if they are using TalkBack.

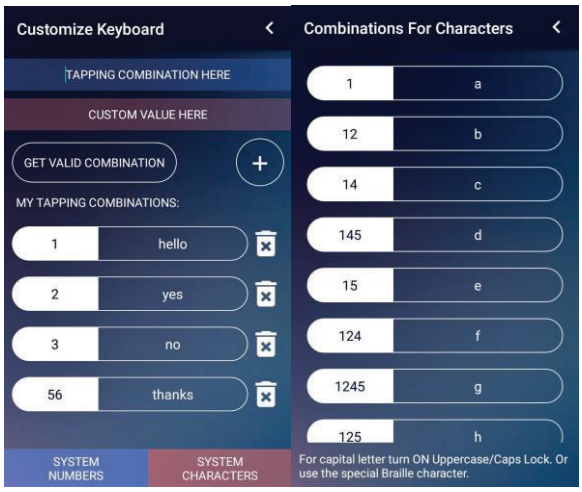


Fig. 3. Customization activity (left) and System characters activity (right).

D. *Provided Tutorial*

The app provides a simple tutorial in which through simple tests users can test or practice all of the features that our keyboard provides. The text, layout and design had been specifically made to be compatible with using TalkBack [6]. The tutorial takes couple of minutes and is interactive so that throughout voice confirmation the users can quickly get the hang of the keyboard functionalities. The tutorial is implemented with a multipage view and the users can move from test to test with swipes. It starts with simple tests for both the letters and the numbers, continues to explaining the movements implemented in the keyboard and finishes with a final test that wraps all of the functionalities explained. The tutorial can be repeated at any time and we strongly recommend to be passed before the first usage of the keyboard. A few steps from the tutorial are shown in Fig. 4.

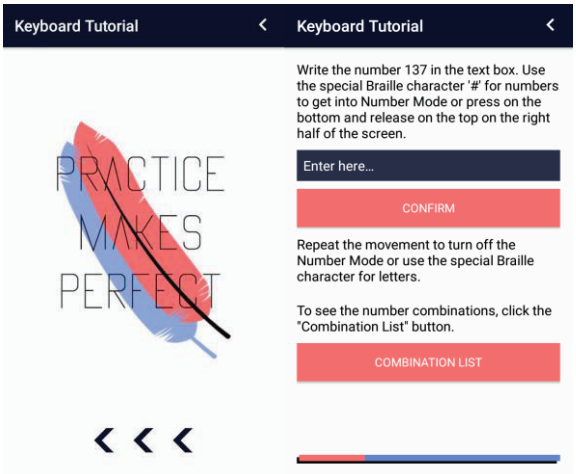


Fig. 4. Two steps of the tutorial.

III. TEXT READER

Google provides a set of very useful APIs [7] that can be implemented in any app for various purposes. Although they are not primarily aimed towards developing accessible apps, we feel that their usage in this field is the strongest. The APIs are mainly used to detect different shapes from pictures, starting with fruits, household objects, faces, letters, etc. Because our app is designed to be a handy tool for communication, we incorporated the API for detecting letters. Currently it only works for the German family of languages. It scans the image and extracts the letters it recognizes grouping them in words, sentences and paragraphs. In order to get a more natural approach with a couple of twitches in the code we designed it to work directly with the phone camera. In that way, different frames are treated like regular images and the detected text is transferred in a text box shown on the bottom of the screen, as shown in Fig. 5. Once there is a content in the text box it is instantly read to the user using Text-to-Speech. The only problem with this design is that with different movement a new frame is being processed. This results in the text being constantly read to the user. Aside this minor problem this feature works as it is intended – To be a helpful tool to our user group that can help them read the text that is all around them but not available to them.

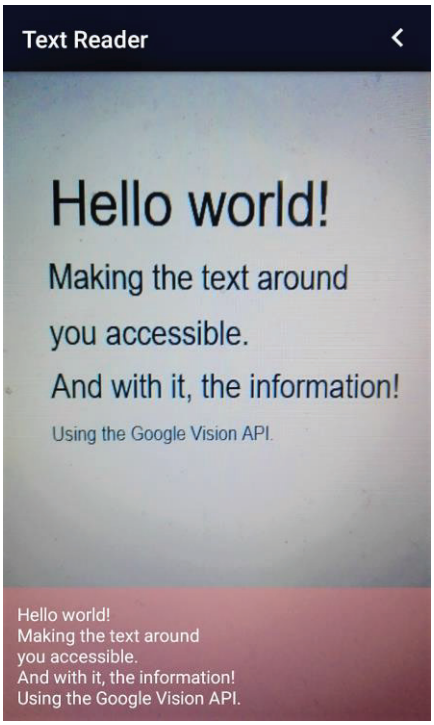


Fig. 5. Example of the Text Reader functionality.

IV. DOCUMENTATION

The documentation section in our app is a short-written guide on what you have to know about our app. We briefly explain what our app provides and point out some tips on how to use our app for our users to have a better understanding. There is also a guide on how to enable and set the Braille keyboard to your default mobile keyboard. Accordingly, there is a button which when pressed opens a list of Input Methods installed on your phone.

V. TESTING

To test our app's efficiency one visually impaired person tested it. The results were better than expected. He had majority of the knowledge needed for using the keyboard. The biggest part of it is knowing the Braille combinations and phone usage with TalkBack. The only little setback were the swipes because normally with TalkBack on, users have to swipe with two fingers. However, he quickly learned them and with time got used to them. The fast progress while testing gave us an understanding on what our app can provide. Since the keyboard takes the whole screen, mistyping mistakes were very rare and he easily understood and worked with the layout of the whole application. Even though the app was tested on only one person, it provided him with privacy, efficiency while typing and more control over his phone. We are working towards a cooperation with associations for blind and visually impaired people to help us perfect the user experience so that it can be a part of their everyday life.

VI. CONCLUSION

In conclusion, Smart Braille provides a better, more private and more effective mobile texting communication for the blind and visually impaired people, it is a great tool and we are eager to see how far it will go. However, we are not planning to stop here. Our next goal is to make this app available for iOS. We want to make implementations for different language interpretations of the Braille Alphabet. Most important of all, constant updates are necessary in which we will consult with our users, so they have a better experience. This paper is based on the first version of the app and stating that the app will progress through time with constant needed updates and with user surveys and reviews.

REFERENCES

[1] http://www.afb.org/info/living-with-vision-loss/using-technology/cell-phones-tablets-and-other-mobile-technology-for-users-with-visual-impairments/touchscreen-smartphone-accessibility-for-people-with-visual-impairments-and-blindness/1235
[2] http://www.thisisinsider.com/how-blind-people-use-smartphones-2017-2
[3] Braille, Louis (1829). Method of Writing Words, Music, and Plain Songs by Means of Dots, for Use by the Blind and Arranged for Them.
[4] https://en.wikipedia.org/wiki/Braille
[5] https://www.interaction-design.org/literature/article/accessibility-usability-for-all
[6] https://www.androidcentral.com/what-google-talk-back
[7] https://cloud.google.com/vision/

Visualizing Real-time Global Assaults

Stefan Dzalev
Faculty of Computer Science and Engineering
Ss. Cyril and Methodius University
dzalev.stefan@finki.ukim.mk

Dajana Stojchevska
Faculty of Computer Science and Engineering
Ss. Cyril and Methodius University
stojchevska.dajana@students.finki.ukim.mk

Ivana Stojkovska
Faculty of Computer Science and Engineering
Ss. Cyril and Methodius University
stojkovska.ivana@students.finki.ukim.mk

Dimitar Trajanov
Faculty of Computer Science and Engineering
Ss. Cyril and Methodius University
dimitar.trajanov@finki.ukim.mk

***Abstract*—In today's world there is a constant, ever rising threat of terrorist attacks which is why our task, first as human beings and second as future engineers, is to give our contribution to raise awareness about suppressing this threat. By analyzing several applications regarding this topic, we realized we would stand our having real-time results. A crawler downloads and extracts data on a daily basis from a single reliable source – GDELT Project, which collects data by scraping numerous news websites, providing us real-time data. We update our PostgreSQL database with this data. After preprocessing the data, we opened a possibility for getting insightful knowledge about certain categories of assaults, mainly using spatial-temporal and linked data. As a final result of the task we created a user-friendly application which displays the knowledge retrieved from that data, as map and related links towards news. We have also put our focus on possible future improvements.**

Keywords—real-time visualization; GDELT project; assaults;

I. INTRODUCTION

Assaults has become a common social issue. It is defined as the unlawful use of violence and intimidation, especially against civilians, in the pursuit of political aims. It seems tough as we hear about assaults and terrorism more and more in the news, which is why it seems closer to us. These are the reasons we chose to explore this area and discover how terrorism has been developing and spreading over the years, as well as the dangers it presents over the world in general. Historical data from 1970 to 2015 was adopted to train and evaluate machine learning models. The model performed fairly well in predicting the places where terror events might occur in 2015, with a success rate of 96.6%. Moreover, it is noteworthy that the model with optimized tuning parameter values successfully predicted 2,037 terrorism event locations where a terrorist attack had never happened before [1].

According to the Global Terrorism Database (GTD), more than 14800 terrorism events of different types occurred globally in 2015 alone [2] [9], which not only caused the deaths of 38430 people but also caused the world to panic [3][4][5].

The GDELT[1] Project [6] is a "global database of society", supported by Google Jigsaw it monitors and gathers data from the world's broadcast, print and web news. It is one of the best free sources of real-time events [10]. It stores the world's events and daily updates its web site with the data from the events in the form of .csv files.

Data analytics and visualization have become an instrument for conquering the world's challenges. Having that in mind we created a system that gathers data about assaults and makes a visualization on a world map. Our data source is the GDELT Project web site.

With this endeavor we aspire to give our support to the long-lasting war against assaults by extracting, processing and using the data from the GDELT Project, related to assault events, and make a visualization of the different categories of assaults on the world map in a web-based application.

II. RELATED WORK

BioWar [7] is an effort to develop a scalable and precise simulation tool to examine disease propagation and agent behavior in response to disease and illness. We believe it will serve to help researchers understand, predict, and analyze weaponized biological attacks at the city level and engage in "what-if" analyses to help inform decision-making in this complex socio-technical policy domain. For example, it can be used in a "what-if" mode to examine the impact of and response to various weaponized attacks for contagious and non-contagious diseases under high-alert and no-alert conditions. Unlike traditional models that look at hypothetical cities, in BioWar the analyst can examine real cities using census, school track, and other publicly available information.

[1] The Global Database of Events, Language, and Tone https://www.gdeltproject.org

Moreover, rather than just providing information on the number of infections, BioWar models the agents as they go about their lives - both the healthy and the infected. This enables the analyst to observe the repercussions of various attacks and containment policies on factors such as absenteeism, medical web hits, medical phone calls, insurance claims, death rate, and over the counter pharmacy purchases.

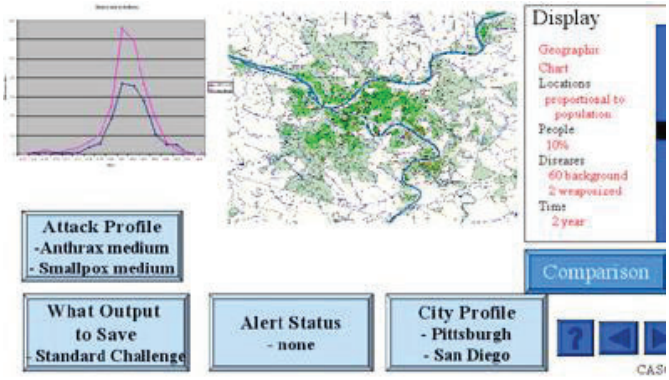


Fig 1. BioWar

Terrorism monitor [8] - The Jamestown Foundation's mission is to inform and educate policy makers and the broader community about events and trends in those societies which are strategically or tactically important to the United States and which frequently restrict access to such information. Utilizing indigenous and primary sources, Jamestown's material is delivered without political bias, filter or agenda. It is often the only source of information which should be, but is not always, available through official or intelligence channels, especially in regard to Eurasia and terrorism.

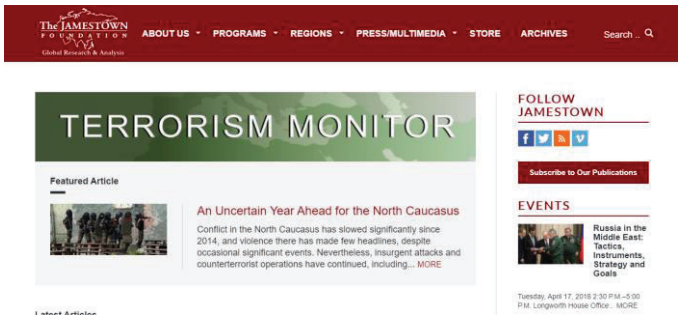


Fig 2. Terrorism monitor application

III. Collecting, Filtering and Storing the Data

The data on the GDELT Project web site is uploaded and daily updated in the form of .csv files in .zip folders. These .csv files contain all of the events that took place in the world and were reported by the online news media. Our first challenge was to create an automated mechanism that downloads these .zip folders, extracts the .csv files, filters the data and stores it in a PostgreSQL database.

TABLE I. Some of the .zip folders on the GDELT Project web site

Filename	*Filesize*	*MD5*
20180417.export.CSV.zip	14.2MB	d1e2c2f30be760c26a72ad29e5a3e1de
20180416.export.CSV.zip	12.7MB	7d451050bc5799958687649a5f622dd3
20180415.export.CSV.zip	7.4MB	81a5de73d489b5124ec173484f7cae07
20180414.export.CSV.zip	9.3MB	624a21a780a84e8e2e12b6dba7f8d9be

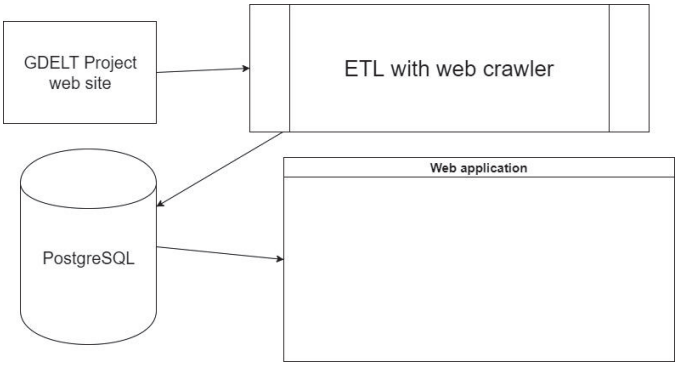


Fig 3. System architecture

A Python crawler downloads the .zip folder of the current day, extracts the .csv file and creates a temporary table in our database. The temporary table has attributes that are of type text. Then the crawler inserts all of the data in the temporary table and calls a stored function in the database that filters the data related to terrorist attacks from the temporary table, casts the attributes into the proper types and inserts it in the main table.

The several categories of assault that we filter out and use in our application are:

- Use unconventional violence
- Abduct, hijack, or take hostage
- Physically assault
- Sexually assault
- Torture
- Kill by physical assault
- Conduct suicide, car, or other non-military bombing
- Carry out suicide bombing
- Carry out car bombing
- Carry out roadside bombing
- Use as human shield
- Attempt to assassinate
- Assassinate

The data attributes that we filter out and use in our application are:

- Date
- Event code
- Goldstein scale
- Number of sources
- Country

- Country code
- Latitude
- Longitude
- Source urls

We have filtered these attributes from the 58 in the original data, because they are sufficient for the use in our application and are a good foundation for upgrading this project in further research on this topic.

IV. WEB APPLICATION

Using the Java Spring framework, we built the back-end of the web application. The front-end of the application, the user interface is built using Angular and node.js.

Our approach and philosophy to building this application was more pointed to generating a user interface that everyone could use. We aimed for the direction of creating this kind of user interface for a number of reasons, but in this paper, we will clarify the one we feel is of most importance. It is our perception of the world that guided us into thinking that the people that live in the countries that have the most terrorist attacks are the people who are most likely to be less computer literate than other people in the world. Having that in our mindset, while brainstorming for this particular project we decided that it is best to refrain from a complex user interface which, in part, is illustrated in Fig 4.

Fig 4. Visualization of assaults on the world map

A user can access our web application, choose the assault category he wishes to analyze and become aware of the assaults that happened the day before he accessed the web application. This way a user can analyze parts of the world, before he makes a trip somewhere, or a reporter can get not only the information that a particular event happened, but the spatial information as well. The geographic, spatial, visualization is made using Google Maps with customated points that show a picture at every point defining the particular category or type, if you will, of the particular assault.

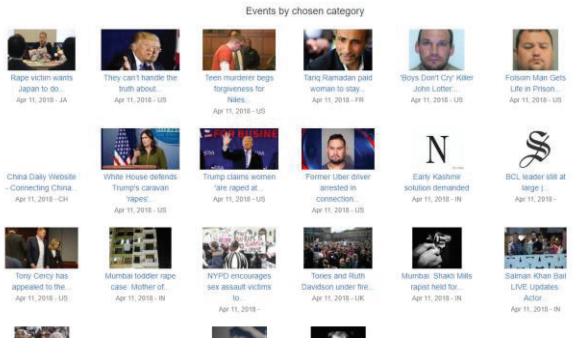


Fig 5. Analysis of assaults portrayed in the media

Because we live in a world smothered in fake, partisan, subjective news, a application that encourages multiple news sources is more than welcomed in the effort to improve, and in this case inform our society.

In the making of our application we felt bound to give the users of our application an objective, reliable source of information related to this particular topic. Guided by this phylosophy we feel we achieved our goal by allowing the user to filter out yesterday's assault events by category of attack and receive numerous online media sources regarding the particular type, category of assault, ordered by date. Regarding this matter, it is important to note that the user is receiving news from multiple online media sources about every assault event.

V. FUTURE WORK

In the modern-day world, especially in the IT sector, it is always an extraordinary aspiration to be engaged in creativity and improvements. Striving for better quality and extrapolation of the main features of the current web application during the many discussions that occurred we came to an agreement that there are several things that should be done that will improve the application.

A reasonable direction we could go in the further development of this application is creating an API that could be accessed by other software systems in order to get certain data analytics. These data analytics could include, risk forecasting, historical analysis of certain parts of the world and automated alert sending for subscribed systems. Another element that is exciting to add to the whole analysis of assaults in the later future is a temporal element, so we could analyze trends, patterns and more clear and exact predictions. A simple explanation of our objectives is not suffice to present the "big picture" so we analyze each of them in greater detail.

- Risk forecasting could be used by travel agencies, airline companies, even companies that work in the trade business. Our vision is this service to be free and open to everyone
- Historical analysis gives us knowledge of assault trends and how they have evolved and spread over time. The idea is to incorporate major events in

economy and politics into this analysis in the effort to find “triggers” for the different kinds of attacks in the different regions, countries of the world.

- The automated alert for subscribed users is mainly for people that live in a region and have subscribed for this service to be alerted if there is a high probability for an attack in the region where they live.

Our future work in the direction of the system architecture involves this system architecture to be redefined, restructured, and upgraded in a way. Our idea is to use the same crawler, incorporate more data sources and use the Hadoop Distributed File System (HDFS) along with Apache Spark for batch and stream data processing and PostgreSQL for analysis result storage.

VI. Conclusion

This project is an effort to help combating the threat of assaults that is reshaping our world for the worse. We are conscious that this help for which we strived will not be able to make much difference in the war against assaults at this moment. What we are sure of, and we are sure many will concur is that this application is a starting point for a much bigger software system, as the one described in our future work aspirations. By raising the awareness, opening the possibilities with the application we developed, we are certain more people, such as ourselves, will follow and embark in the process of developing and using software systems that will reduce the number of lives that are ruined because of assaults.

References

[1] Ding, Fangyu, Quansheng Ge, Dong Jiang, Jingying Fu, and Mengmeng Hao. "Understanding the dynamics of terrorism events with multiple-discipline datasets and machine learning approach." PloS one 12, no. 6 (2017): e0179057.

[2] LaFree, Gary, and Laura Dugan. "Introducing the global terrorism database." Terrorism and Political Violence 19, no. 2 (2007): 181-204.

[3] Stecklov, Guy, and Joshua R. Goldstein. "Terror attacks influence driving behavior in Israel." Proceedings of the National Academy of Sciences of the United States of America 101, no. 40 (2004): 14551-14556.

[4] Sharot T, Martorella EA, Delgado MR, Phelps EA. How personal experience modulates the neural circuitry of memories of September 11. Proceedings of the National Academy of Sciences of the United States of America, 2007

[5] Hersh ED. Long-term effect of September 11 on the political behavior of victims' families and neighbors. Proceedings of the National Academy of Sciences of the United States of America, 2013

[6] Leetaru, Kalev, and Philip A. Schrodt. "Gdelt: Global data on events, location, and tone, 1979–2012." In ISA annual convention, vol. 2, no. 4, pp. 1-49. 2013.

[7] Carley, Kathleen M., Douglas B. Fridsma, Elizabeth Casman, Alex Yahja, Neal Altman, Li-Chiou Chen, Boris Kaminsky, and Démian Nave. "BioWar: scalable agent-based model of bioattacks." IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans 36, no. 2 (2006): 252-265.

[8] Jamestown Foundation, https://jamestown.org/programs/tm/, accessed Mart 2018.

[9] LaFree, Gary. "The global terrorism database: Accomplishments and challenges." Perspectives on Terrorism 4, no. 1 (2010).

[10] An, Haewoon Kwak Jisun. "Two Tales of the World: Comparison of Widely Used World News Datasets GDELT and EventRegistry." (2016).

Adversary Model for Machine Learning

Borche Davchev
Ss. Cyril and Methodius University
Faculty of Computer Science and Engineering (FCSE)
1000 Skopje, Macedonia
RandomAdversary@gmail.com

***Abstract*— Despite the big growth in popularity, the security evaluation of products which rely on machine learning remains a challenge. This research is step towards a solution, by presenting an adversary model and showing how to model an adversary against 3 malware classifiers.**

Keywords—machine learning; threat modeling; adversarial machine learning; AI security

I. INTRODUCTION

The recent rise in popularity of machine learning lead to development and adoption of many products based on its principles. These products can be found everywhere, from video games and personal assistants to autopilots and security solutions. However, despite the big growth, the security of these products remains a challenge. Proactive approach as shown in fig. 1 can help detect and fix security issues at an early development stage.

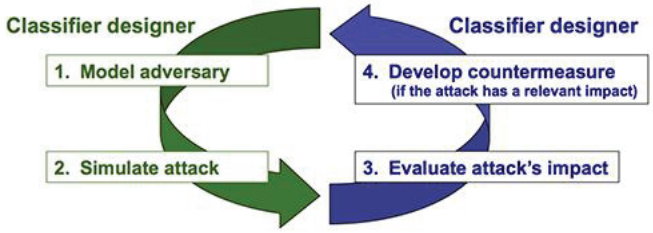


Fig. 1 - Conceptual representation of the proactive arms race. The designer tries to anticipate the adversary by simulating potential attacks, evaluating their effects, and developing countermeasures if necessary [1].

The key component in this approach is the adversary model. The lack of good adversary model limits our ability to compare different learning algorithms and leads to incomplete threat models.

II. PRIOR AND RELATED WORK

Can Machine Learning Be Secure? [2] is the research that opened the question of machine learning security and provided the taxonomy of different types of attacks on machine learning. This taxonomy is often discussed under the context of adversarial machine learning, the study of effective machine learning techniques in a presence of malicious adversary. With new attacks came the need for new adversary models and authors began extending the model from [2] or creating new adversary models closely related to their research. Two examples where this can be observed are "Security evaluation of pattern classifiers under attack" [1] where the authors extend the model from [2] and "Is feature selection secure against training data poisoning?" [3] where we have new model based on the taxonomy from [2] adapted to the context of adversarial feature selection.

III. ADVERSARY MODEL

The model can be seen as extension of [1], [2]. The key differences are focus on machine learning as process and appeal to broader audience that includes not only machine learning researchers but also anyone that might be involved in a development of product that relies on machine learning.

The adversary model consists of several attributes:

- Goal of the attack
- Knowledge of the system under attack
- Capability for manipulating data
- Attack strategy

A. Capability

1) Causative (poisoning) attacks

Under this scenario it's assumed that the adversary controls a percentage of the training and test data via specially-crafted attack samples.

2) Exploratory (evasion) attacks

In evasion attacks, the adversary manipulates the data at test time with the goal of avoiding detection.

B. Goal

1) Availability

Availability is violated when the system is compromised, causing disruption of service.

2) Integrity

Integrity is compromised when the adversary is able to do malicious activities without being detected and without compromising the functionality of the system.

3) Privacy

The adversary may want to trick the system into disclosing private details about its users.

C. Knowledge

1) Perfect knowledge

The worst-case scenario in which the adversary knows the data, feature space, model and the algorithm.

2) Limited knowledge

A more realistic scenario in which the adversary knows partial information.

a) Knowledge of the training data

The adversary may have access to the training data, a subset of the training data, or access to surrogate dataset which is collected from the same or similar source as the original dataset.

b) Knowledge of the feature representation

The adversary may know which subset of features is used.

c) Knowledge of the feature selection algorithm

It's possible for an adversary to know which feature selection algorithm is used.

d) Knowledge of the learning algorithm

The adversary may know the learning algorithm and its parameters.

e) Black box approach

In this scenario the adversary can submit samples to the model and observe the response.

D. Attack strategy

The attack strategy describes the interaction between the adversary and the system. The proper format of this part, depends on the target audience. It can be simple explanation of the steps and the risk involved, maybe followed by a research paper. If required, formal models can be used.

IV. CONNECTING THE MODEL TO THE REAL WORLD

Case study: PDF Malware detection

A. Evading known malware detectors

In this scenario we would like to evade detection from 2 malware detectors which are specialized for detecting PDF malware, Hidost [5] and PDFrate [6], [7]. PDFrate focuses on metadata and the structure of the document. Hidost builds model based on the file structure and its content. Both classifiers claim resistance to evasion and mimicry attacks.

The adversary model can be created by answering the following questions:

1. Can you add, modify or delete samples from the training dataset?
2. How well do you know the system under attack?
3. What do you know about the feature space?
4. What do you know about the learning algorithm?
5. What is your goal?
6. What is your attack strategy?

In this case we don't want to modify the training set. Our attack strategy will fall into the evasion attacks category. The exact features are unknown but we know that the focus is on the file structure. Both detectors can be used with custom classifiers, however, based on experiments random forest produced the best results. For the attack strategy, we want to make changes to the file structure while preserving the malicious nature of the file. EvadeML [8] can be used to achieve this. EvadeML is an evolutionary framework based on genetic programming for automatically finding variants that evade detection by machine learning-based malware classifiers. For this example, a simplified version of the attack strategy presented in [9] will be used. The simplified attack strategy is:

1. Grab malicious pdf file. This could be single pdf for targeted attacks or randomly selected file for mass campaigns.
2. Do a random mutation. To do this, generate the abstract syntax tree for the selected sample, then pick random object in the tree and do one of the following actions: add another object below the select item, delete the object or replace it with another object.
3. Check if file is classified as malicious
4. Check if file is malicious. Typically, this involves running the sample in a sandbox and looking at the list of network and api calls to verify the correct malicious behavior.
5. Observe the results from step 3 and 4. If the file is no longer malicious, go to step 1. If the file is malicious and classified as malicious go to step 2.
6. Repeat for all samples.

The result is set of files which hide their malicious nature from the selected classifiers. These files are called adversarial examples. Adversarial examples are inputs formed by applying small but intentionally worst-case perturbations to examples from the dataset, such that the perturbed input results in the model outputting an incorrect answer with high confidence [10].

B. Blackbox attacks against Gmail

For this scenario we have bunch of malicious files we want to send as email attachments without being detected by Gmail. The adversary model can be a bit of a challenge. The details for the malware detection engine used by Gmail are not available to the public. Given the popularity of Gmail, poisoning attacks are unlikely to succeed. This leaves us with evasion type of attack. We have no information about the feature space or the learning algorithm. Without information about the learning algorithms, the attack strategy revolves around finding evasion patterns and using adversarial examples from other models.

1) Part 1 - Exploiting the transferability property of adversarial examples

Adversarial examples often transfer between models, one input can be used to trick more than one model. An attacker may train their own substitute model, craft adversarial examples against the substitute, and transfer them to a victim model, with very little information about the victim [11], [12]. The authors of the paper which was used in the previous example [9] submitted the adversarial examples to Gmail and found out that 0.6% of the examples can be transferred to Gmail [13]. The amount of samples may seem low, but this is enough to give us clues into possible blind spots in the model

which can be exploited.

2) Part 2 - Finding evasion patterns

Analyzing the different paths that lead to the creation of adversarial examples that work on Gmail can give insight into the weak spots of the target classifier. In this case it was revealed that making simple change such as declaring new JavaScript variable inside the pdf file is enough to evade detection. With this new insight they were able to raise the success rate to 47.1% [13].

C. Alternative attack strategy

Changing the classifier from Hidost and PDFrate to the web service offered by Gmail will eventually lead to creation of adversarial example. The problem with this approach is the time and number of api calls needed. Another approach is to follow the Seed-Explore-Exploit framework [14] where you make api calls and observe the response. Based on the responses, a surrogate model can be made and used as replacement for the web service. Adversarial examples generated based on the surrogate model are likely to transfer into the real one.

V. CONCLUSION

With the growing adoption of machine learning and artificial intelligence in general, we need tools to develop models and security evaluation frameworks accessible to anyone. The first step is making model of your adversary. This can be done by answering the following questions:

1. Can you add, modify or delete samples from the training dataset?
2. How well do you know the system under attack.
   - What do you know about the feature space?
   - What do you know about the learning algorithm?
3. What is your goal?

These answers combined with current research as different attack strategies enable us to compare the security of different learning algorithms and build better defense strategies.

REFERENCES

[1] Biggio, Battista & Fumera, Giorgio & Roli, Fabio. (2013). Security Evaluation of Pattern Classifiers Under Attack. IEEE Transactions on Knowledge and Data Engineering. 99. 1. 10.1109/TKDE.2013.57

[2] Marco Barreno, Blaine Nelson, Russell Sears, Anthony D. Joseph, and J. D. Tygar. 2006. Can machine learning be secure?. In Proceedings of the 2006 ACM Symposium on Information, computer and communications security (ASIACCS '06). ACM, New York, NY, USA, 16-25.

[3] Huang Xiao, Battista Biggio, Gavin Brown, Giorgio Fumera, Claudia Eckert, and Fabio Roli. 2015. Is feature selection secure against training data poisoning?. In Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37 (ICML'15), Francis Bach and David Blei (Eds.), Vol. 37. JMLR.org 1689-1698.

[4] Wei Liu and Sanjay Chawla. 2009. A Game Theoretical Model for Adversarial Learning. In Proceedings of the 2009 IEEE International Conference on Data Mining Workshops (ICDMW '09). IEEE Computer Society, Washington, DC, USA, 25-30.

[5] Nedim Šrndić and Laskov Pavel. (2016). Hidost: a static machinelearning-based detector of malicious files. EURASIP Journal on Information Security. 2016. . 10.1186/s13635-016-0045-0.

[6] Smutz, Charles & Stavrou, Angelos. (2012). Malicious PDF detection using metadata and structural features. ACM International Conference Proceeding Series. 239-248. 10.1145/2420950.2420987.

[7] PDFrate. https://csmutz.com/pdfrate/. Accessed 3 September 2017

[8] http://evademl.org/gpevasion/ Accessed 3 September 2017.

[9] Weilin Xu, Yanjun Qi, and David Evans. Automatically Evading Classifiers A Case Study on PDF Malware Classifiers. Network and Distributed Systems Symposium 2016, 21-24 February 2016, San Diego, California.

[10] Goodfellow, Ian & Shlens, Jonathon & Szegedy, Christian. (2014). Explaining and Harnessing Adversarial Examples. arXiv 1412.6572.

[11] Papernot, Nicolas & McDaniel, Patrick & Goodfellow, Ian. (2016). Transferability in Machine Learning: from Phenomena to Black-Box Attacks using Adversarial Samples.

[12] Liu, Yanpei & Chen, Xinyun & Liu, Chang & Song, Dawn. (2016). Delving into Transferable Adversarial Examples and Black-box Attacks.

[13] David Evans. Classifiers under Attack. USENIX Enigma 2017

[14] Tegjyot Singh Sethi, Mehmed Kantardzic. Data Driven Exploratory Attacks on Black Box Classifiers in Adversarial Domains

An overview of the lightweight cryptography in the new concept IoT

Dime Boshkovski
Faculty of Computer Science and Engineering
Ss. Cyril and Methodius University
Skopje, Macedonia
boshkovski.dime@students.finki.ukim.mk

***Abstract*— The Internet of Things (IoT) being a promising technology of the future is expected to connect billions of sensors and small devices. In this paper is given an introduction of cryptography of IoT and what is Lightweight cryptography. After that is discuss about the division of symmetric key cryptography, usage of symmetric key cryptography in IoT and the disadvantages of public key cryptography. Also one part is about similarities and differences in construction and performance of two block ciphers from which one with lightweight properties, Present algorithm which is used in IoT and on the end is given the importance of cryptography in IoT.**

***Keywords*— security, encryption, lightweight.**

I. INTRODUCTION

Today the grown of small object that are connected on the internet rapidly increases. This devices are constrained in their performacese because the biggest part of them are with small dimensions. The data that send this devices is crucial for the customers but from the other side also from big importance is the volume of the devices. The balance is some kind of algorithm that will sutisfies the security needs and from other side will not complicate normal functioning of that device. This environment of small devices that must sutisfies some level of security is called "Internet of Things (IoT)" [1].

The biggest reason of rapidly growing of this environment IoT is growing of Internet usage. The price of Internet every day is lower and for small amount of money you can get faster internet and bigger amount of traffic but from other side the traffic that goes throught internet is less secure than the private network [2]. Every day the number of gadgets and the sensors that are connect on the internet rises and this provides suitable ground for the grown of IoT [3].

The rising of cryptoanalysis produces some new kind of attack that and the cryptography must to solve this attack. This solution produces more complex kinds of crypting, matrix that are used for permutation, bigger keys than usual, bigger sboxes that are used etc. To be suttisfied this needs the chips that are used needs bigger memory and bigger computation performance. The balance that help is the technique so called "Lightweight cryptography (LWC)". With this kind of tecnique are produced algortihm that can be used in devices with constrained resources. Lightweight cryptography mades balance between security from one side and energy consumption from other side [1].

II. SYMMETRIC KEY CRYPTOGRAPHY

A. Block Cyphers

The two algorithms that are used in the practical system are Clefia and Present. This algorithm are part of the standard ISO/IEC 29192 "Lightweight Cryptography" [1]. The reason of using this algorithm in devices with small performance is small complexity and small needs for memory from one side and satisfing cryptographical security from other side. For example the AES (Advanced Encryption Standard) and DES (Data Encryption Standard) both block ciphers that are with high level of security especially AES, but they need more complex chips and bigger memory because they have bigger sbox, more complicated permutation matrices, bigger number of cycles etc [4].

1) CLEFIA

This algorith is block cipher. The size of blocks that are crypted in every cycle is 128 bits. The lengths of the keys that are used can be of: 128, 192 and 256 bits. This algorithm achieves enough immunity against known attacks and flexibility for efficient implemetation in both hardware and software by adopting several new design tecniques.

a) Design strategy of CLEFIA

The three fundamental characteristics for every lightweight chiper are: security, speed and cost. Clefia block cipher has all of them.

Clefia is structured from Feistel Type-2 functions. Feistel functions are important because they can be used for crypting and decrypting just with changing the order of keys. Also the other advantage of Type-2 Feistel Functions from author [5] are:

- F-functions are smaller than that of the traditional Feistel structure
- Plural F-functions can be processed simultaneously

- Tends to require more rounds than the traditional Feistel structure [5]

With smaller functions is possible to made smaller chips and this is from big importance. The second feature allow to the processors parallel execution of the task and the consequence of this is shorter time. From other side the third characteristis is disadvatege of the Type-2 Feistel function but when the sum of advantage and disadvantage is made the characteristics are positive. Also are used some tecniques that reduces the number of cycles and the consequnce of the this are better performance.

With the F-functions used in the Clefia algorithm is included diffusion switching mechanism. This mechanisam uses matrices that gives stronger immunity against different kind of attacks. Also with this tecnique the number of rounds is reduces but the security characteristics dont lose to much.

The algebraic immunity is on the needed level because of using two different S-boxes. In DES algorithm are used 8 sboxes but Clefia satisfies the needs just with two.

In Table I [5] are given summarized results for effiecent implementation. One important think is that Clefia can be implemented efficiently in hardware and in software. Some of the advantages of Clefia are mentioned in the nexts paragraphs.

TABLE I. DESIGN ASPECTS FOR EFFICIENT IMPLEMENTATIONS

Design technique	***Characteristics***
Generalized Feistel Network	Small size F-functions (32-bit in/out) No need for the inverse F-functions
SP-type F-function	Enabling fast table implementation in software
DSM	Reducing the numbers of rounds
S-boxes	Very small footprint of S_0 and S_1 in hardware
Matrices	Using elements with low hamming weights only
Key Schedule	Same structure with the data processing part Only a 128-bit register is required for CLEFIA with 128-bit keys Small footprint of DoubleSwap function

b) Advantages of CLEFIA

Clefia is enough secure and can counterattack against all known cryptanalyses. With research is confirmed that this algoritam is secure, and has no weakness from any kind of attack. One of the reason for this immunity is using two different types of S-boxes allocated in F-functions. With this s-box the algorithm is immune on algebraic attacks including the XSL attack (method for solving equation with variables and from solution recovering the key).

From the research made by the authors [5] the results that are achieved from software version of Clefia on 2.4 GHz AMD Athlon 64 are 13 cycles/byte, 1,48 Gbps. From other side the hardware implemetation of Clefia is very small, occupying less than 5000 gates by 0.09 μm. Clefia is in the group of the fastest block ciphers while the hardware implementation is on the faster than other blockciphers [5].

2) PRESENT

Present algorithm is in the group of block ciphers. He i used for security in application that are used in tag based environment. The block that is crypted is with size of 64 bits and can be used keys with length of 80 and 128 bits. Number of rounds for crypting is 31 and for crypting is used SP-network.

Every round of crypting is consisted of one xor operation with round key. After that is included one linear bitwise permutation and finally there is substitution layer, sboxs. Sboxes are matrix that can accept on input four bits and gives four bits on the output. In Table II the authors [6] shown up comparision of lightweight block ciphers.

TABLE II. COMPARISION OF LIGHTWEIGHT BLOCK CIPHER IMPLEMENTATIONS

	Key size	Block size	Cycles per block	Throught put 100 kHz	logic process	Area rel.
Present	80	64	32	200	0,18	1570
DES	56	64	144	44,4	0,18	2309
AES	128	128	1032	12,4	0,35	3400

From the results shown in the table is possible to conclude that the number of cycles per block for present is smallest that is expected because he has lightweight characteristics.

B. Stream Ciphers

From researching of stream ciphers the Enocoro algorithm has the lightweight properties and also this algorithm is mentioned in ISO/IEC 29192 “Lightweight cryptography” standard. This algorithm is prepared for use in the practical systems. Some of the charactheristics of Encoro are shown below [7]:

1) Enocoro

Like other stream ciphers this algorithm at the begin must have Initialization vector IV. Choosing the IV is from special importance because someone that make cryptoanalyses can obtain using the same key and several different IV the result be several different keystreams [8]. The operations thate are made in the algorithm are internal state and functions, one for updating the current state and one for producing output from current state. These opearion are repeating for every new result [9].

C. Hash function

Today there are many hash algorithms that but not all of them has lightweight properties. The most atractive of hash algoritms is SHA-3. This algorithm is expected to be hash function that can be used in every environment. But after testing this algorithm he dont give expected results. Research of the lightweight hash function is at the begin and greater result are expected in future. Maybe the best results will be achieved if hash function are constructed with base of block ciphers [1].

Algorithms that are the best lightweight properties and also are mentioned in the ISO/IEC 29192 “Lightweight

cryptography" are: Photon, Spongent and Lesamnta-LW. This hash functuin are prepared for use in the practical systems [10]. In the next paragraphs will be shown some of the characteristics of this algorithms first Photon, second Spongent and final Lesamnta.

1) Lightweight hash-function optimized for hardware implementation

a) Photon-function family

The Photon family of functions is the smallest hash function that is desigened, and also he has excelent area/throughput trade offs. To design this function for base is used AES that is one of the most used and most secure block ciphers.

The photon function gives outputs with different lengths. The lengths can be from 64 to 256 bits. The internal state depends from the output and can be 5 dinstinct values: 100, 144, 196, 256 and 288 bits. Because of different needs of application there are different versions of Photon one for each internal state. The Photon 80/20/16 is used for application were the results from the hash from 64 bits considered to be sufficient. While the version Photon 256/32/32 is used for collision of 128 bits [11].

b) Spongent – function family

There are five versionof spongent hash function. The output in this functions can be from: 88, 128, 160, 224 and 256 bits. The first with out from 88 bits is used in scenarios where is not need high security and where the performace of the devices are on low level (RFID). The versions with output of 128 and 256 bits are used where is needed middle collision security and the devices are with constrained performance. The last two versions with output of 224 and 256 bits are used where is needed high collision security. This two are on the same level with the SHA-2 and SHA-3 hash functions.

In Figure 1 the authors [12] gives an overview of Sponget hash function ie how the result is generated. The result is

Spongent relies on a sponge construction – a simple iterated design that takes a variable-length input and can produce an output of an arbitrary length based on a permutation π_b operating on a state of a fixed number b of bits. The size of the internal state b = r + c >= n is called width, where r is the rate and c the capacity. The sponge construction proceeds in three phases:

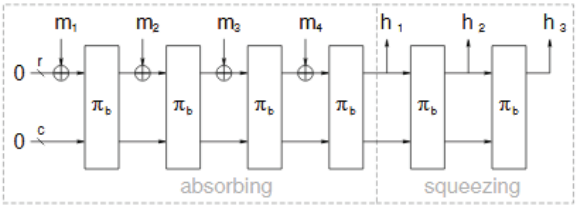


Fig. 1. Sponge construction based on a b-bit permutation π_b with capacity c bits and rate r bits. m_i are r-bit message blocks. h_i are parts of the hash value.

This algorithm has three phases [12]:

Initialization phase: First is made initializaion of the bits. If the number of bits is not diveded with r at the top of bits is added one bit "1" followed with "0" bits while the length of all bits is not diveded with r. After that the bits are diveded in groups of r bits.

Absorbing phase: the block are xored with first r bits of the state and after that are interleaved with the permutation π_b.

Squeezing phase: the first r bits of the state are returned as output, interleaved with applications of the permutation π_b, until n bits are returned [12].

2) Lightweight hash-function optimized for software implementation

a) Lesamnta-LW

This hash function is optimized for software implementation. This function can be used from the smallest devices to the high-end servers. The reason why this algoritham can be used in all of the devices is in that, that is made for 8 bits CPUs that are integrated in the smallest devices. This algoritham gives better result between speed and cost than SHA-256 on an 8-bit CPU [13].

III. Public Key Cryptography

While lightweight public key primitives are in demand for key management protocols in smart objects networks, the required resource for public key primitives is much larger than that of symmetric key primitives [1]. At this time, there are no promising primitives that meet enough security and lightweight properties compared with the conventional primitives such as RSA (Rivest Shamir Adleman) and ECC (Elliptic Curve Cryptography). Some public key primitives (e.g. ECC) can be implemented with relatively small footprint, but they cannot execute with in a reasonable time [14].

IV. Commparision of Performance of Block Ciphers DES and Present

Those algorithms are choose because they are in the same group of cryptography protocols (block cyphers) but just one of them is in the group of lightweight ciphers (Present). Also those algorithms has the same size of block that is crypted at one time (8 bytes = 64 bits). The other similaity is both uses sboxes and also both uses permutations with which the encrypted data have better security against linear and differential attacks.

They also have differences DES algorithm uses Feistel function while Present uses Substitution-Permutation Network. In DES algorithm are used eight sboxes with size each of tham of 64 values, each value of 6 bits while Present uses one sbox with size of 16 values of 4 bits. From this we can conclude just from size of sbox DES is more complex ie 3072 bits are needed for storage of all of the sboxes while for Present are needed 64 bits. The consequence of this is also more complex logical operation, more complex processor, more complex memory etc.

The tests are made on laptop which is not part of IoT and have the characteristics: processor Intel(R) Core(TM) i5-2430M CPU 2,4 GHz, RAM 8 GB and 64 bits operating system. The code for both algorithms DES and Present is written in Intellij IDEA 2016.2.5.

TABLE III. COMPARISION OF DES AND PRESENT BLOCK CIPHERS

Block Size / *Algorithm*	*2,5 KB*	*5KB*	*10KB*
DES	125ms	182ms	368ms
PRESENT	86ms	146ms	251ms

Tests are made on blocks with size of 2,5KB, 5KB and 10KB from 8 to 10 time and the middle time in milliseconds are shown in Table III. From the results is possible to conclude that the time for crypting the same block with Present and DES are differnet i.e. the time for Present is shorter that was expected.

V. WHY IS LIGHTWEIGHT CRYPTOGRAPHY REQUIRED FOR INTERNET OF THINGS

The propose to adopt new advancing technology, "Lightweight Cryptography", in the IoT. The most important two reasons are:

1. Efficiency of end-to-end communication

Two aspects are the most importart. The first one is the security of communication to be on the needed level and second one is the energy consumption to be on low level because the biggest number of devices that are used in IoT the source of energy is battery. For this reason is used symmetric cryptography ie application of the lightweight symmetric key algorithm allows lower energy consumption for end devices.

2. Applicability to lower resource devices

The footprint of the lightweight cryptographic primitives is smaller than the conventional cryptographic ones. The lightweight cryptographic primitives would open possibilities of more network connections with lower resource devices.

VI. CONCLUSION

The cryptography in Internet of Things is at the begin of the development because in the future the IoT will continue rapidly to increase and the number of small devices that will be connected on the Internet will need more efficient and effective cryptographic tecniques. There are two important thinks that must be considered when are designed cryptography algorithms: the first one is algorithm to satisfy the security against every kind of attack and the second think is usage of power ie crypting of data to be an activite that will not spend the energy resources.

REFERENCES

[1] Masanoby Katagi and Shiho Moriai, "Lightweight cryptography for the Internet of Things". Sony Corporation, 10. 2008. pp. 1-4.

[2] Muhammad Usman, Irfan Ahmed, M. Imran Aslam, Shujaat Khan and Usman Ali Shah, "A lightweight encryption algorithm for secure Internet of Things", International Journal of Advanced Computer Science and Applications, Vol. 8, No. 1, 2017, pp. 1-10.

[3] Uday Kumar, Tuhin Borgohain and Sugata Sanual, "Comparative analysis of cryptography library in IoT". pp. 1-5.

[4] "Information technology – security techniques – lightweight cryptography – part 2: block ciphers", p. 41, pp. 1-41, January 2012.

[5] "The 128-bit blockcipher Clefia", Sony Corporation, pp. 1-52, January 2010.

[6] A. Bogdanov, L. R. Knudsen, G. Leander, C. Paar, A. Poschmann, M. J. B. Robshaw, Y. Seurin and Vikkelsoe, "Present: an ultra-lightweight block cipher", pp. 1-18, September 2007.

[7] "Information technology – security techniques – lightweight cryptography – part 3: stream ciphers", p. 26, pp. 1-26, October 2012.

[8] "Pseudorandom number generator Enocoro", Hitachi, Ltd., pp. 1-14, February 2010.

[9] Martin Hell and Thomas Johansson, "Security evaluation of stream cipher Enocoro-128v2", pp. 1-31.

[10] "Information technology – security techniques – lightweight cryptography – part 5: hash functions", p. 26, pp. 1-26, August 2016.

[11] Jian Guo, Thomas Peyrin and Axel Poschmann, "The Photon family of lightweight hash functions", pp. 1-22.

[12] Andrey Bogdanov, Miroslav Knezevic, Gregor Leander, Deniz Toz, Kerem Varici and Ingrid Verbauwhede, "Spongent: the design space of lightweight cryptographic hashing", pp. 1-23.

[13] Shoichi Hirose, Kota Ideguchi and Hidenori Kuwakado, "An AES based 256-bit hash function for lightweight application: Lesamnta-LW", iecie trans. fundamentals, vol.e95-A, no.1, pp.89-99, january 2012.

[14] "Information technology – security techniques – lightweight cryptography – part 4: Mechanisms using asymmetric techniques", p. 1-26, june 2013.

ISBN 978-608-4699-08-8