Volume 38 Number 1 March 2014

ISSN 0350-5596

Informatica

An International Journal of Computing and Informatics

Special Issue:

Advances in Semantic Information Retrieval

Guest Editors:

Vitaly Klyuev

Maxim Mozgovoy

1977

Editorial Boards, Publishing Council

Informatica is a journal primarily covering intelligent systems in the European computer science, informatics and cognitive community; scientific and educational as well as technical, commercial and industrial. Its basic aim is to enhance communications between different European structures on the basis of equal rights and international refereeing. It publishes scientific papers accepted by at least two referees outside the author's country. In addition, it contains information about conferences, opinions, critical examinations of existing publications and news. Finally, major practical achievements and innovations in the computer and information industry are presented through commercial publications as well as through independent evaluations.

Editing and refereeing are distributed. Each editor from the Editorial Board can conduct the refereeing process by appointing two new referees or referees from the Board of Referees or Editorial Board. Referees should not be from the author's country. If new referees are appointed, their names will appear in the list of referees. Each paper bears the name of the editor who appointed the referees. Each editor can propose new members for the Editorial Board or referees. Editors and referees inactive for a longer period can be automatically replaced. Changes in the Editorial Board are confirmed by the Executive Editors.

The coordination necessary is made through the Executive Editors who examine the reviews, sort the accepted articles and maintain appropriate international distribution. The Executive Board is appointed by the Society Informatika. Informatica is partially supported by the Slovenian Ministry of Higher Education, Science and Technology.

Each author is guaranteed to receive the reviews of his article. When accepted, publication in Informatica is guaranteed in less than one year after the Executive Editors receive the corrected version of the article.

Editors's Introduction to the Special Issue on "Advances in Semantic Information Retrieval"

Semantic technologies and information retrieval hold a firm place among topical research directions of modern computer science. Advances in this field define the ways we use computers in the age of Internet and mobile technologies. We see high level of interest to semantics and information retrieval at our annual International Workshop on Advances in Semantic Information Retrieval (ASIR) workshop that continues to attract researchers from different parts of the world.

In this special issue of Informatica journal we introduce four revised and extended papers, presented at the workshop.

The first paper entitled *Automatic Detection of Antisocial Behaviour in Texts* by Myriam Munezero, Maxim Mozgovoy, Tuomo Kakkonen, Vitaly Klyuev, and Erkki Sutinen is devoted to detection of antisocial behavior (ASB) manifestations in written documents. The authors search for linguistic features that pertain to ASB in order to use those features for the automatic identification of ASB in text. They used a collected ASB text corpus as a machine learning resource and approach the detection of ASB in text as a binary classification problem. The results from the experiments show that by exploiting the emotional information together with Bag-of-Words (BoW) the accuracy of over 90% in the classification of ASB in text is reached. These findings will have positive implications in the early detection of potentially harmful behavior.

The next paper entitled *Leveraging User Experience through Input Style Transformation to Improve Access to Music Search Services* by Marina Purgina, Andrey Kuznetsov, and Evgeny Pyshkin addresses the problems of music searching and main tasks the developers face in the domain of music information retrieval. The authors introduce the architecture of the software and the data model for integrated access to existing music search web services. The authors illustrate their approach by developing a mobile accessed software prototype that allows the users of Android-running touch screen devices to access several music search engines including Musipedia, Music Ngram Viewer, and FolkTuneFinder. The designed application supports various styles of music input query. The authors pay special attention to input style transformation aimed to fit well the requirements of the supported search services.

The third paper entitled *User Annotations as a Context for Related Document Search on the Web and Digital Libraries* by Jakub Ševcech, Róbert Móro, Michal Holub, and Mária Bieliková proposes a method for query construction enabling search for other documents related to the currently studied one using not only the document's content, but also user created annotations as indicators of user's interests. In the proposed approach, annotations are used to activate nodes in a graph created from the document's content employing spreading activation algorithm. The authors evaluate the proposed method in Annota — a service for bookmarking and collaborative annotation of Web pages and PDF documents displayed in a web browser. Along with its main purpose, Annota is designed to support scenarios useful for a novice researcher working together with his or her mentor. Based on Annota usage data the authors also analyzed properties of various types of annotations. Discovered annotation properties served as a basis for simulation performed to determine optimal parameters of the query construction. The authors compared the proposed method to the commonly used tf-idf based method that was outperformed with the method introduced in the paper. Therefore, annotations proved to be a viable source of information for user's i nterest detection.

The fourth paper entitled *SOAROAD: an Ontology of Architectural Decisions Supporting Assessment of Service Oriented Architectures* by Piotr Szwed, Paweł Skrzynski, Grzegorz Rogus, and Jan Werewka describes SOAROAD (SOA Related Ontology of Architectural Decisions) developed to support the evaluation of architectures of information systems based on the Service-Oriented Architecture (SOA) approach. The main goal of the ontology is to provide constructs for documenting architecture. However, it is designed to support future reasoning about architecture quality and fulfilling the nonfunctional system requirements such as scalability, ease of maintenance, reuse of software components, etc. Another important reason is building a common knowledgebase. When building the ontology, the Architecture Tradeoff Analysis Method (ATAM) was adopted which was chosen as a reference methodology of architecture evaluation.

As ASIR chairs, we are strongly committed to our basic aim: to create an atmosphere of friendship and cooperation for everyone, interested in computational linguistics and information retrieval. The workshop is firmly established as an event within Federated conference on computer science and information systems (FedCSIS), annually organized by the System Research Institute of the Polish Academy of Sciences and the Polish Information Processing Society, and sponsored by the IEEE.

In its turn, ASIR is supported by the University of Aizu (Japan), known as Japan's first university solely dedicated to computer science engineering. The University of Aizu is a major center of international education and the home of several conferences, sponsored by the ACM and the IEEE.

We would wish to acknowledge selfless efforts of our committee members and FedCSIS conference organizers, who ensured high quality of publications and flawless arrangement of the forum. We would like to specially mention professors Marcin Paprzycki, Maria Ganzha, and Halina Kwasnicka, responsible for FedCSIS.

We had a great support from our international team of reviewers, consisting of:

Grzegorz J. Nalepa, Ryszard Tadeusiewicz (AGH University of Science and Technology, Poland); Eloisa Vargiu (Barcelona Digital Technology Centre, Spain); Larisa Soldatova (Brunel University, United Kingdom); Shih-Hung Wu (Chaoyang University of Technology, Taiwan); Cristian Lai (CRS4, Italy); Ahsan Morshed (CSIRO ICT Centre, Commonwealth Scientific and Industrial Research Organisation, Australia); Roman Shtykh (CyberAgent Inc., Japan); Krzysztof Goczyła (Gdansk University of Technology, Poland); Yannis Haralambous (Institut Telecom - Telecom Bretagne, France); Katarzyna Budzynska (Institute of Philosophy and Sociology of the Polish Academy of Sciences, Poland); Piotr Kulicki, Robert Trypuz (John Paul II Catholic University of Lublin, Poland); Stefano Borgo (Laboratory for Applied Ontology, Italy); Janusz Kaczmarek (Lódz University, Poland); Simone Ludwig (North Dakota State University, United States); Mari Carmen Suárez de Figueroa Baonza (Ontology Engineering Group, Scool of Computer Science at Universidad Politécnica de Madrid, Spain); Raúl Palma (Poznan Supercomputing and Networking Center, Poland); Jolanta Cybulka, Jacek Martinek, Agnieszka Ławrynowicz (Poznan University of Technology, Poland); Evgeny Pyshkin (Saint Petersburg State Polytechnical University, Russia); Vladimir Dobrynin (Saint Petersburg State University, Russia); Haofen Wang (Shanghai Jiao Tong University, China); Slawomir Zadrozny (Systems Research Institute, Poland); Massimiliano Carrara (Universita di Padova, Italy); Nikolay Mirenkov, Alexander Vazhenin (University of Aizu, Japan); Marek Reformat (University of Alberta, Canada); Tuomo Kakkonen (University of Eastern Finland, Finland); Miroslav Vacura (University of Economics, Czech Republic); Sabina Leonelli (University of Exeter, United Kingdom); Wladyslaw Homenda (Warsaw University of Technology, Poland); Qun Jin (Waseda University, Japan); Maciej Piasecki (Wroclaw University of Technology, Poland).

We also thank Professor Matjaz Gams (managing editor of Informatica), who supported the publication of this special issue.

In 2014, we are organizing ASIR workshop within FedCSIS in Warsaw, Poland. We will continue to maintain high standards of quality and organization, set by the first workshops. We welcome all the researchers, interested in semantics and information retrieval, to join our event.

Vitaly Klyuev
Maxim Mozgovoy

Editors of the special issue

Automatic Detection of Antisocial Behaviour in Texts

Myriam Munezero, Calkin Suero Montero, Tuomo Kakkonen and Erkki Sutinen
School of Computing, University of Eastern Finland
P.O.Box 111, FI-80101, Joensuu, Finland
E-mail: {firstname.lastname}@uef.fi

Maxim Mozgovoy and Vitaly Klyuev
The University of Aizu, Tsuruga, Ikki-machi
Aizu-Wakamatsu, Fukushima, 965-8580 Japan
E-mail: {mozgovoy, vkluev}@u-aizu.ac.jp

Keywords: antisocial behaviour, machine learning, emotion detection, language analysis

Received: November 23, 2013

A considerable amount of effort has been made to reduce the physical manifestation of antisocial behaviour (ASB) in communities. However, the key to the early detection of ASB is, in many cases, in observing its manifestations in written language, which has not been studied in detail. In this work, we search for linguistic features that pertain to ASB in order to use those features for the automatic identification of ASB in texts. We use an ASB text corpus we have collected as a machine learning resource and approach the detection of ASB in texts as a binary classification problem where discriminating features are taken from the linguistic representation of texts in the form bag-of-words and ontology-based emotion descriptors. Results from preliminary experiments show that by exploiting the emotional information together with Bag-of-Words (BoW) over 90% accuracy in the classification of ASB in texts is reached. Our findings have positive implications in the early detection of potentially harmful behaviour.

Povzetek: Pri analizi asocialnih besedil v omrežjih dosežejo napredek v kvaliteti prepoznavanja z uporabo ontologij čustev.

1 Introduction

Text mining allows for the automatic assessment of linguistic features in texts. Based on the analysis results, it is possible to analyse, for instance the topics that the texts deal with, as well as linguistic styles used in the texts. Language syntax and semantics are tools that are used to express thoughts, opinions, beliefs and emotions through words. The words used can reveal important aspects of someone's social and psychological worlds [33]. Of interest to us, are words and linguistic features that express thoughts or feelings of harming another member of the community. In this paper, we analyse and discover the linguistic features that pertain to ASB based on *machine learning* (ML) and the *antisocial behaviour* (ASB) corpus we introduced in Munezero et al. [27]. Identifying these features will allow us to detect new instances of ASB

ASB is broadly defined as any unconsidered action taken against individuals or groups of individuals that may cause harm or distress to society [5]. Often individuals involved in ASB have disclosed in advance their feelings and plans through oral or written language [30]. The Internet has been used as the outlet for the expression of such emotional states and / or plans of violent acts through the use of blogs or video sites [9]. Moreover, online communication is often used as a way of shouting out people's intentions before engaging in their acts of violence [21, 2].

The growth of the volume of harmful material on the Web has resulted in increased research for its automatic detection [8]. Being able to automatically detect negative material is beneficial, for instance, to managers of websites that allow users to post content or as part of an early warning system to authorities on possible threats to public safety. The automatic detection of ASB could also give rise to self-awareness systems for the individuals that are expressing thoughts or emotions related to ASB.

This paper investigates the linguistic features used in texts that relate to ASB. By employing ML algorithms we explore the linguistic features that can be used to reliably classify texts containing ASB. For our initial experiments, we explore the impact that BoW and emotions as linguistic features have on the classification of ASB.

This paper is based on: M. Munezero, M. Mozgovoy, T. Kakkonen, V. Klyuev and E. Sutinen, *Antisocial Behavior Corpus for Harmful Language Detection*, published in the Proceedings of the 3rd International Workshop on Advances in Semantic Information Retrieval (part of the FedCSIS' 2013 conference).

2 Related work

Much of the research work on ASB has been performed in the realm of social sciences and psychiatry. There have also been efforts towards detecting and preventing physical manifestations of ASB (such as violence) in communities (e.g. the Home Office in United Kingdom[1]). As such, this problem has not been particularly tackled from the perspective of computational linguistic analysis for early detection and intervention.

As no previous general models for detecting ASB from text exist, we provide an overview of the work done in the context of detecting cyberbullying, terrorism and criminal behaviour which all can be considered as specific forms of ASB.

Perhaps the most notable related work has been carried out in a research project entitled "Intelligent information system supporting observation, searching and detection for security of citizens in urban environment" [41]. The project aimed at automatic detection of terrorist threats and recognition of serious criminal behaviour or violence based on multi-media content. Within the context of INDECT, criminal behaviour is defined as "behaviour related to terrorist acts, serious criminal activities or criminal activities in the Internet".

Our work differs from the one done in the INDECT project in the focus of the research. While INDECT aims at using the analysis of images, video, and text, our focus is on the analysis of text data.

In their cyberbullying study, Dinakar et al., [12] made use of YouTube comments that involved sensitive topics related to race and culture, sexuality and intelligence. Moreover, Yin et al. [45] made use of online forums for detecting online harassment. The cyber-pedophilia research by Bogdanova et al. [3] made use of perverted online journal texts based on which to learn models to discriminated pedophiles from non-pedophiles.

While the corpora used in the studies reported above contain some forms of negative behaviours, their focus is more than ours. We make use of a broader ASB corpus that contains text related to ASB ranging from suicide notes to terrorism and online threats.

2.1 Language expressivity in ASB

Fitzgerald [15] describes the language of ASB as being "deeply value laden, implying purposeful negative action and or behaviour harmful to others". In addition, some researchers have suggested that certain emotions are closely associated with ASB. Some of these emotions include anger, frustration, arrogance, shame, anxiety, depression, sadness, low levels of fear, and lack of guilt [7]. Based on these descriptions, it is reasonable to expect some distinguishing linguistic features in the ASB corpora that may include the use of words that are deemed threatening, harmful or related to violence and emotions that are perceived as overly negative.

2.2 Detecting emotions in texts

Emotions have long been investigated in several studies ranging from social psychology to computational linguistics [19]. Lists of primary or "basic" emotions have been put forward in the psychological field prominently by Frijda [16], Ekman [13] and Plutchik [34] among others. The basic emotion categories used in these lists include: anger, sadness, joy, love, surprise, happiness, fear, and disgust (see [28] [37]), for a detailed compilation of primary emotion lists). Within the Natural Language Processing (NLP) research community, more often than not researchers use Ekman [13] six basic emotion categories: anger, disgust, fear, happiness, surprise and sadness [1] [39].

Performing emotion analysis on various types of text can help us understand and measure the emotions expressed in them. Broadly speaking, two main methods exist for the analysis of emotions within the NLP community: word lists-based and ML-based. Word list based methods use lexical resources such as lists of emotion-bearing words, lexicons or affective dictionaries [29] [14], and databases of commonsense knowledge [20], The *General Inquirer* (GI) [38], the *Affective Norms for English Words* (ANEW) [4], the *WordNet-Affect* [40] [42], and more recently the *NRC word-emotion association lexicon* [25] [24], are all well-known lexical resources.

Whereas ML-based methods cast the problem as a multi-class classification problem, for instance, the automatic emotion classification of news headlines into emotion categories [10]. A significant amount of annotated data is required that represents each of the emotions that are used as the classes. In this work, we use ML to classify texts as containing ASB or not. Our aim is to investigate which features are the best for indentifying instances of ASB in texts.

3 Experimental design

For an exploratory purpose, we conducted four experiments using the ASB corpus. We approached the classification task as a binary classification task, that is, a document is classified as either containing or not containing ASB. We compared the positive ASB texts first with each of the three negative sets of examples (Sect 3.1) and then all the corpora together. We approached it in this manner firstly because the corpora are written in different styles and we wanted to observe whether ASB texts show some distinct characteristics allowing for successful classification from each of the three negative sets, secondly because between the sets there was a balance in terms of the number of documents and average size in characters. We experimented with three supervised ML classifiers for the classification task (Sect 3.2) using three sets of features (Sect 3.3). Furthermore, with each experimental corpus, we used ten-cross validation, that is, the entire corpus was first

[1] http://www.homeoffice.gov.uk/crime/anti-social-behaviour/

partitioned into a training set and test set of 90% and 10% respectively, this process was performed ten times. The average results of the 10-cross validation are reported in Section 4.

3.1 Corpora

The following subsections describe each of the four text corpora used in the experimental study. As we are firstly concerned with the binary classification analysis, we compared both positive (those with ASB) (Sect 3.1.1) and negative (Sect 3.1.2 – 3.1.4) examples (non-ASB texts). In order to obtain the negative examples, we used two popular sentiment corpora, movie reviews [31], the emotion annotated corpus (ISEAR) [36] and factual Wikipedia article extracts [44]. Table 1 summarises the documents in the four corpora.

3.1.1 ASB corpus

The ASB corpus is a collection of aggressive, violent, and hostile texts. The texts were collected from various blog posts and news-websites which Munezero et al. [27] could conclusively identify as being ASB. In total 148 documents were identified as ASB. The collection is all English texts, having topics such as: serial killer manifestos, antisocial texts, terrorism, violence-based texts, and suicide notes.

Important to us, the messages in these documents are reflective of the author's thoughts and emotions. The corpus was collected specifically for the purpose of detecting ASB, conflict, crime and violence behaviour from text documents. The collection is based on the research on ASB that has shown that aggression, violence, hostility, and lack of empathy are among the traits that are most directly associated with ASB [6] [32].

3.1.2 International Survey on Emotion Antecedents and Reactions

The ISEAR corpus is a collection of student reports on situations in which the respondents felt any of the seven major emotions: joy, fear, anger, sadness, disgust, shame, and guilt. The responses include descriptions of how they appraised the situation and how they reacted [36].

3.1.3 Movie reviews

This collection consists of 2000 movie reviews. They are labelled in respect to their polarity: negative and positive. The corpus was first used in [31], and now is often applied in sentiment analysis and opinion mining research as a standard development and test set.

3.1.4 Wikipedia text extracts

We searched and collected Wikipedia articles by using similar concepts such as those we found to be characteristic ASB: killing, terror, violence, aggression, and frustration. The aim of including these texts was to observe how well our classification algorithms could distinguish between ASB texts and informative texts containing similar keywords.

3.2 Classifiers

For classifying the documents into the two classes, we experimented with three supervised ML classifiers: Multinomial Naïve Bayes, SMO for the implementation of Support Vector Machines, and J48 for Decision Trees. The three selected algorithms have shown to be effective in various text classification studies. We made use of the WEKA tool [17] to implement the classifiers used in our study.

Multinomial Näive Bayes (MNB). The NB classifier is a probabilistic model that assumes independence of the attributes used in the classification. The classifier has shown good performance even when the sample size is small [11]. We used the MNB classifier implemented in WEKA, which uses a multinomial distribution for each of the features.

Support Vector Machine (SVM) is based on the maximum margin hyperplane rather than probabilities as the Naïve Bayes [23]. The SVM classifier aims to find a hyperplane, represented by a vector that maximally separates the document vectors in one class from those in the other [31].

J48 Decision Tree (J48) is an implementation of the C4.5 decision tree in WEKA. Decision trees are predictive models that are used for classification tasks by starting at the root of tree and moving through it until a leaf is encountered [35]. The decision tree is built from the input training data using the property of information gain or entropy to build and divide nodes of the decision tree in a manner that best represents the training data and the feature vector [12].

3.3 Classification features

3.3.1 Bag-of-Words

As a first experiment with the ASB corpus, we used the Vector Space Model approach so as to consider the words as independent entities. The model makes an implicit assumption that the order of words in document does not matter, which is also referred to as the Bag-of-Words (BoW) assumption. The approach is sufficient for many classification tasks, as the collection of words appearing in the document (in any order) is usually sufficient to differentiate between semantic concepts [23]. Each document in the corpora was represented as a feature vector composed of binary attributes for each word that occurs in the file.

Let $\{f_1,\ldots, f_m\}$ be a predefined set of m features that can appear in a document. Let $n_i(d)$ be the number of times f_i occurs in a document d. Then each document d is represented by the document vector $d{:}=(n_1(d), n_2(d),\ldots,n_m(d))$ [31]. If a word appears in a given document, its corresponding attribute is set to 1; otherwise it is set to 0. Generally, the BoW approach works well for text classification. However, it does not take into consideration any semantic and contextual information.

Moreover, in order to reduce the number of words in the BOW representation we used the LovinsStemmer [22] in order to replace each word by its stem.

Table 1: Corpora description with source, number of documents and average document size.

Corpus	Source	No. of Documents	Avg. Document Size (characters)
ASB	[27]	148	680
ISEAR	[36]	265	110
Movie reviews	[31]	178	390
Wikipedia extracts	[44]	212	680
Total		803	1860

3.3.2 Emotions

Emotions reveal connections of individuals to values in the social world and hence, are the triggers of many social psychological phenomena, such as altruism, antisocial behavior and aggression [32]. In our experiments, we analyse in particular, emotions that might be present in the ASB corpus and analyse whether they are reliable classification features.

To identify the emotions presented in the corpora, we made use of an emotion ontology introduced in [26]. It is an ontology of emotion categories whereby each category contains a set of emotion classes and emotion words. Figure 1, demonstrates the negative emotion with samples from the disliking class.

The emotion ontology is based on WordNetAffect and it contains 85 classes and 1,499 words. On average, the ontology contains 17.6 words per emotion class which gives a relatively wide coverage of emotion classes and emotion words. This together with the fact that the ontology was not fitted on to any particular dataset or text corpus makes it suitable to be used in our experiments as a basis for ASB classification.

For the classification, we made use of two types of emotion-based features: ontology-dependent and ontology-independent emotion features. The ontology-dependent features are collected through a tagging process using the emotion ontology. Through the tagging process, we collected tags such as the sum of all the relative frequencies of the emotion classes that belong to the emotion categories represented in the ontology. While the ontology-independent emotion features were obtained by using the SentiStrength system [42] to calculate the emotion strength of a text.

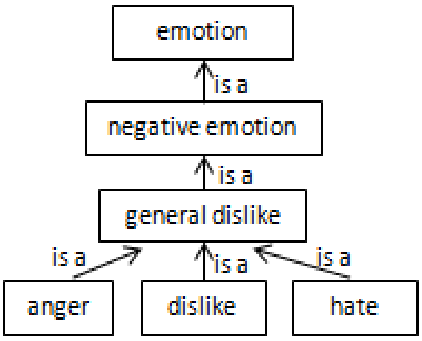


Figure1: Sample from the negative-emotion category section of the ontology.

4 Results

For an exploratory purpose, we conducted four experiments using the ASB corpus and three corpora as negative examples of ASB (Subsection 3.1.2 - 3.2.4). We explored the impact that BoW and emotions as classification features have on the detection of ASB texts. In the first experiment, binary classifiers using the three classifiers described in Subsection 3.2 were trained on ASB+ISEAR, in the second on ASB+Movie reviews, and in the third on ASB+Wikipedia extracts. Finally, all the corpora were combined into a single data set.

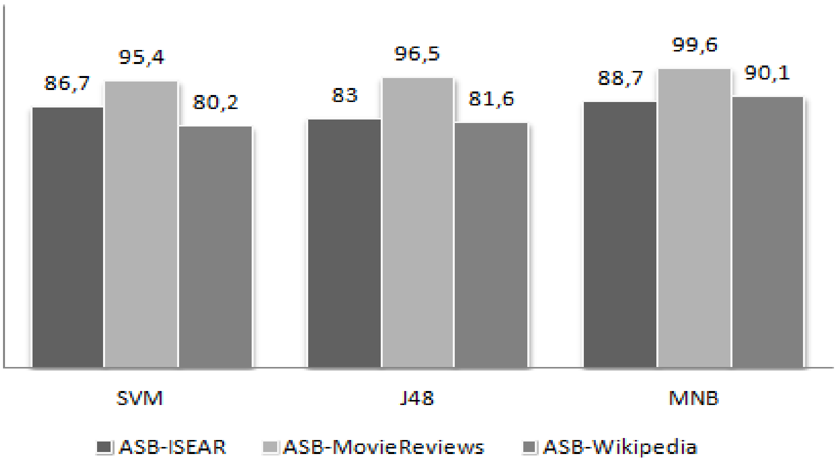


Figure 2: Accuracy results of SVM, J48 and MNB classifier with emotions+BoW as classification features (%).

Table 2: Results from SVM classifier (%).

Corpora	**Features**	**Accuracy**	**Precision**	**Recall**	**F-measure**
ASB + ISEAR	Emotions	84.9	85.8	84.9	84.9
	BoW	86.2	88.7	86.2	86.0
	Emotions+BoW	**86.7**	**89.1**	**86.8**	**86.7**
ASB + MovieReview	Emotions	81.1	82.5	81.2	79.1
	BoW	**95.8**	**95.8**	**95.8**	**95.7**
	Emotions+BoW	95.4	95.5	95.4	95.3
ASB + Wikipedia	Emotions	69.7	70.3	69.7	69.0
	BoW	79.6	80.2	79.6	79.6
	Emotions+BoW	**80.2**	**80.3**	**80.3**	**80.3**

Table 3: Results from J48 classifier (%).

Corpora	**Features**	**Accuracy**	**Precision**	**Recall**	**F-measure**
ASB + ISEAR	Emotions	**84.9**	**84.9**	**84.9**	**84.9**
	BoW	79.2	79.5	79.2	79.2
	Emotions+BoW	83.0	83.2	83.0	83.0
ASB + MovieReview	Emotions	84.6	84.6	84.6	84.6
	BoW	**98.1**	**98.1**	**98.1**	**98.1**
	Emotions+BoW	96.5	96.6	96.5	96.5
ASB + Wikipedia	Emotions	62.5	62.3	62.5	62.3
	BoW	**82.9**	**83.6**	**82.9**	**82.9**
	Emotions+BoW	81.6	82.1	81.6	81.6

Table 4: Results from MNB classifier (%).

Corpora	**Features**	**Accuracy**	**Precision**	**Recall**	**F-measure**
ASB + ISEAR	Emotions	71.7	72.0	71.7	71.5
	BoW	**88.7**	**90.0**	**88.7**	**88.6**
	Emotions+BoW	**88.7**	**89.6**	**88.7**	**88.6**
ASB + MovieReview	Emotions	76.9	77.0	76.9	74.2
	BoW	98.4	98.4	98.4	98.4
	Emotions+BoW	**99.6**	**99.6**	**99.6**	**99.6**
ASB + Wikipedia	Emotions	58.5	59.4	58.6	58.5
	BoW	89.5	90.4	89.5	89.3
	Emotions+BoW	**90.1**	**90.8**	**90.1**	**90.1**

The performances of the classifiers were then compared in terms of accuracy, precision, recall and F-measure. We made use of ten-fold cross validation whereby samples of data are randomly drawn for analysis and the classification algorithm then computes predicted values [23]. Table 2, 3 and 4 show the average of the ten-fold cross validation results on the corpora for each of the ML classifiers with a) BoW, b) emotions, and c) emotions + BoW, as features.

Using SVM classifier, the emotion + BoW features performed better in two of the experiments (ASB+ISEAR and ASB+wikipedia). With J48, the emotions were the better discriminator for the ASB+ISEAR. However the BoW model performed better for the other sets. In looking at the MNB classifier, the emotions + BoW feature set performed better in all the three sets. Hence, in majority of the cases, the addition of emotions to BoW provided better accuracy results. We note that our experiments are preliminary, especially as there is no standard ASB corpus available and the number of documents in the ASB corpora is relatively small.

Figure 2 summarizes the accuracy results of the three classifiers using both the emotions + BoW as features.

From Figure 2, we see that accuracy results in all three classifiers are over 85% which indicates a relatively high accuracy. The best accuracy (99,6%) was reached on the ASB + Movie reviews set with the MNB classifier.

We further noted that when all the four corpora (ASB + ISEAR + MovieReviews + Wikipedia) were combined, the classifiers were not learning. This was due to the imbalance in the class distribution, i.e. the majority of the texts were from the negative (not ASB) class, which then causes ML algorithms to perform poorly on the minority class [18]. However, with the MNB classifier, we observed that it was able to learn in spite of the imbalance.

A closer look at the most predictive features revealed emotion classes such as 'general-dislike', 'hate', 'anxiety', and 'sadness' as expected based on the known connection between ASB and negative emotions. Surprisingly, however the emotion class 'affection' also appeared as a contributing attribute.

5 Conclusion and future work

In this paper, we applied text classification techniques for the analysis and detection of ASB. We reported on experiments where the linguistic features, BoW and emotions were used for the classification of ASB. Our experimental results illustrated that linguistic features such as BoW and emotions can be used successfully to classify ASB in text. We found that the performance of MNB was consistently better than that of J48 and SVM when using the emotions + BoW features. In comparison, when using emotion features alone, the J48 and SVM had the highest accuracy on the ASB+ISEAR (84,9%) and with the BoW features alone, J48 had the highest accuracy with the ASB+MovieReview (98,1%). Thus both features are essentially for achieving high classification accuracy.

Deeper analysis of the features further revealed subsets of emotion features that most contributed to the classification accuracy.

ASB is a growing concern to the society, and in some instances to the government and law enforcement agencies around the world. In line with creating a safer community, identifying the individuals who pose a danger to a community involves analysing the information they put forward. Thus future work involves exploiting the identified linguistic features to build a model to classify new instances of ASB in text as part of an early detection system. Using the features, we would also like to explore the categorizations between different types of ASB, for example physical manifestations of ASB such as violence to other individuals, and non-physical acts such as cyberbullying.

Additionally, with the identified features, we would like to extend the corpus. A larger corpus would allow us to have a larger training set for ML algorithms allowing for learning of new features for building a classification model. In this case, fewer than 200 records were used that could be confidently identified as ASB, and due to this amount, we observed that the SVM and J48 classification models were not learning due to the imbalance in the data.

These experiments were our first attempt at automatically detecting ASB in texts. The results we demonstrated are promising, but experiments on large-scale date are necessary to confirm the robustness of our approach.

Moreover, in this paper, we investigate BoW and emotions as features, but in future we plan to include semantic analysis which could additionally reveal features for ASB identification.

Regardless, in our work, we have found that NLP techniques have potential for the early detection of ASB while the harmful behaviour might still be at its planning stage. Our results have direct applications for national and local security.

Acknowledgement

This work was supported partially by the grant for "Detecting and visualizing their changes in Text" project No. 14166, funded by the Academy of Finland and JSPS KAKENHI Grant Number 25330410.

References

[1] Alm, C. O., & Sproat, R. (2005). Emotional sequencing and development in fairy tales. *Springer*, (pp. 668–674).

[2] Böckler, N., Seeger, T., Sitzer, P., & Heitmeyer, W. (2013). *School Shootings: International Research, Case Studies, and Concepts for Prevention.* New York, USA: Springer.

[3] Bogdanova, D., Rosso, P., & Solorio, T. (2012). Modelling fixated discourse in chats with cyberpedophiles. *Proceedings of the Workshop on Computational Approaches to Deception Detection* (pp. 86–90). Association for Computational Linguistics.

[4] Bradley, M. M., & Lang, P. J. (1999). *Affective norms for English words (ANEW). Instruction Manual and Affective Ratings.* Technical report, University of Florida, The Center for Research in Psychophysiology.

[5] Card, R., & Ward, R. (1998). *The Crime and Disorder Act*. Retrieved 3 28, 2013, from legislation.gov.uk: http://www.legislation.gov.uk/ukpga/1998/37/contents

[6] Clarke, D. (Abingdon, UK). *Pro-Social and Anti-Social Behaviour.* 2003: Taylor & Francis.

[7] Cohen, L. J. (2005). Neurobiology of Antisociality. In C. Stough, *Neurobiology of Exceptionality* (pp. 107-124). New York, USA: Kluver Academic/Plenum Publishers.

[8] Correa, D., & Sureka, A. (2013). *Solutions to Detect and Analyze Online Radicalization: A Survey.* Delhi, India: Indraprastha Institute of Information Technology.

[9] Crowley, S. (2007, 11 7). *Finland Shocked at Fatal Shooting*. Retrieved 3 28, 2013, from BBC News: http://news.bbc.co.uk/1/hi/world/europe/7084045.stm

[10] Danisman, T., & Alpkocak, A. (2008). Feeler: Emotion Classification of Text Using Vector Space Model. *AISB 2008 Convention, Communication, Interaction and Social Intelligence. 2*, pp. 53-59. Aberdeen, UK. Affective Language in Human and Machine.

[11] De Ferrari, L., & Struart, A. (2006). Mining housekeeping genes with a Naive Bayes classifier. *BMC Genomics, 7*(277).

[12] Dinakar, K., Reichart, R., & Lieberman, H. (2011). Modeling the detection of textual cyberbullying. *International Conference on Weblog and Social Media-Social Mobile Web Workshop.*

[13] Ekman, P. (1993). Facial Expression and Emotion. *American Psychologist, 8*(4), 376-379.

[14] Elliot, C. (1992). *The affective reasoner: A process model of emotions in a multi-agent system.* Ph.D. thesis,, Northwestern University, Institute for the Learning Sciences.

[15] Fitzgerald, M. (2011). *Submission to the Department of Human Services on behalf of Public Housing Tenants in relation to Human Rights concerns raised by the Anti-Social Behavior Pilot.* Fitzroy Legal Service.

[16] Frijda, N. H. (1986). Emotional Behavior. In *The Emotions. Studies in Emotion and Social Interaction.* Cambridge University Press.

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

[18] Hulse, J. V., Khoshgoftaar, T. M., & Napolitano, A. (2007). Experimental Perspectives on Learning from Imbalanced Data. *Proceedings of the 24th International Conference on Machine Learning.* Corvallis, OR.

[19] Liu, B. (2010). Sentiment analysis and subjectivity. In N. Indurkhya, & F. J. Damerau (Eds.), *Handbook of Natural Language Processing, (2nd ed.).* Boca Raton, Florida, USA: CRC Press, Taylor and Francis Group.

[20] Liu, H., Lieberman, H., & Selker, T. (2003). A Model of Textual Affect Sensing using Real-World Knowledge. *Proceedings of the 2003 IUI*, (pp. 125-132).

[21] Logan, M. (2012, July). *Case Study: No More Bagpipes*. Retrieved March 15, 2013, from The Threat of the Psychopath: http://www.fbi.gov/stats-services/publications/law-enforcement-bulletin/july-2012/case-study

[22] Lovins, J. B. (1968). Development of a stemming algorithm. *Mechanical Translation and Computational Linguistics., 11*, 22-31.

[23] Miner, G., Elder, J., Hill, T., Nisbet, R., Delen, D., & Fast, A. (2012). *Practical text mining and statistical analysis for non-structured text data applications (1st ed).* Waltham, MA: Academic Press.

[24] Mohammad, S. M. (2012). Portable Features for Classifying Emotional Text. *Proceedings of the 2012 NAACL HLT*, (pp. 587–591).

[25] Mohammad, S. M., & Turney, P. D. (2010). Emotions Evoked by Common Words and Phrases: Using Mechanical Turk to Create an Emotion Lexicon. *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, (pp. 26-34).

[26] Montero, C. S., Kakkonen, T., & Munezero, M. (2014). Investigating the Role of Emotion-based Features in Author Gender Classification of Informal Text. A. Gelbukh (Ed.): *CICLing 2014, Part II, Lecture Notes in Computer Science 8404*, (pp. 98–114), Springer-Verlag Berlin Heidelberg 2014.

[27] Munezero, M., Mozgovoy, M., Kakkonen, T., Klyuev, V., & Sutinen, E. (2013). Antisocial behavior corpus for harmful language detection. *Federate Conference in Computer Science.* Krakow, Poland.

[28] Ortony, A., Clore, G. L., & Collins, A. (1994). The Structure of the Theory. In *Chapter 2 in The Cognitive Structure of Emotions* (pp. 15-33). Cambridge University Press.

[29] Ortony, A., Clore, G. L., & Foss, M. A. (1987). The Referential Structure of the Affective Lexicon. *Cognitive Science, 11*, 341-364.

[30] O'Toole, M. E. (2000). *School Shooter: A Threat Assessment Perspective. National Center for the Analysis of Violent Crime.* Quantico, Virginia, USA.: Federal Bureau of Investigation.

[31] Pang, B., Lee, L., & Vaithyanatha, S. (2002). Thumbs up?: sentiment classification using machine learning techniques. *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10:* (pp. 79-86). Association for Computational Linguistics.

[32] Parrot, G. W. (2001). *Emotions in Social Psychology.* Philadelphia, Pennsylvania, USA: Taylor & Francis.

[33] Pennebaker, J. W., Mehl, M. R., & Niederhoffer, K. G. (2003). Psychological Aspects of Natural Language Use: Our Words, Our Selves. *Annual Review of Psychology, 54*, 547-577.

[34] Plutchik, R. (2001). The Nature of Emotions. *American Scientist, 89*(4), 344-350.

[35] Quinlan, J. R. (1993). *C4.5: Programs for machine learning.* San Mateo, Calif: Morgan Kaufmann Publishers.

[36] Scherer, K. R., & Wallbott, H. G. (1994). Evidence for universality and cultural variation of differential emotion response patternin. *Journal of personality and social psychology, 66*, 310.

[37] Shaver, P., Schwartz, J., Kirson, D., & O'Connor, C. (1987). Emotion knowledge: Further exploration of a prototype approach. *Journal of Personality and Social Psychology, 52*, 1061-1086.

[38] Stone, P. J., Dunphy, D. C., Smith, M. S., & Ogilvie, D. M. (1966). *The General Inquirer: A Computer Approach to Content Analysis.* Cambridge, Massachusetts, USA: The MIT Press.

[39] Strapparava, C., & Mihalcea, R. (2008). Learning to Identify Emotions in Text. *Proceedings of the ACM SAC'08*, (pp. 1556-1560).

[40] Strapparava, C., & Valitutti, A. (2004). WordNet-Affect: an Affective Extension of WordNet. *Proceedings of the 4th LRE*, (pp. 1083-1086).

[41] The INDECT Consortium. (n.d.). *XML Data Corpus: Report on Methodology for Collection, Cleaning and Unified Representation of Large Textual Data from Various Sources: News Reports Weblogs Chat.* Retrieved 10 10, 2010, from http://www.indect-project.eu/files/deliverables/public/INDECT_Deliverable_4.1_v20090630a.pdf (2010, Dec. 10).

[42] Thelwall, M., Bucley, K., Paltoglou, G., & Cai, D. (2010). Sentiment Strength Detection in Short Informal Text. *Journal Of The American Society for Information Science And Technology., 61*(12), 2544–2558.

[43] Valitutti, A., Strapparava, C., & Stock, O. (2004). Developing Affective Lexical Resources. *PsychNology, 2*(1), 61-83.

[44] Wikimedia Foundation. (2013, May 08). Wikipedia: The Free Enceclopedia.

[45] Yin, D., Xue, Z., Hong, L., Davison, B. D., Kontostathis, A., & Edwards, L. (2009). Detection of harassment on web 2.0. *Proceedings of the Content Analysis in the WEB*, *2.*

Leveraging User Experience through Input Style Transformation to Improve Access to Music Search Services

Marina Purgina, Andrey Kuznetsov and Evgeny Pyshkin
St. Petersburg State Polytechnical University
Institute of Computing and Control
Polytechnicheskaya ul. 21, St. Petersburg 194021, Russia
E-mail: mapurgina@gmail.com, andrei.n.kuznetsov@gmail.com, pyshkin@icc.spbstu.ru
http://kspt.ftk.spbstu.ru/info/staff/pyshkin/en

Keywords: information retrieval, music search, mobile application, query transformation, input style transformation

Received: November 23, 2013

We analyze problems of music searching and main tasks the developers face in the domain of music information retrieval. We introduce the architecture of the software and the data model for integrated access to existing music searching web services. We illustrate our approach by developing a mobile accessed software prototype which allows users of Android running touch screen devices accessing several music searchers including Musipedia, Music Ngram Viewer, *and* FolkTuneFinder. *The application supports various styles of music input query. We pay special attention to input style transformation aimed to fit well the requirements of the supported search services.*

Povzetek: Opisana je metoda iskanja glasbenih posnetkov na androidnih napravah.

1 Introduction

A variety of multimedia resources constitutes considerable part of the present-day Web information content. Numerous search services usually provide special features to deal with different types of media such as books, maps, images, audio and video recordings, software, etc. In addition to general-purpose searching systems there are solutions using specialized domain sensitive interfaces. Truly, quality of a search service depends both on the efficiency of algorithms it relies on, and on user interface facilities. Such interfaces may include special syntax forms, user query visualization facilities, interactive assisting tools, components for non-textual query input, interactive and "clickable" concept clouds, and so on [1]. Depending on the searching tasks, specialized user interfaces may support different kinds of input like mathematical equations or chemical changes, geographic maps, XML-based resource descriptions, fragments of software source code, editable graphs, etc.

In text searching such aspects as morphological, synonymic and grammar variations, malapropisms, and spelling errors condition particular difficulties of a searching process. In the music searching domain there are specific complications like tonality changes, omitted or incorrectly played notes or intervals, time and rhythmic errors. Thus, although there are eventual similarities between text and music information retrieval, they differs significantly [2].

Human ability to recognize music is strongly interrelated to listener's experience which may be considered itself to be a product of music intelligent perception [3, 4]. Recently (see [5]) we also analyzed internal models of music representation (with most attention to function-based representation) being the foundation of various algorithms for melody extraction, main voice recognition, authorship attribution, etc. Music processing algorithms use the previous user experience implicitly. As examples, we could cite the *Skyline* melody extraction algorithm [6] based on the empirical principle that the melody is often in the upper voice, or *Melody Lines* algorithms based on the idea of grouping notes with closer pitches [7].

The remaining text of the article is organized as follows. In section 2 we review music searching systems and approaches of the day. We also introduce our experience in the domain of human centric computing and refer to some recent related works. In section 3 we describe music query input styles and analyze possible transformations of music input forms so as to fit the requirements of search services. Section 4 contains the description of the developed Android application architecture. We show how it works and make an attempt to analyze the searching output from the point of a musicologist.

2 Background and related works

In general, we are able to search music either by metadata description, or by music content. Searching by metadata

This paper is based on M. Purgina, A. Kuznetsov and E. Pyshkin *An Approach for Developing a Mobile Accessed Music Search Integration Platform* published in the proceedings of the 3rd International Workshop on Advances in Semantic Information Retrieval (part of the FedCSIS'2013 conference).

seems to be very similar to textual searching. Metadata is not necessarily to be restricted by bibliographical data like author, title, artist, conductor, editor, date of publication, etc. They may also include information about performance itself like time signature, tempo, musical instruments, tonality, lyrics and so on. In some situations searching by metadata and searching by content are not so different. Let us consider "A Dictionary Of Musical Themes" [8] which includes short snippets (transposed to C-dur tonality) of musical themes of a composition. These snippets can be used to seek unknown composition by its theme. On one hand, these themes extracted from the original composition constitute metadata, on the other hand, they enable searching a composition by its content. In fact, user can use an ordinary text-based search engine to retrieve compositions represented in such a dictionary. Nowadays exactly the same technique are being used in indexing algorithms and fingerprint algorithms with only few differences: a) metadata are extracted automatically, and b) being sort of pure mathematical abstractions (e.g. hashcodes, fingerprint vectors, etc.) metadata may have no sense for humans.

Since the time when the first dictionaries of musical themes were created, the world dramatically changed. People developed new multimedia carriers requiring more complicated search scenarios:

1. Searching music information by existing audio fragment considered as an input.
2. Searching compositions by human remembrance represented in a form of singed, hummed, tapped or anyhow else defined melody or rhythm fragment.
3. Searching music by lyrics.
4. Searching music by bibliographical data (e.g. title, author etc.).
5. Searching music by keywords (e.g. "scary Haloween music").

Searching by given audio fragments is supported by many specialized search engines like *Audiotag*, *Tunatic* or *Shazam*[9]. As a rule, it is implemented on the basis of so called audio fingerprinting technique. The idea of such an approach is to convert an audio fragment of fixed length to a low-dimensional vector by extracting certain spectral features from the input signal. Then this vector (being a kind of audio spectral fingerprint) is compared to fingerprints stored in some database [10, 11].

In the case of *Searching by human remembrance* scenario we can distinguish two situations. In the first one the search engine deals a monophonic user query representing a main voice, a rhythm, a melodic or rhythmic contour. In the second case the user query represents polyphonic music fragment (e.g. while searching by note score). Errors in user queries condition the main problems of *Searching by human remembrance*. Thus, music fragments comparison algorithms have to be robust in regards to the most popular user errors like expansion, compression, omission or repetition [12, 13]. The another complication is that it is impossible to search directly within the audio resources' binary contents since we usually have no exact faithful audio fragment[2].

Searching by lyrics is not fully automated. An end user is able to use general purpose text search machine (like google, yahoo or yandex) for this task, but text based tools search in existing textual data which is usually published either by author or by music lovers. In theory it's possible to use speech recognition engine for the purpose of lyrics extraction [14], but in practice the recognition quality is not good enough for automatic lyrics transcription. We experimented with Google Voice service and it showed good results for lyrics recognition if a user is singing in silent environment with no background music (it successfully recognized 17 of 20 songs). But if we try to play a broadcasted recording of popular artists, the Google Voice simply ignored the input just like it was nothing sang at all. Indeed, the Google Voice was designed as a speech recognition service, not lyrics transcription service. The problem of automatic recognition of lyrics in singing exists, but this topic is actually out of scope of this research.

The first three cited scenarios are related to the so called *cover song identification* task. However the final goal of such a kind of searching process is not always a song itself (which may be an object of copyright restrictions). The user may be satisfied with obtaining music bibliographical metadata that can be used to look for the song in an online music store. It is exactly what the fourth scenario *Searching music by bibliographical data* means.

Finally, *Searching by keywords* is often implemented through tags annotations. In this case a search query is being parsed to find keywords that could be considered to be tags. Then these tags (i.e. words from a predefined vocabulary of genres, moods, instruments etc.) are used for searching through an annotated database. For example, such technique is used by Last.fm or Pandora online radios.

The search scenarios we explained here cover only the most common tasks. Besides such kinds of usual tasks, there are many other music information retrieval problems including searching compositions by their emotional properties (this task appears in recommendation systems), identifying exact particular performance instead of cover song retrieval, locating a position inside a song (used in score following, for instance), and many others[3].

Similar to other kinds of IR systems, a MIR system usually contains frontend and backend components. In this paper we pay attention mostly to a frontend part communicating with existing searchers, considering a backend system as an *Application Program Interface* (API).

[2] For the user provided sequence "A4 B4 C4" as an input it may happened that a melody that the user is actually looking for does not contain such notes at all.

[3] The overview of the most popular MIR tasks together with descriptions of the recent algorithms can be found at Music Information Retrieval Evaluation eXchange (MIREX) home page (see `http://www.music-ir.org/mirex`).

The extensive description of input styles used by music search engines may be found in [15]. Presently there are many searching web services like *Midomi*, *Musipedia*, *Ritmoteka*, *Songtapper*, *Music Ngram Viewer*, and *FolkTuneFinder* where customers use one of several possible styles to input a music query.

Table 1 represents possible ways to access different music web search services and pays attention to the following facilities:

- **Voice** Using voice recorded from microphone
- **Rhythm** Using tapping/clicking with keyboard, mouse or other input device
- **Tags** Support for keywords or tags
- **Exmpl** Using uploaded audio fragment as an example
- **Lyr** Searching by lyrics
- **Notes** Music score or pitch notation
- **VKB** Virtual keyboard generating note sequence with rhythm
- **URD** Parsons code
- **API** External API (SOAP, REST, etc.)

Nowadays many services are accessible via browsers since they support Web interface features. Another important issue is the possibility to access some services from inside the software applications by using open protocols. It gives the way to create tools which allow users not to be limited by only one service at a time.

Particularly, *Musipedia* service uses SOAP protocol described in [16]. *FolkTuneFinder* and *Music Ngram Viewer* (both are also used in our work as target search services) are based on the REST architectural style and their responses are wrapped in JSON format. Detailed description of the API usage rules and examples for *Music Ngram Viewer* service may be found in [17].

With respect to music inputs styles, existing tools support the following opportunities to define a music fragment:

- To sing or to hum the theme and to transfer the recording to the music search engine.
- To write notes by using one of known music notations directly (e.g. music score, Helmholtz or American pitch notation, MIDI notation, etc.).
- To tap the rhythm.
- To play the melody with the use of a virtual keyboard.
- To use MIDI-compatible instrument or it's software model.
- To enter Parsons code, or to set the melody contour by using "U, R, D" instructions [4] as shown in Figure 1.
- To define keywords or enter text query.
- To record a piece of original composition (e.g. record a composition played on a radio with use of microphone) and to transfer the recording to the music search engine.

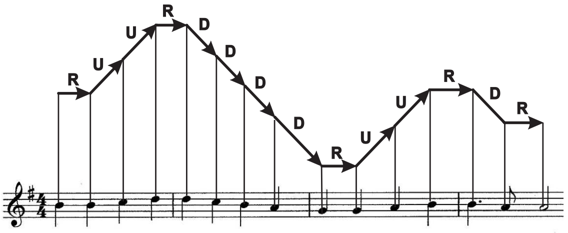


Figure 1: Beethoven's "Ode to Joe" fragment represented in Parsons code.

3 Music query input styles

As shown in the above section, we may define the music query by using different input styles. For a searching framework, the important issue is not only featuring different input interfaces but transforming one query form to the another depending on search service availability and it's communication schema.

Different input styles are useful since the user music qualification differs. Melody definition by using a virtual or real keyboard is one of the most exact ways to represent the query, since it accumulates most melody components. However it is not common that users are skilled enough to use the piano keyboard as well as to write adequate note score.

Contrariwise, tapping a rhythm seems to be relatively simple way to define music searching query. The problem is that the number of possible rhythm patterns is evidently less than the number of compositions. It means that even if we succeed to tap the rhythm correctly, we may apparently have a list of thousands titles in return [5].

If a user didn't record a fragment while the composition was playing, then the only choice is to sing a melody by voice. Recording a voice (so called query-by-humming or query-by-singing) requires both user's singing skills and support for such a facility from the search system. It is important to note that query-by-example and query-by-humming are quite different tasks of MIR[5]).

[4] Each pair of consecutive notes is coded as **U** ("sound goes Up") if the second note is higher than the first note, **R** ("Repeat") if the consecutive pitches are equal, and **D** ("Down") otherwise. Some systems use **S** ("the Same") instead of **R** to designate pitch repetition. Rhythm is completely ignored.

[5] `http://www.music-ir.org/mirex/wiki/2013:Main_Page`

Table 1: Accessing Music Searching Web Services

Name	*Access*								
	Voice	*Notes*	*VKB*	*URD*	*Rhythm*	*Tags*	*API*	*Exmpl*	*Lyr*
Audiotag[a]	–	–	–	–	–	–	–	+	–
Tunatic[b]	–	–	–	–	–	–	–	+	–
Shazam[c]	–	–	–	–	–	+	–	+	–
Midomi[d]	+	–	–	–	–	–	–	–	–
Musipedia[e]	+	+	+	+	+	+	SOAP	–	–
Ritmoteka*f*	–	–	–	–	+	–	–	–	–
Songtapper[g]	–	–	–	–	+	–	–	–	–
Music Ngram Viewer[h]	–	–	+	–	–	+	REST	–	–
FolkTuneFinder[i]	–	–	+	+	+	+	REST	–	–
Google, Yandex, Yahoo!	–	+	–	–	–	+	+	–	+

[a] http://www.audiotag.info
[b] http://www.wildbits.com/tunatic/
[c] http://www.shazam.com
[d] http://www.midomi.com
[e] http://www.musipedia.org
[f] http://www.ritmoteka.ru
[g] http://www.bored.com/songtapper
[h] http://www.peachnote.com
[i] http://www.folktunefinder.com

Mobile devices with touch screens affect strongly usage aspects of music searching interfaces. Such devices make possible simulating many kinds of music instruments, although the virtual piano-style keyboard remains the most popular interface.

3.1 Transformation of input styles

We represent relationships between selected music query input styles in form of an oriented graph. In Figure 2 blue nodes correspond to query representation, red nodes correspond to input methods. *Notes* are used to represent both a query and an input method, and denoted by using grey color. Every transition arc denoted by latin letters (*a* to *r*) shows the possible transformation from one input style to another, namely: a) synthesis & automatic notes transcription; b,c,g) equivalent symbolic transformation; d) pitch estimation; e) pitch sequence with fixed rhythm pattern; f) rhythm with fixed pitch pattern; h) keep only pitch values; i) keep only time intervals; j) calculate pitch intervals[6]; k) calculate inter offset intervals[7]; l) compare pitches; m) compare time intervals; n) pitch sequence with fixed pitch interval pattern; o) rhythm with fixed IOI pattern; p) compare pitch intervals; q) compare IOI; r) onset time estimation.

Since the virtual keyboard based query implicitly includes such note attributes as it's start time, duration and pitch value, there is no much difficulty to transform the keyboard input into the rhythm or pitch notation. The same information (notes) could be extracted from the voice input. Query-by-humming searching machines usually don't need such kind of transformations and use hummed or singed input "as is". However it's still possible to transform singed input into symbolic form in order to create queries for search machines that don't support query-by-humming method. Most recent comparative study of pitch extraction algorithms can be found in [18], and most recent results for multiple fundamental frequency estimation task can be found on MIREX page[8].

Regardless of how we get the notes, we can easily transform them into a rhythm or a pitch notation. This transformation is not lossless. When we transform notes into rhythm we lose information about pitch values, and when we transforming to pitch sequence we lose information about rhythm. Next we can reduce absolute values of pitches and time intervals, and we get sequence of pitch-intervals or sequence of Inter Offset Intervals (IOI). Then we can reduce interval values and get melodic or rhythmic contour. Again this is one-way transformation because we lose an information about the value of an interval and leave only sign of the value encoded with letters 'U', 'R' and 'D'.

Clear that walking through the graph from left to right we reduce the user query, and therefore it seems we couldn't expect better searching results. However such transformations may have sense for at least two reasons:

- We attempt to emphasize the meaning of special

[6] Strictly, this is one way transformation (back transformation produces many transpositions, but as we said before, a comparison algorithm should be robust against transpositions)

[7] Strictly, this is one way transformation (back transformation produces many different tempos, but as we said before, a comparison algorithm should be robust against tempo fluctuation)

[8] http://www.music-ir.org/mirex/wiki/2013:Multiple_Fundamental_Frequency_Estimation_\%26_Tracking_Results

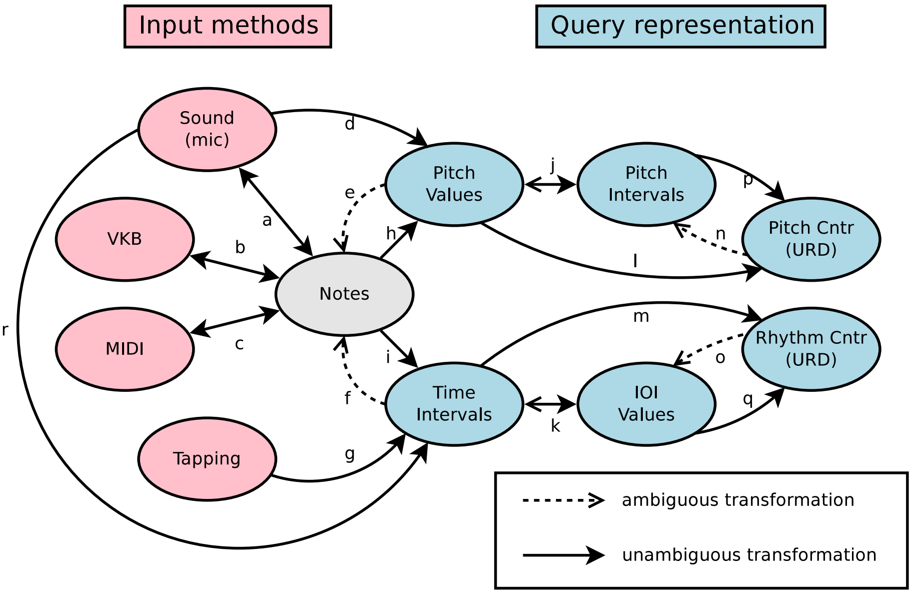


Figure 2: Graph of music query transformations.

melody attributes.

- We would like to try to connect a search service which probably uses quite different music database (e.g. specialized on some music genre[9]) although it supports only restricted input methods (e.g. rhythm or pitch notation).

Regarding to the user interface issues, the ability to move from one input style to another renders possible to switch easily between different searching systems within the framework of one mobile or web application without re-entering the query.

3.2 Models to represent queries

Usually user queries are relatively short (and it is true not only for the case of music [19]), so we use sequence of *Note* objects to represent the searching query.

The attributes of a *Note* object are the following:

- *Note name* according to the American pitch notation
- Its *Octave number*
- Its *Onset time*
- Its *End time*

This representation is equivalent to a subset of MIDI subset of MusicXML format. It means that we can get query in MusicXML format directly from our representation in order to upload it to statistics server. XML based standard is always a good choice for future compatibility. An ability to upload user queries to a server is extremely important feature for the domain of MIR research, because this is the cheapest way to receive information for further investigations [10]. For instance, collected information can be used for modeling user errors.

This representation is equivalent to notes representation so we can easily transform it into the desired form like a sequence of pitches or a rhythm pattern.

4 Introducing mobile application for accessing music search services

Nowadays, people are happy to use their mobile devices to access different search services at any time from any place. They use different types of such devices which may have different input mechanisms like phone keys, *qwerty*-keyboards, touch screens, voice recognition, and so on. The variety of devices running on Android operating system is rapidly increasing during last years, so we decided to use Android platform for our music searching application.

For our implementation we selected some music searchers which may be accessed programmatically, particularly: *Musipedia*, *Music Ngram Viewer*, and *FolkTuneFinder*. For three searching systems we implemented four user query input styles:

- Note score editor supporting one voice definition
- Parsons code
- Rhythm tapping

[9] We turn our attention to the example of such a case in the following section of this paper.

[10] This approach is very similar to a *Game With A Purpose* (GWAP) that is widely used to collect annotations for data. As an example we can cite "Major Miner – music labeling game" for audio data or "Google Image Labeler" for pictures.

- Piano style virtual keyboard with additional representation of American pitch notation

4.1 Application architecture

General application construction ideas are common for various operating platforms. Despite the fact that eventually we developed an application for Android operating system, here we describe common application architecture in platform independent way, but keeping in mind that target device is a mobile device. The Android application can serve as a model for implementing flexible human centric interface which is oriented to present-day style of using hardware and software facilities of various mobile devices.

Figure 4 represents main components of our music searching application.

There are following UI and non UI components shown in Figure 4:

- UI views (boxes with blue background)
- Query transformer
- Search system adapters (one adapter for each search system, boxes with red background)
- Serializers (either JSON or XML)
- Connectors (only one connector supported so far: HTTP)

This is a sort of scalable architecture. We can easily add new query input methods, new search machines or new communication protocols (like FTP or even SMTP if required). The theory of operations is the following. The main view *InputStyleSelection* provides the interface for input style choice. According to the selected input style the respective view (*MelodyContour*, *MusicScore*, *VirtualPiano*, or *RhythmTapper*) opens and provides the corresponding input interface. Depending on the input style the user query could be either a sequence of *Note* objects (Music Score and Virtual Piano produce this output), *Rhythm* (produced by RhythmTapper) or *Melody contour*. Then the application iterates through a collection of available adapters for search engines. Depending on the adapter capabilities the input query may be transformed to the acceptable representation. Then the adapter performs a request to a search service. The request is serialized with a serializer (JSON or XML) and performed through one of available connectors. The response is parsed by the appropriate adapter and added to a list of search results to be displayed by the SearchResults view.

With respect to search services' application interfaces mentioned in section 2, the web information exchange protocol adapters have been implemented.

The SOAP protocol is not recommended for mobile devices since it uses verbose XML format and may be considerably slower in comparison with other middleware technologies. Unfortunately it is the only way to communicate with the *Musipedia* system. In our case, the mentioned SOAP disadvantages shouldn't case concern since the exchange occurs relatively rarely, only when the respective button is pressed by a user, and there is small amount of information being transferred. We use *org.ksoap2* Java package [20] containing classes required for handling SOAP envelopes and literal XML content. To implement interaction with other search services (based on the REST architecture and wrapping their responses in JSON format which is typically more compact in comparison with XML) we use *Google Gson* Java library [21]. It allows converting Java objects into their JSON representation as well as backward converting JSON strings to equivalent Java objects.

4.2 Usage example

The application starts with a welcome screen for the preferred input style selection (Figure 4).

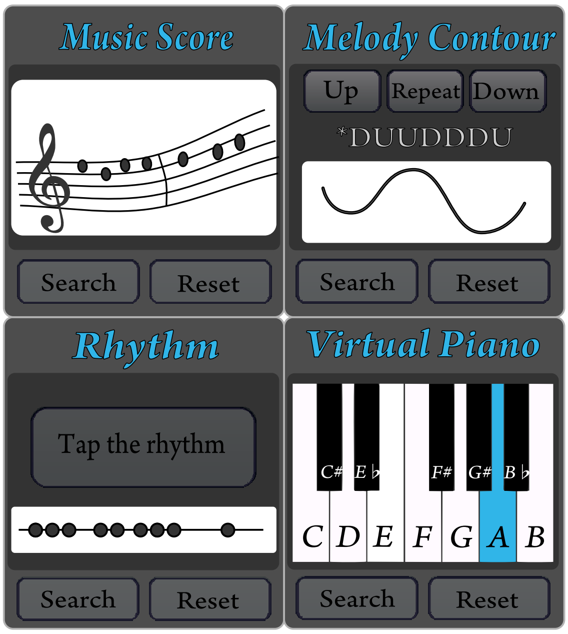


Figure 4: Main activity: input style selection.

Then a user selects an input style. For example if a user selects the virtual keyboard interface, the virtual piano is displayed on the screen. The melody is stored in form of a note sequence with respect to the following related data: a *pitch* represented in the American pitch notation (note name and octave number), *onset time*, *end time*.

Other properties may be computed depending on the requirements of a music searcher. Let us illustrate this by the input represented in form of a simplified timing chart (with respect to the note names rather than sound frequencies). The chart in Figure 5 represents some first notes of the well known Russian folk song "Birch Tree".

For the reason that *Musipedia* searcher requires a sequence of triplets containing an onset time, a MIDI pitch and its duration, the user input shown in Figure 5 is converted to the following query data:

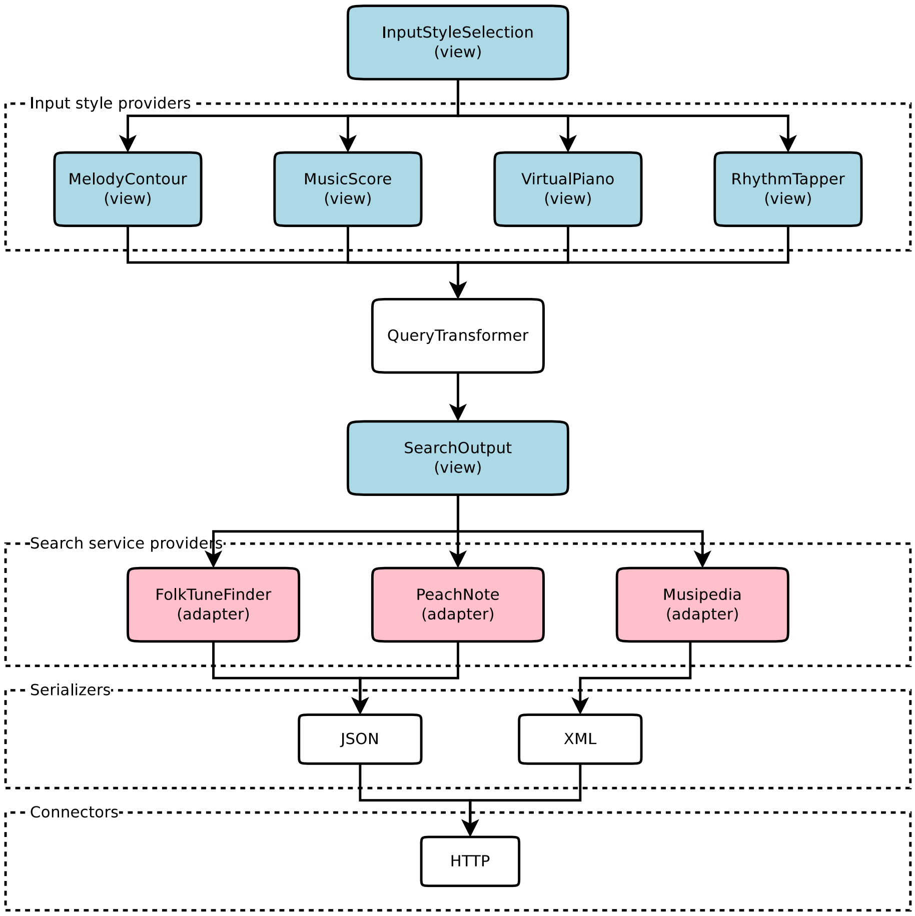


Figure 3: Mobile music searching application architecture.

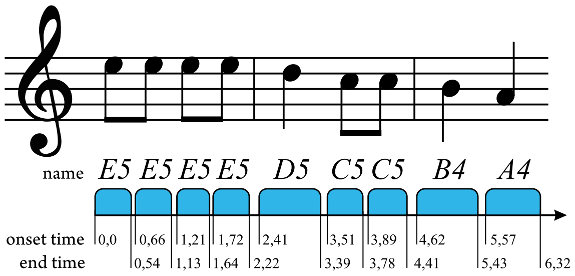
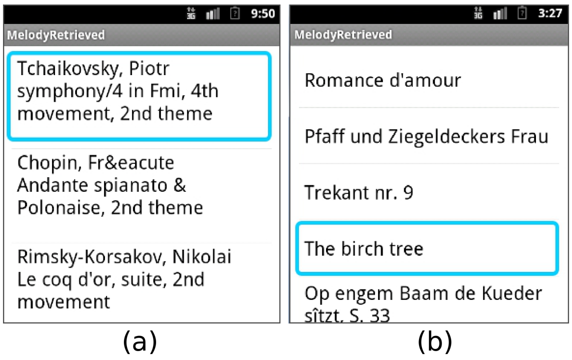


Figure 5: Test melody: note score representation and timing chart.

0.0, 76, 0.54; 0.66, 76, 0.47; 1.21, 76, 1.43; 1.72, 76, 0.50; 2.41, 74, 0.98; 3.57, 72, 0.27; 3.89, 72, 0.52; 4.62, 71, 0.81; 5.57, 69, 0.75;

After this the respective information is included to the SOAP request which is subsequently sent to the *Musipedia* server. As a result, the searching system returns the list of retrieved compositions as shown in Figure 6(a) [11]. We see the confirmation of the known fact that this melody was used by Piotr Tchaikovsky in the 4th movement of his Symphony No.4 in F-moll (compare with the fragment of the symphony note score shown in Figure 7).



Tchaikovsky, Piotr symphony/4 in Fmi, 4th movement, 2nd theme

Chopin, Fr&eacute Andante spianato & Polonaise, 2nd theme

Rimsky-Korsakov, Nikolai Le coq d'or, suite, 2nd movement

Romance d'amour

Pfaff und Ziegeldeckers Frau

Trekant nr. 9

The birch tree

Op engem Baam de Kueder sîtzt. S. 33

Figure 6: Results retrieved by *Musipedia* and *FolkTuneFinder* searchers.

Using other searching engines may enhance searching results by taking into account other music genres. Let us take the *FolkTuneFinder* service which requires a sequence of MIDI pitches. Hence the user input is transformed into the sequence of MIDI pitches as follows:

[11] We selected Tchaikovsky's work, but as you can see, the similar theme may also be recognized in some other known compositions.

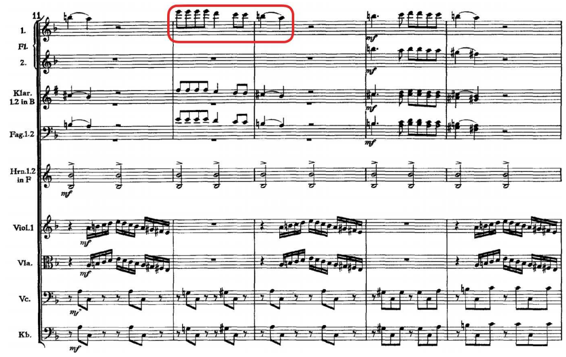


Figure 7: Birch Tree song cited by Tchaikovsky in his 4th symphony.

76, 76, 76, 76, 74, 72, 72, 71, 69

For the case of the melody contour defined with using Parsons code, the user input is the following *"RRRD-DRDD"*. We implemented the interface component which allows constructing the URD-query by pushing buttons *Up*, *Down* and *Repeat* with synchronous demonstration of the respective graphical contour which is being generated automatically[12]. As you see in Figure 6(b) the resulting output also contains the "Birch Tree" among other compositions. Note that since the melody contour is a less exact input method (comparing to direct melody definition), it is normal that we don't have the desired melody in the very first lines.

The example we selected for the illustration shows well one important aspect of music searching process, although in a slightly simplified manner. When we discover the composition corresponding to the given request, we may expect obtaining even more information than simply a desired piece of music. Fast every Russian knows the "Birch Tree" song since the early childhood years. But only those who listen to the classical music discover this theme in one motive of Tchaikovsky's symphony. In contrast to this, western music lovers may listen this motive first just in the Tchaikovsky's work, and after a while recognize it as a citation of the Russian folk song. Isn't it a kind of process similar to a music perception in terms of musicology?

5 Conclusion and future work

In the domain of human-centric computing much attention is paid to the facilitating user interface features in relation with a kind of data being processed. As a special type of information retrieval systems, music retrieval systems demand special ways to interact with users. They include not only traditional text or media based queries but specific forms of user input facilities such as note score representations, virtual or MIDI-compatible instruments, as well as composing queries based on melody humming or rhythm tapping which may contain errors of human interpretation. Such approaches may help to overcome limitations of fingerprinting techniques which require exact or nearly exact audio fragments to proceed with searching in the databases of stored music compositions. In our work we investigated styles of user inputs used in various music search services and applications. We applied transformation rules of query conversion from one input style to another to a software tool communicating with programmatically accessible music search services from mobile devices running on the Android operating system.

In the current implementation we supported only those music queries which are representable in symbolic form (e.g. note score, pitch notation, note sequences, or contour symbolic description). User interface facilities may be improved if we consider other ways to interact with the user having a touch screen device. It may include, for example, melody contour or rhythm drawing facilities. Even for the search services that we used currently, there are input styles which are still not incorporated into the existing software prototype. We investigate possibilities to support interfaces for melody singing or humming. Actually we faced the problem to pass the audio query to the searching engines via existing data transfer protocols that we are allowed to use.

Another way to extend the interface is to provide additional filters like http://www.folktunefinder.com/search/melody/ does. As described in [22], we can provide additional filters or search criteria to experienced users. If a user can provide an information about time signature, tonality, instruments, or define other metadata, there is no objective to prevent the user from doing that. Probably two problems that might appear here is a) our UI will be overcomplicated and b) target search machine may not support such kind of searching. Anyway there is an area for research, how to provide this capability without strong coupling with any of target music search machines.

Ways to extend the interface may also include a support for connected MIDI-compatible devices and text-based searching facilities aimed to explore music metadata information. The another interesting improvement which could fit well especially mobile equipment interfaces is to support music tagging as described for example in [23]. Hence the key idea is to connect different kinds of search services with rich user input facilities so as to follow better the usage style of modern mobile devices.

References

[1] E. Pyshkin and A. Kuznetsov, "Approaches for web search user interfaces," *Journal of Convergence*, vol. 1, no. 1, 2010.

[2] Z. Mazur and K. Wiklak, "Music information retrieval on the internet," in *Advances in Multime-*

[12] We consider to investigate possibility to support a melody contour drawing interface in future implementations.

dia and Network Information System Technologies. Springer, 2010, pp. 229–243.

[3] B. Snyder, *Music and Memory: An Introduction*. Cambridge, Mass. [u.a.]: MIT Press, 2000.

[4] D. Deutch, "Music perception," *Frontiers in Bioscience*, 2007.

[5] A. Kuznetsov and E. Pyshkin, "Function-based and circuit-based symbolic music representation, or back to Beethoven," in *Proceedings of the 2012 Joint International Conference on Human-Centered Computer Environments*. ACM, 2012, pp. 171–177.

[6] A. L. Uitdenbogerd and J. Zobel, "Manipulation of music for melody matching," in *Proceedings of the sixth ACM international conference on Multimedia*, Bristol, United Kngdm, September 1998.

[7] C. Isikhan and G. Ozcan, "A survey of melody extraction techniques for music information retrieval," in *Proceedings of 4th Conference on Interdisciplinary Musicology (SIM'08)*, Thessaloniki, Greece, July 2008.

[8] H. Barlow and S. Morgenstern, *A dictionary of musical themes*. Crown Publishers, 1948. [Online]. Available: http://books.google.ru/books?id=jZ5HAAAAMAAJ

[9] A. Wang, "The shazam music recognition service," *Communications of the ACM*, vol. 49, no. 8, pp. 44–48, 2006.

[10] W. Hatch, "A quick review of audio fingerprinting," McGill University, Tech. Rep., March 2003. [Online]. Available: http://www.music.mcgill.ca/~wes/docs/finger2.pdf

[11] P. Cano, T. Kalker, E. Batlle, and J. Haitsma, "A review of algorithms for audio fingerprinting," *The Journal of VLSI Signal Processing*, vol. 41, no. 3, pp. 271–284, 2005.

[12] R. J. McNab, L. A. Smith, I. H. Witten, C. L. Henderson, and S. J. Cunningham, "Towards the digital music library: Tune retrieval from acoustic input," 1996.

[13] S. Downie and M. Nelson, "Evaluation of a simple and effective music information retrieval method," in *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, ser. SIGIR '00. New York, NY, USA: ACM, 2000, pp. 73–80. [Online]. Available: http://doi.acm.org/10.1145/345508.345551

[14] A. Mesaros and T. Virtanen, "Automatic recognition of lyrics in singing," *EURASIP J. Audio Speech Music Process.*, vol. 2010, pp. 4:1–4:7, Jan. 2010. [Online]. Available: http://dx.doi.org/10.1155/2010/546047

[15] A. Nanopoulos, D. Rafailidis, M. M. Ruxanda, and Y. Manolopoulos, "Music search engines: Specifications and challenges," *Information Processing & Management*, vol. 45, no. 3, pp. 392–396, 2009.

[16] "Musipedia SOAP interface." [Online]. Available: http://www.musipedia.org/soap_interface.html

[17] "Music ngram viewer API." [Online]. Available: http://www.peachnote.com/api.html

[18] O. Babacan, T. Drugman, N. d'Alessandro, N. Henrich, and T. Dutoit, "A comparative study of pitch extraction algorithms on a large variety of singing sounds," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, 2013, pp. 7815–7819.

[19] V. Klyuev and Y. Haralambous, "A query expansion technique using the ewc semantic relatedness measure," *Informatica: An International Journal of Computing and Informatics*, vol. 35, no. 4, pp. 401–406, 2011.

[20] "Package org.ksoap2." [Online]. Available: http://ksoap2.sourceforge.net/doc/api/org/ksoap2/package-summary.html

[21] I. Singh, J. Leitch, and J. Wilson, "Gson user guide." [Online]. Available: https://sites.google.com/site/gson/gson-user-guide

[22] A. Kuznetsov and E. Pyshkin, "Searching for music: from melodies in mind to the resources on the web," in *Proceedings of the 13th international conference on humans and computers*. University of Aizu Press, 2010, pp. 152–158.

[23] K. Bischoff, C. S. Firan, and R. Paiu, "Deriving music theme annotations from user tags," in *Proceedings of the WWW 2009*, 2009.

User Annotations as a Context for Related Document Search on the Web and Digital Libraries

Jakub Ševcech, Róbert Móro, Michal Holub and Mária Bieliková
Faculty of Informatics and Information Technologies, Slovak University of Technology
Ilkovičova 2, 842 16 Bratislava, Slovakia
E-mail: {jakub.sevcech, robert.moro, michal.holub, maria.bielikova}@stuba.sk

Keywords: annotation, related document, search, query construction

Received: November 23, 2013

In this digital age, a lot of documents that people read are accessed through the Web and read on-line. There are various applications and services which enable creating bookmarks, tags, highlights and other types of annotations while reading these electronic documents. Annotations represent additional information on a particular information source and indicate that documents or their sections are somehow interesting for the document reader. However, existing approaches lack immediate reward for content annotation. We propose a method for query construction enabling search for other documents related to the currently studied one using not only the document's content, but also user created annotations as indicators of user's interests. In our proposed approach, annotations are used to activate nodes in a graph created from the document's content employing spreading activation algorithm. We evaluate the proposed method in Annota - a service for bookmarking and collaborative annotation of Web pages and PDF documents displayed in a web browser. Along with its main purpose, Annota is designed to support scenarios useful for a novice researcher working together with his or her mentor. Based on Annota usage data we also analyzed properties of various types of annotations. Discovered annotation properties served as a basis for simulation we performed to determine optimal parameters of the query construction. We compared the proposed method to the commonly used tf-idf based method which our method outperformed when using annotations in the query construction process by improving the overall precision of the document retrieval. Therefore, annotations proved to be a viable source of information for user's interest detection.

Povzetek: Razvita je metoda, ki pri delu s spletnimi besedili uporablja oznake v besedilu.

1 Introduction

While reading printed documents, a common practice is to write down various types of notes. We use them as means of storing our thoughts, to highlight interesting parts of the document and to ease navigation in the printed document. Many tools and services allow us to create similar notes in electronic documents as well. We can create various bookmarks, tags, highlights and other types of annotations while surfing the Web or when reading electronic documents. In contrast to notes written in printed documents, electronic annotations are often objects of further processing and they can serve as means of improving intra and inter document navigation, to organize personal collections of documents, to search for documents, etc.

There is active research in the field of utilization of annotations [1] and patterns [2] their users follow when creating or making use of these annotations. Various types of annotations can be used for user interest identification [3], user modeling [4] and subsequently for personalization or additional support while searching for resources.

Annotations, created by user, can be considered a form of user's context he or she creates while reading documents and traveling in digital space [5]. This context can take various forms depending on the used annotation type, such as thoughts stored as short notes attached to the document as a whole, comments to specific sections of the document, or highlighted document sections that are in some way interesting to the reader. Many applications use annotations as a means of navigation between documents and for organizing content. For example, in [6] the authors describe an organization of learning materials and collaboration of students while learning to use an educational system that provides students the possibility to attach various types of annotations to learning objects. The study of various search tasks supported by a social bookmarking service

[*] This paper is based on J. Ševcech and M. Bieliková, *Query Construction for Related Document Search Based on User Annotations* published in the proceedings of the 3rd International Workshop on Advances in Semantic Information Retrieval (part of the FedCSIS'2013 conference).

deployed in a large enterprise is presented in [7]. The authors concluded that bookmarking services and annotations attached to documents can enhance document organization and social navigation.

User generated tags are one of the most commonly used methods for organizing the content, because of their utility and applicability for various content types. They have been successfully used to organize various media files, e.g. photos, videos, and documents in many real world applications such as bookmarking services Diigo[1] or Delicious[2]. Further types of annotations, such as highlights and comments can serve to create custom in-document navigation. The users can categorize or describe resources [8] and thus create navigation that fits their needs without relying on navigation provided by document's author.

User created annotations can be used not only to support inter or intra document navigation. Tags are used for folksonomy construction [9], annotations can play an important role in content enrichment and content quality improvement, e.g. in an educational system, as presented in [6]. In this system the authors use content error reports, user generated comments and questions, to improve course content and other types of annotations, such as tags and highlights, for the navigation and even the content summarization [10].

Currently, there are many services allowing users to annotate the documents. However, all of these applications motivate users to create annotations by a prospect of future improvement of inter or intra document navigation, i.e. users benefit from created annotations only after there are enough annotated documents, or when returning to previously annotated document. Problem with this approach is that there is a lack of immediate reward after annotation is created.

The rest of the paper is structured as follows. In section 2 we further analyze different approaches for utilization of annotations in the search process. Annota - a service for web page bookmarking and annotation, that allows users to insert various types of annotations to Web pages and PDF documents displayed in Web browser, is introduced in section 3. We describe multiple applications and usage scenarios that are supported by annotations the users attach to documents emphasizing Annota's unique features compared to existing similar systems. In section 4 we propose a method for query construction from currently studied document and its attached annotations as one of document annotation applications. This method produces a query that can be used in related document retrieval where the query is taking into account user's interest provided by created annotations. The query is created while the user is reading the documents and it is used to search for related documents to the currently studied one. The reward for user creating annotations is thus provided during the time of annotation creation. We evaluate the proposed method using synthetic as well as online experiments in the Annota system in section 5 and conclude by discussing the method's properties and implications for the area of research in section 6.

2 Related work

One of the possible employments of annotations in information processing is the document search. There are two possible approaches for exploitation of annotations in the search process. One is to use annotations while indexing documents by expanding documents in a similar way anchor texts are used [11], or using bookmarks and annotations as document quality indicators while ranking documents [12].

The second possible application of annotations is in the query expansion or in query construction process. An example of annotations used for query expansion is presented in [13], where tags attached to search results are used to expand initial query similarly to pseudo-relevance feedback. Multiple methods for query expansion in folksonomies are presented in [14]. Of particular interest are methods expanding queries by tags from folksonomies on the basis of semantic similarity between words of the query and these tags.

An example of annotations used as queries to retrieve related documents is presented in [15]. The authors asked users to read a set of documents and to create annotations in documents using a tablet. They used these annotations as queries in related document search. They compared search precision of these queries with relevance feedback expanded queries. Queries derived from user's annotations produced significantly better results than relevance feedback queries.

More often, when creating queries for related document retrieval, the document's content is used instead of attached annotations. In [16] authors used the most important phrases from the source document as queries for document retrieval. Another work dealing with search for related documents is described in [17] where the authors use related document search as a means of recommendation of citations into unpublished manuscripts. They use text-based features of the document to retrieve similar documents and citation features to establish authority of documents. Similar document retrieval has also its application in document recommendation. In work presented in [18] a list of documents similar to those visited by the users were used as a form of content based recommendation of related documents.

In popular search engines such as ElasticSearch[3] and Apache Solr[4], term frequency is used in the query construction process. They provide special type of query interface called "more like this" query, which processes source text and returns a list of similar documents. Internally, the search engine extracts the most important words using tf-idf metric from source text and it uses the

[1] Diigo, http://www.diigo.com/
[2] Delicious, https://delicious.com/
[3] ElasticSearch, http://www.elasticsearch.org/
[4] Apache Solr, http://lucene.apache.org/solr/

most important words as a query for related documents search.

In most applications, the similar document retrieval process consists of two phases. In the first step, queries in the form of the most important phrases and, more often, the most important terms are extracted from the document's content. In the second step, these queries are used to retrieve other documents. In order to retrieve these most important terms from document's content, many different methods are used. Mostly, they are based solely on the term frequency in the document (such as already mentioned tf-idf based method) but many other methods are applicable. One possible category of methods for query term extraction are methods based on ATR (automatic term recognition) algorithms [19].

In multiple works authors showed that annotations represent important source of information for document retrieval. Methods for query construction for document retrieval however, use only document's content and information about the document collection in query construction process. They do not utilize user created annotations as user's interest indicators when creating query for document retrieval. We believe that annotations used in query construction process can significantly improve related document retrieval precision.

In our work we propose and evaluate a method for query construction from the document content enhanced by user created annotations. Annotations are used as interest indicators to determine parts of the document the user is most interested in. Using user created annotations our method creates a keyword query for related document search taking into account the user's interests. Proposed method is used in social bookmarking service Annota to retrieve related documents to the currently studied document. Annotations are used in related document retrieval in time of their creation and they provide immediate motivation for additional annotation creation in the form of related document search.

3 Service for Web page annotation

We developed a service called Annota[5] [20], which allows users to attach annotations to arbitrary web pages or PDF documents displayed in a web browser. Annota was created as a system to study methods for document search, navigation and organization on the Web. We uniquely employ annotations created by users in various methods of information retrieval, especially in digital libraries. In this domain, Annota supports various scenarios of collaboration: between a novice researcher and his or her supervisor (mentor), or between more researchers working on a joint project.

A few projects for supporting researchers already exist. Mendeley[6] allows users to organize and annotate documents via a desktop application and web interface. ResearchGate[7] is specialized to connect researchers while allowing them to add their own publications, follow others and ask research-related questions.

Annota provides environment to collaboratively collect documents while attaching annotations to them. Annota's unique features include annotation of documents directly on the Web as well as support for collaborative features such as bookmark sharing within groups and following other users of the service. Annota is realized as a client-server system. Client is represented by a browser extension allowing annotation of web pages. Annotations are stored on the server together with the identification of the resource (its URL) and additional metadata. The browser extension allows users to create various types of annotations, such as:

- tags,
- highlights,
- comments attached to selected text, and
- notes attached to the document as a whole.

Although Annota can be used on every web page, our target domain are digital libraries used by researchers in the field of information technologies, for which we provide additional support and tools. Annota stores metadata on various entities from digital libraries (authors, papers, conferences, etc.). We get this information by parsing web pages of selected digital libraries the users of Annota visit. When a user bookmarks a page containing metadata about a paper, Annota creates bibliographic reference to it. We realized the possibility to insert annotations into arbitrary web pages, articles in digital libraries and PDF documents displayed in web browser, by bookmarking and sharing documents and annotations.

The Annota service allows users to organize documents by tags, folders or faceted trees. It is possible to search in texts of documents contained in the user's library or in the library of bookmarked documents of all users. Besides keyword search, Annota offers various means of information space exploration, such as cloud of important terms, content of which is adapted by users' navigation history, i.e. by their previous queries [21], or navigation leads in the search results' summaries.

An example of a web page annotated using Annota is displayed in Figure 1. The figure shows a widget, where it is possible to bookmark displayed page, insert tags, edit note and share the bookmark with groups the user is member of. Users are able to highlight text fragments of the web page and to attach comments to these text selections.

The basic scenario of the service usage follows a user studying a document. The user has the following possibilities for particular activities:

- Bookmarking documents.
- Organizing the collection of documents using tags attached to individual bookmarks.
- Organizing the collection of documents by inserting the bookmarks into folders.
- Highlighting parts of the text and creating other types of annotations.
- Sharing bookmarked document in a group the user is member of via group sharing feature.

[5] Annota, http://annota.fiit.stuba.sk/

[6] Mendeley, http://www.mendeley.com/

[7] ResearchGate, http://www.researchgate.net/

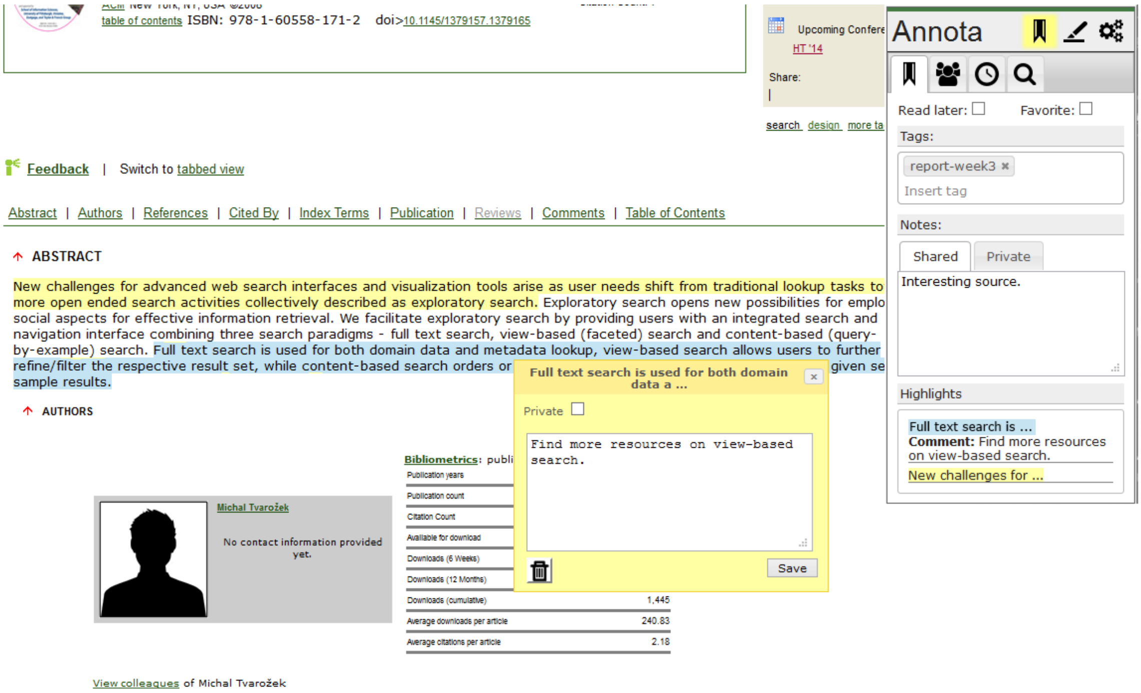


table of contents ISBN: 978-1-60558-171-2 doi>10.1145/1379157.1379165

ABSTRACT

New challenges for advanced web search interfaces and visualization tools arise as user needs shift from traditional lookup tasks to more open ended search activities collectively described as exploratory search. Exploratory search opens new possibilities for emplo social aspects for effective information retrieval. We facilitate exploratory search by providing users with an integrated search and navigation interface combining three search paradigms - full text search, view-based (faceted) search and content-based (query-by-example) search. Full text search is used for both domain data and metadata lookup, view-based search allows users to further refine/filter the respective result set, while content-based search orders or given se sample results.

Full text search is used for both domain data a ...

Private □

Find more resources on view-based search.

AUTHORS

Michal Tvarožek

No contact information provided yet.

Bibliometrics: publi

Publication years	
Publication count	
Citation Count	
Available for download	
Downloads (6 Weeks)	
Downloads (12 Months)	
Downloads (cumulative)	1,445
Average downloads per article	240.83
Average citations per article	2.18

View colleagues of Michal Tvarožek

Annota

Read later: □ Favorite: □

Tags:

report-week3 ×

Notes:

Interesting source.

Highlights

Full text search is ...
Comment: Find more resources on view-based search.

New challenges for ...

Upcoming Confere
HT '14
Share:
search design more ta
Feedback | Switch to tabbed view
Abstract | Authors | References | Cited By | Index Terms | Publication | Reviews | Comments | Table of Contents
×
Save
Insert tag
Shared
Private

Figure 1: Web page in ACM DL annotated using bookmarking service Annota. It can be annotated collaboratively (highlights from different users are displayed in different colors).

- Following activity of other interesting users.

3.1 Annotation usage scenarios

Previously mentioned features are useful when a user works alone. However, research nowadays is being done by teams of collaborating people, sometimes composed by only two researchers (researcher novice and his supervisor or mentor), other times the teams are larger. In order to support collaboration of researchers within such teams, we support several scenarios of using Annota, namely:

- novice researcher scenario,
- paper authors scenario, and
- activity following scenario.

3.1.1 Novice researcher scenario

The novice researchers working on their projects obviously start by doing research on the state of the art in the research area of their interest. They usually read a lot of research articles, some of which are more useful and relevant to the target research topic than others. The researchers need tools to keep them organized in order to reference them later in their work. Moreover, the novice researchers need help from their respective mentors, who are expected to recommend useful resources their mentee should read.

Annota helps the novice researchers to organize the resources they read (using folders or tags) and to annotate the research papers. The researchers can use their own notes later, while writing the papers or preparing presentations.

Annota also allows the supervisors to create a group and invite the researcher they supervise to become a member of it. Then, the supervisors can share papers via this group, thus recommending important study material to their mentees. This is very helpful and thanks to that the novice researchers have a point from which to start searching for more information on their research topic.

Working in groups also enables the novice researchers to report the progress to their supervisor e.g. by using specialized tags (report-week3 for third week in semester as can be seen in Figure 1). Apart from one group per researcher, the supervisor might also pick the tactics of creating a group for all his mentees who share similar research topics. The users then share interesting research articles they have found together with their annotations and comments. The supervisors can add their own notes and help distinguish relevant publications or propose further readings.

In order to help the researchers with finding relevant papers, Annota allows them to search for papers already bookmarked by others. Since they also assign tags to the resources, the researcher might find an interesting paper on a certain topic easier than using only the search features provided by the digital library.

Annota can generate a report for the supervisor showing the activity of the group (or per user) for a selected period of time, containing overview of shared papers together with annotations. This allows the supervisors to continuously monitor the progress of the researchers they manage and effectively help them, a feature which is unique to Annota.

3.1.2 Paper authors scenario

In this scenario we consider a group of researchers doing research together. Part of every research is studying the work already done in the respective field. Collaborating researchers form a group in Annota and they can share interesting publications with each other, comment and annotate them. These annotations can be later used when the researchers need to write a paper about their results, especially the “Related work” section.

A group in Annota needs not to be private. On the contrary, we encourage the groups to be public so that other users of Annota can see interesting resources the group has found together with their opinions. We allow every group to formulate its research goals in the form of tags (similarly to tags attached to documents). Users can find a group of their interest based on these tags.

3.1.3 Activity following scenario

We realize that collaboration and sharing of thoughts is very important for researcher in any field. Therefore, in Annota we allow its users to form social networks by following each other (a concept known mostly from Twitter[8]). When user A follows user B, the user A can see user B’s newly added bookmarks and annotations, as well as other activities (joining of a group, following another user). When user A considers user B to be an authority in a field of his or her interest, this can keep the user A informed about the latest trends.

We do not limit the ability to follow someone just to Annota users. Since we gather freely available metadata about publications from various digital libraries, we allow the user to follow researchers, who are not Annota users. Furthermore, we allow them to follow interesting conferences, journals or publishers. This way the users are notified when their favorite researcher publishes a new paper, new issue of a journal or proceedings of their favorite conference are published, etc.

Moreover, the users of Annota can also follow the whole group, which enables them to see their newly added information. We believe the feature of following various entities (people, groups, publications, etc.) allows the whole community to grow and learn from each other and is an important feature to keep informed about the latest trends. Naturally, all the activities of the Annota users can be set to be private if they wish to keep their privacy. In such a case, nobody (not even the followers) sees them.

3.2 Creation of the Web page annotations

The browser extension created as a part of Annota service allows users to create annotations that link to document as a whole (tags, note) or to particular parts of the document (highlight, comment). As the extension is inserting annotations into web pages and they change frequently and without notification, we had to use a method for annotation linking to specified parts of the document that is resistant to changes in annotated document.

The key element in document annotation is the selection of a method to link documents and created annotations. Multiple systems supporting annotation creation assume that documents will not change after annotations are inserted. This is very strong assumption we cannot make in a domain such as web pages. We have to use method for annotation interlinking with document’s content with regard to documents which may change over time. In [22] multiple criteria, which must be met by a robust method for locating annotations into documents, are defined. Some of the described criteria are:

- The method has to be robust to common changes in the referenced document.
- It has to be based on document’s content.
- It has to work with uncooperative servers.
- The information necessary to locate annotation have to be relatively small compared to the document’s content.

At the same time, in this work the authors suggest several approaches that meet these criteria. One of them is to use annotation context in form of surrounding text to place the annotation into the document. The method using document content to place annotations is defined also in Open Annotation Model [23]. It is tolerant to changes in the document content and when using approximate matching of strings it is also tolerant (to some extent) to changes in annotation context as well.

In order to attach annotations to document parts we use redundant representation of annotation location to support linking annotations into changing documents and to improve stability of annotation location. For locating annotation in the text, we store highlighted text with order of its in-text occurrence together with surrounding text. The combination of selected text and text occurrence order is tolerant to changes in the document’s content except for changes in selected text and some changes before annotation location. With usage of approximate matching this method is to some extent tolerant to changes in selected text as well.

4 Method for query construction

Currently, the most common form of query used when searching for documents on the Web is the list of keywords. That is why the majority of methods for document retrieval using source document as query is transforming the document content into keyword queries. In order to retrieve words from the document to be used as query for related document search it is possible to use multiple different approaches. One of them is to extract most frequently occurring terms using the tf-idf metric or various ATR algorithms [19] as discussed in section 2. The tf-idf based method provides rather straightforward possibility to incorporate user created annotations: the source text of the document is extended by the content of created annotations, possibly with various weights for different types of annotations.

However, the method using the tf-idf for query word extraction takes into account only the number of occurrences of words in the source document (and document corpus). We believe that not only the number of word occurrences but also the structure of the source text is important when constructing a query for related

[8] Twitter, https://twitter.com

documents retrieval. Especially, if we suppose that while reading the document the users are usually interested in only a fraction of the document, this fraction is the place where they most probably attach an annotation.

We use user created annotations to increase weights of annotated parts of the document in query construction process and to attach additional content to the document. We proposed a method based on spreading activation in text of studied document transformed to a graph. The method uses annotations as interest indicators to extract parts of documents the user is most interested in. The proposed method consists of two phases:

1. Text to graph transformation that conserves word occurrence frequency in node degree and text structure in graph edges structure.
2. Graph nodes activation introduced by annotations attached to the document and query word extraction using spreading activation algorithm in created graph.

The text to graph transformation is based on word neighborhood in the text. The graph created from the text using words neighborhood conserves words importance in node degree, but it also reflects the structure of the source text in the structure of edges [23]. Such graph can be used for example the most important terms [24]. We use this graph to extract words are most important from point of view of the document reader and we use them as queries to retrieve similar documents.

4.1 Text to graph transformation

In order to transform text to a graph, it is first processed in several steps: segmentation, tokenization, stop-words removal and stemming. After these steps the initial text has a form of a list of words. Every unique word from this list is transformed into a single node of a graph. The edges of the graph are then created between two nodes if corresponding words in the text are neighbors or they are in the predefined maximal distance. This transformation is described by the following pseudocode:

```
tokens = text.downcase().split()
words = tokens.removeStopwords().stem()
length = words.size
nodes = words.uniq
edges = []
for(i=0;i<length;i++){
  for(j=i;i<min(i+dist,length-1);j++){
    edges.add(words[i], words[j])
  }   }
graph = Graph.new(nodes,edges)
```

As settings for maximal distance between words we used options described in [24], where two passages through the text with maximal distance set to two words and five words are used. By using these setting, the words with greater distance were connected and at the same time close words are better connected by bigger number of common edges. Created edges have the same weight but to speed up spreading activation process, we connected multiple edges between the same nodes and we set weights of the resulting edges to the number of connected edges.

4.2 Query word extraction

We use the graph representation of the text in order to find the most important nodes/words using spreading activation algorithm. This algorithm is commonly used for example to find the most related nodes in a graph to the initially activated node. The activation introduced into the initial node is spreading through the nodes of the graph and after the change in nodes activation is smaller than a specified threshold, the greatest amount of activation is concentrated in the most related nodes.

It is possible to use this algorithm for related nodes search, but also for other applications, such as keyword extraction [25]. We use the same intuition to extract the most important words to sections user is most interested in. We utilized user created annotations as their interest indicators. These annotations are used to introduce initial activation to nodes annotations are attached to. The initial activation is propagating through the graph and it is concentrating in most important words of the text. An example visualization of text transformed to graph and nodes activated by attached annotations is displayed on Figure 2. The node size reflects activation level and edge thickness number of edges between nodes. Colored nodes represents nodes with highest activation level, thus words selected as query for related document search.

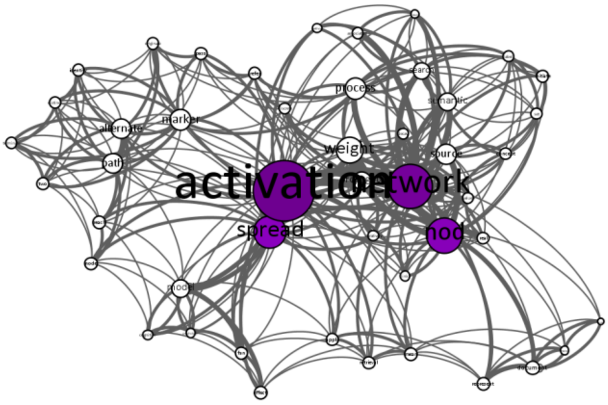


Figure 2: Example of text transformed to graph with activation spread across nodes.

When using annotations to insert initial activation into the document graph we consider separately annotations that are:

- highlighting parts of the document and
- inserting additional content into the document.

The proposed method takes into account both types. Those, which highlight parts of the document, contribute by activation to nodes representing words of highlighted part of the document and those enriching content of the document are extending the document graph by adding new nodes and edges and they are inserting activation to this extended part of the graph. When inserting activation to extended parts of the document we assume that some portion of the words used in the annotation content are located in the document text as well. The activation from the extended part of the graph can then pass to the rest of the graph through common nodes.

When initial activation is spreading through the created graph, the nodes where activation is concentrating are the most important words of the graph and are considered words fit for the query. In our case, the activation is inserted into the graph through annotations attached to the document by its reader.

The proposed method is able to extract words, which are important for annotated part of the document, but it is also able to extract globally important words, that are important for document as a whole. The portion of locally and globally important words can be controlled by the number of iterations of the algorithm. With increasing number of iterations the activation is spreading from activated part of the document and extracted locally important words are changed to globally important words. When using this method it is thus important to determine when to stop the algorithm to find the best portion of globally and locally important words. It is also important to determine the right amount of activation inserted into the graph by various types of annotations. We determine these settings using simulation based on real user data while evaluating proposed method. The simulation is described in the next section of this paper.

The method for query word extraction uses annotations to insert initial activation into text transformed to graph. In case when no annotations are attached to the document, it is possible to extract globally important words from the document by activating the whole document's text.

5 Evaluation

In order to evaluate related document retrieval we performed both synthetic tests on dataset extracted from Wikipedia articles and online experiment with users of Annota bookmarking service.

5.1 Related document retrieval

We analyzed behavior of users of Annota while annotating documents using browser extension. Our experiments are based on usage data of 82 users who created 1 416 bookmarks and 399 in-text annotations during 4 months long period of using Annota on day-to-day basis. We studied multiple parameters of created annotations and we derived probabilistic distributions of these parameters. We studied properties such as the note length, number of highlights per user and per document, highlighted text length or probability of comment to be attached to highlighted text. We used extracted annotations properties and knowledge about their distributions in further evaluation. All observed parameters were following logarithmic or geometric distributions. Figure 3 displays an example of derived distribution for number of highlighted texts per document that follows logarithmic distribution.

Using various attributes of annotations and their probabilistic distributions we created a simulation, to find optimal weights for various types of annotations and number of iterations of proposed method for query construction from document text and attached annotations. We optimized query construction for document search precision.

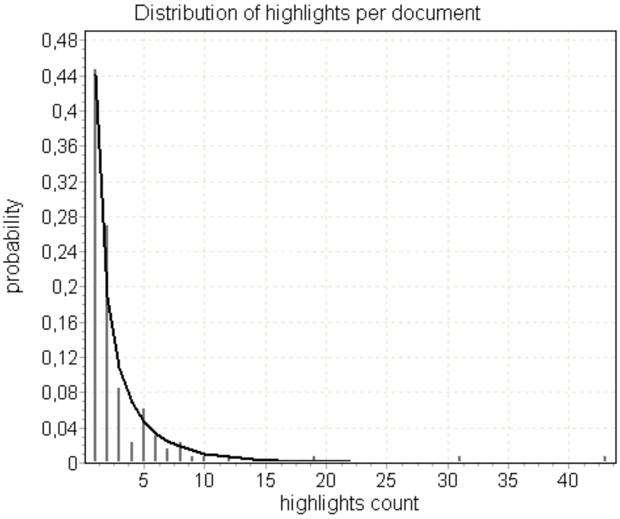


Figure 3: Logarithmic distribution of highlighted texts number per document.

The simulation was performed on the dataset we created by extracting documents from Wikipedia articles written in English. We constructed the source documents with aim to create documents containing several similar sections (from the point of view of used words) and with different topics. These generated documents simulate documents, where the user is interested in only a fraction of the content. In order to create such documents we used disambiguation pages in Wikipedia. The disambiguation page disambiguates multiple meanings of the same word and contains links to pages for each of these meanings. By combining abstracts of pages describing different meanings of the same word into single document, we simulate sections of the text describing multiple topics.

We downloaded all disambiguation pages and we selected random subset of these pages for which we downloaded pages they are linking to. Along with these disambiguated documents we downloaded all documents, having common category with at least one of disambiguated documents. We used search engine ElasticSearch to create an index of all downloaded documents and to search within this index. The parameters of created dataset are summarized in Table 1.

Table 1: Parameters of dataset used in simulation

Attribute	Number
All disambiguation pages	226 363
Selected disambiguation pages	86
Pages disambiguation pages are linking to	629
Categories	2 654
All downloaded pages	232 642

In the simulation we generated annotations in a way to correspond with probabilistic distributions extracted from the annotations created by users of the Annota service. From every disambiguation page and the pages it was linking to, we created one source document by combining abstracts of all pages in random order. For every source document we selected one of the composing abstracts, which simulated one topic user is most interested in. We generated both annotations highlighting parts of the document and annotations inserting additional content for selected abstract. The highlights were randomly distributed over the whole abstract. To simulate the content of annotations extending content of the document (notes, comments) we used random parts of the page annotated abstract was extracted from.

Generated annotations along with source document content were used to create query using proposed method for query construction. The created query was used for related documents search in the index of all downloaded documents. When evaluating relevance of retrieved documents, we considered document to be relevant if it was from the same category as the page annotated abstract was extracted from.

We performed a simulation with several combinations of parameters and we implemented hill climbing algorithm to optimize parameter weights combination for the highest document search precision. Single iteration of performed simulation is described by following pseudocode:

```
for disambig in disambiguations do
  abstracts = disambig.pages.abstracts
  for abstract in abstracts do
    text = abstracts.shuffle.join(" ")
    graph = Graph.new(text)
    annot = Annotation.create(abstract)
    graph.activate(annot, weights)
    graph.spreadActivation()
    query = graph.topNodes
    results = ElasticSearch(query)
    cat = abstract.page.categories
    relevant = results.withCategory(cat)
  end
end
```

We compared search precision for proposed method and for tf-idf based method (“more like this” query) provided by ElasticSearch when searching for 10 most relevant documents. For the purpose of comparison of the proposed method with method based on tf-idf when using annotations in the query construction process, we performed an extension of the tf-idf based method to use annotations in query word extraction process. This method uses word frequency to find the most important words in the text. We extended the text of the document by text annotations were attached to and annotations content. We provided different weights for different annotations types by repeated extension of document by highlighted text and annotations content. We determined the optimal number of repetitions using parameter optimization with hill climbing algorithm, similarly to simulation for parameter estimation for method based on spreading activation in text transformed to graph.

Along with simulation using generated annotations for methods comparison, we performed two experiments to determine retrieval precision with no annotations and when whole abstract of the source document was highlighted. We aimed to determine the precision of compared methods when no annotations are available and while having complete information about user’s interests. Results for simulations with generated annotations along with experiments with no annotations and with whole document fragment annotated are summarized in Table 2.

Table 2: Simulation results for spreading activation based method and tf-idf based method.

Method	Precision
Tf-idf based with no annotations	21.32%
Proposed with no annotations	21.96%
Tf-idf based with generated annotations	33.64%
Proposed with generated annotations	37.07%
Tf-idf based with whole fragment annotated	43.20%
Proposed with whole fragment annotated	53.34%

Proposed method based on spreading activation obtained similar or better results to tf-idf based method in all performed experiments. The results of experiments with no annotations, where only the content of the document was used to create query, suggests that proposed method provides similar, even better results for query word extraction. These results were achieved despite the fact that proposed method is using only information from the document content and not the information about other documents in the collection by contrast to tf-idf based method. The proposed method can thus be used as an alternative to tf-idf based method when creating query from document content.

The experiments with generated annotations and whole text fragment annotated suggests that proposed method outperforms tf-idf based method when annotations are used in query construction process. We performed a Student’s t-test on 5% level of significance for pairs of proposed method and baseline method which showed statistically significant difference in mean precisions for compared methods when using generated annotations and whole abstracts annotated in query construction process (p-value < 0.01%).

The comparison of both methods without using annotations and using generated annotations in query construction process proved that annotations can increase precision of related documents retrieval. The experiment with whole document fragments annotated suggests that with increasing number of annotations the precision of generated queries increases for both used methods.

5.2 Online experiment in Annota bookmarking service

In order to compare the real increase of precision of document retrieval method with and without annotations we performed a qualitative user study where 8 volunteers were asked to annotate documents of their choice stored in Annota. Afterwards, we generated two queries, one with and another one without annotations. We retrieved two lists of documents using these queries and we presented them to volunteers in random order. They were asked to select documents describing a topic related to the topic of source document from displayed lists and to select more relevant from two presented lists.

The volunteers annotated 11 unique documents. In 9 cases they selected the list created by the method using annotations as more relevant one. In one case the method using annotations created query in Slovak we found no relevant documents. This was caused by the fact that in this document all annotations were written in Slovak and all documents we searched in were in English. In one case the method not using annotations obtained better results. We obtained 34 relevant documents using method with annotations compared to only 15 documents returned by method without annotations.

Part of volunteers were writing annotations in Slovak, but to keep conditions the same as during document annotation out of the experiment, we allowed them to write annotations the same way they are used to. We asked one user to repeat the experiment on one document after he translated created annotations written in Slovak to English. When translated annotations were used in query construction all retrieved results were related to the source document.

In one case we asked the volunteer to repeat the experiment with increased number of annotations attached to the document. During this repeated experiment, the volunteer doubled the number of attached annotations. In the second retrieved list of documents, the number of relevant documents retrieved increased and the list included one exact match with the topic user was most interested in. With increasing number of annotations attached to document the precision of related document retrieval increases.

When using annotations to create a query, the proposed method retrieved more relevant documents in greater number than in the case when annotations were not used. Also, using annotations in to create queries, we retrieved more documents describing the same, as well as related topic as the source document.

We used a questionnaire about user's habits when annotating documents to determine how users of Annota are creating annotations into studied documents. The majority of participants are using annotations while reading printed or electronic documents. When annotating electronic documents, they use various tools to create bookmarks, to-do lists, saving documents for later, to insert highlights, comments and other types of annotations into documents. The most frequently used types of annotations are tags and in-text highlights. The purpose of creating annotations such as notes, comments and highlights is to summarize studied documents, describe documents, highlight most important sections, and store their thoughts about studied documents and as a form of in-document navigation to support fast recollection of document when returning to previously studied document. Interviewed volunteers confirmed our assumption that using annotations users are indicating those parts of the document they are most interested in.

6 Conclusions

Annotations represent a significant source of information on interesting or important parts of the documents. Their importance increases with possibilities for manipulating documents on the Web in the same way as printed documents and with possibilities for further processing and utilizing of the created annotations. We introduced Annota – a service for bookmarking and annotating Web documents while focusing on the domain of digital libraries. We described several scenarios where annotations can be useful. We studied users' behavior while annotating documents on the Web and proposed a method for query construction from document's content and attached annotations. For this purpose we considered document's content and its structure by using text to graph transformation and query terms extraction using spreading activation introduced by attached annotations.

The simulation based on probabilistic distributions of various parameters of annotations created by the users of Annota proved, that using annotations when creating queries for related document retrieval can increase retrieval precision and with increasing number of attached annotations the precision rises.

We compared two methods for query word extraction. The method based on spreading activation in document text transformed to graph outperforms tf-idf based method when creating query for related documents search from source document and attached annotations. The proposed method achieved comparable results to tf-idf based method when no annotations were used in the process of query construction. It is thus possible to use it even when no annotations are attached to the document with comparable precision as commonly used method when extracting words suitable for query for related document retrieval. The spreading activation based method outperformed baseline method when annotations attached to documents were used in query construction process. The proposed method does not use information from other documents, only information from the content of the source document and its attached annotations. It is thus search engine independent and can be used to create queries for any search engine accepting queries in the form of a list of keywords.

Performed user study showed that users insert annotations into document sections they are most interested in and they use annotations to summarize documents, highlight most important parts of documents and to store their thoughts. In connection with comparison of related document retrieval precision of proposed method and commonly used method when using annotations in query construction process and when no annotations were used, we showed that annotations can be used as user interest indicators in query construction for related document retrieval and that they improve related document retrieval precision.

Acknowledgement

This work was partially supported by the projects VG1/0675/11, VG1/0971/11 and APVV-0208-10. The authors wish to thank other members of Annota team, namely Roman Burger, Martin Lipták, Juraj Kostolanský, Peter Macko and Samuel Molnár for their contribution to design and implementation of selected components of Annota.

References

[1] M. Agosti, N. Ferro (2007). A formal model of annotations of digital content. *ACM Trans. Inf. Syst.*, vol. 26, no. 1.

[2] S.A. Golder, B.A. Huberman (2006). Usage patterns of collaborative tagging systems." *Journal of Information Science*, vol. 32, no. 2, pp. 198-208.

[3] X. Wu, L. Zhang, Y. Yu (2006). Exploring social annotations for the semantic web. *Proc. of the 15th Int. Conf. on World Wide Web (WWW '06)*, ACM, pp. 417-426.

[4] R. Wetzker, C. Zimmermann, C. Bauckhage, S. Albayrak (2010). I tag, you tag: translating tags for advanced user models. *Proc. of the 3rd ACM Int. Conf on Web Search and Data Mining (WSDM '10)*, ACM, pp. 71-80.

[5] P. Návrat (2012). Cognitive traveling in digital space: from keyword search through exploratory information seeking. *Central European Journal of Computer Science,* vol. 2, no. 3, pp. 170-182.

[6] M. Šimko, M. Barla, V. Mihál, M. Unčík, M. Bieliková (2011). Supporting Collaborative Web-based Education via Annotations. *World Conf. on Educational Multimedia, Hypermedia and Telecommunications*, pp.2576-2585.

[7] D. Millen, M. Yang, S. Whittaker, J. Feinberg (2007). Social bookmarking and exploratory search. *ECSCW 2007*, Springer, London, pp. 21–40.

[8] C. Körner, R. Kern, H. P. Grahsl, M. Strohmaier (2010). Of categorizers and describers: An evaluation of quantitative measures for tagging motivation. *Proc. of the 21st ACM Conf. on Hypertext and Hypermedia*, ACM, pp. 157-166.

[9] C. Cattuto, C. Schmitz, A. Baldassarri, et al. (2007). Network properties of folksonomies. *AI Comm.*, vol. 20, no. 4, pp. 245-262.

[10] R. Móro, M. Bieliková (2012). Personalized Text Summarization Based on Important Terms Identification. *23rd Int. Workshop on Database and Expert Systems Applications*, IEEE, pp. 131–135.

[11] X. Zhang, L. Yang, X. Wu, et al. (2009). sDoc: exploring social wisdom for document enhancement in web mining. *Proc. of the 18th ACM conf. on Inf. and knowledge management*, ACM, pp. 395–404.

[12] Y. Yanbe, A. Jatowt, S. Nakamura, K. Tanaka (2007). Can social bookmarking enhance search in the web? *Proc. of the 7th ACM/ IEEE-CS joint conf. on Digital libraries*, ACM, pp. 107–116.

[13] C. Biancalana, A. Micarelli (2009). Social tagging in query expansion: A new way for personalized web search. *Computational Science and Engineering*, vol. 4. IEEE, pp. 1060-1065.

[14] R. Abbasi (2011). Query expansion in folksonomies. *Semantic Multimedia*, Springer Berlin Heidelberg, pp. 1-16.

[15] G. Golovchinsky, M.N. Price, B.N. Schilit (1999). From reading to retrieval: freeform ink annotations as queries. *SIGCHI Bulletin*. ACM Press, 1999, pp. 19–25.

[16] Y. Yang, N. Bansal, W. Dakka, et al. (2009). Query by document. *Proc. of the 2nd ACM Int. Conf. on Web Search and Data Mining (WSDM '09)*, ACM, pp. 34–43.

[17] T. Strohman, W. B. Croft, D. Jensen (2007). Recommending Citations for Academic Papers. *Proc. of the 30th Annual Int. SIGIR Conf. on Research and Development in Inf. Retrieval*, ACM, pp. 5–6.

[18] M. Kompan, M. Bieliková (2010). Content-based News Recommendation. *E-Commerce and Web Technologies*, Lecture Notes in Business Information Processing, vol. 61, part 2, Springer, pp.61-72.

[19] Z. Zhang, J. Iria, C. A. Brewster, F. Ciravegna (2008). A comparative evaluation of term recognition algorithms. *Proc. of 6th Int. Conf. on Language Resources and Evaluation*, Marrakech Morocco.

[20] J. Ševcech, M. Bieliková, R. Burger, M. Barla (2012). Logging activity of researchers in digital library enhanced by annotations. *Proc. of 7th Workshop on Int. and Knowledge oriented Tech.*, pp. 197-200. (in Slovak)

[21] S. Molnár, R. Móro, M. Bieliková (2013). Trending words in digital library for term cloud-based navigation. *Proc. of the 8th Int. Workshop on Semantic and Social Media Adaptation and Personalization (SMAP '13)*, IEEE CS, to appear.

[22] T. A. Phelps, R. Wilensky (2000). Robust intra-document locations. *Computer Networks*, vol. 33, no. 1, pp. 105-118.

[23] P. Ciccarese, M. Ocana, L. J. Garcia Castro, S. Das, T. Clark (2011). An open annotation ontology for science on Web 3.0. Journal of Biomedical Semantics, vol. 2, no. 2.

[24] D. Paranyushkin (2011). Visualization of Text's Polysingularity Using Network Analysis. *Prototype Letters*, vol. 2, no. 3, pp. 256–278.

[25] G. K. Palshikar (2007). Keyword extraction from a single document using centrality measures. *Pattern Recognition and Machine Intelligence*, Springer Berlin Heidelberg, pp. 503-510.

SOAROAD: An Ontology of Architectural Decisions Supporting Assessment of Service Oriented Architectures

Piotr Szwed, Pawel Skrzyński, Grzegorz Rogus and Jan Werewka
Department of Applied Computer Science
AGH University of Science and Technology
al. Mickiewicza 30, 30-059 Kraków, Poland
E-mail: {pszwed, skrzynia, rogus, werewka}@agh.edu.pl

Keywords: software architecture, ontology, SOA, ATAM, architecture assessment, architecture evaluation, enterprise architecture

Received: November 23, 2013

Enterprise architecture (EA) management has become a widely discussed approach in both industry and academia due to the inefficiency of current IT architectures to cope with rapid changes in business environments. On the other hand Service Oriented Architecture (SOA) is widely accepted as a state of the art approach to the design and implementation of enterprise software. However software design and development according to SOA paradigm is a complex task, often integrating various platforms, technologies, products and design patterns. Hence, it arises a problem of early evaluation of a software architecture to detect design flaws that might compromise expected system qualities. Decisions related to software architecture have a great impact on the business value of a software product under development and influence the software company competitiveness. Usually, a software architecture is developed by a company team, whose experience is limited to narrow set of solutions and technologies. This is a motivation for developing a methodology for the assessment of architectural solution that can be performed in more independent way. Such assessment requires extensive knowledge gathering information on various types of architectural decisions, their relations and influences on quality attributes. In this paper SOAROAD (SOA Related Ontology of Architectural Decisions) is described, which was developed to support the evaluation of architectures of information systems based on SOA approach. The main goal of the ontology is to provide constructs for documenting architecture. However, it is designed to support future reasoning about architecture quality and fulfilling the non functional system requirements such as scalability, ease of maintenance, reuse of software components etc. The last important reason is building a common knowledge base. When building the ontology Architecture Tradeoff Analysis Method (ATAM) was adopted which was chosen as a reference methodology of architecture evaluation.

Povzetek: Opisana je ontologija, ki omogoča evaluacijo arhitektur za informacijske sisteme SOA.

1 Introduction

Service Oriented Architecture (SOA) might be treated as a state of the art approach to the design and implementation of enterprise software, which is driven by business requirements. Within the last decade a number of concepts related to SOA have been developed, including ESB (Enterprise Service Bus), web services, design patterns, service orchestration and choreography and various security standards. Due to the fact that there are many technologies that cover the area of SOA and the fact that SOA is not related to any specific technology, the development and evaluation of SOA compliant architectures is especially interesting and problematic.

SOAROAD has been designed as a methodology for the assessment of software architectures developed according to SOA principles. During system development several stages and corresponding architecture evaluation goals can be identified. The first stage is related to the formulation of a strategy for a new system development or an integration of existing software. The next stage consists in proposing competitive architectural approaches and assessing them with respect to selected quality attributes. The third stage has as an input the assumed system architecture and aims at identifying requirements (usually expressed as scenarios) and determining risks and costs for achieving the assumed scenario responses. This step can be referred as *early* architecture evaluation. The last stage, that can be considered as a *late* architecture evaluation, is related to software verification and validation resulting in the specification of test

This paper is based on P. Szwed, P. Skrzyński, G. Rogus and J. Werewka *Ontology of architectural decisions supporting ATAM based assessment of SOA architectures* published in the proceedings of the 3rd International Workshop on Advances in Semantic Information Retrieval (part of the FedCSIS'2013 conference).

cases used for TDD (Test Driven Development) or BDD (Behaviour Driven Development) approach.

In this paper we focus on the third stage of architecture evaluation. SOAROAD has been designed as a methodology for the assessment of software architectures developed according to SOA principles. It is based on the Architecture Tradeoff Analysis Method (ATAM) [18, 7], which is a mature, scenario-based, early method for architecture assessment. ATAM defines a quality model and an organizational framework for evaluation process. Expected system qualities are represented as mappings between scenarios and quality attributes. System architecture being an input for ATAM is expressed in form of views describing components and their connections. During the evaluation a team of experts analyzes selected properties of components and connections to detect sensitivity points, tradeoffs and assigns risks. In the evaluation process, the first information on expected system qualities, architectural approaches and decisions is collected from architecture documentation and interviews with stakeholders, then a team of experts analyze selected properties of components to identify sensitivity points and evaluate risks. A limitation of the ATAM method is that it depends on experts knowledge, perception and previous experience. It may easily happen that an inexperienced evaluator overlooks some implicit decisions and risks introduced by them.

In the SOAROAD approach the very basic set of ATAM terms used to describe architecture is enriched by including common terminology and relationships between concepts related to various aspects of service oriented architecture design and development. The gathered knowledge, formalized as an ontology, facilitates performing an assessment in more exhaustive manner, helping to ask questions, revealing implicit design decisions and obtaining more reliable results.

The contribution of the paper is a proposal of a SOAROAD ontology as a tool supporting scenario based assessment of systems following a service-orientation paradigm and service design, development and deployment.

The paper is organized as follows. In Section 2 related works are discussed. Section 3 gives an overview of ATAM methodology. Section 4 introduces a concept of ontology application in architecture evaluation. Section 5 provides the ontology description. Section 6 discusses an example of SOAROAD methodology application. Section 7 summarizes the paper and presents conclusions together with future works planned.

2 Related works

Architecture evaluation has attracted many researchers and practitioners during the last 20 years. A survey paper on this topic [26] lists 37 methods of architecture evaluation, classifying them according to two dimensions: location in the software lifecycle (early vs. late) and element being analyzed (system architecture, isolated architectural style or a design pattern). The paper suggests that scenario-based methods, including SAAM [20] and ATAM [18, 7] can be considered as a mature, reliable and easy to implement in practical situations.

There are several reports on successful applications of ATAM for assessment of a battlefield control system [19], wargame simulation [17], product line architecture [10], control of a transportation system [4], credit card transactions system[24] and a dynamic map system [32]. Recently, a few extensions of ATAM were proposed, including a combination with the Analytical Hierarchy Process [36] and APTIA [21].

Despite the fact that the area of enterprise architecture (EA) and service oriented architecture (SOA) has been gaining significant attention there have not been much research on SOA architecture assessment.

Song and Song [30] proposed EA institutionalization processes and its metric based assessment for implemented EA based on the currently available EA frameworks. In the EA processes, we define institutionalization strategies specific to an organization's goals, target architecture based on their baseline architecture, and transition plan for institutionalization. The assessment is based on changes made on existing architecture which describes the current or as-is state of an enterprise. To describe the baseline architecture they suggest organizing information structure according to the architectural views as in ANSI/IEEE Standard 1471-2000 [13].

Javanbakht, Pourkamali, Feizi [16] observed that in some enterprises, particularly in developing countries, baseline is not a suitable basis for creating target architecture and they proposed improvement and correction of organizational architecture by using enterprise architecture maturity. They used multifactor systems to provide a practical method for the assessment of any given organization and making accurate decisions on the improvement or redesign of its architecture based on missions, goals and restrictions of the organization. With the use of their method they claimed that the enterprise architectures can be assessed and an accurate decision about the development of the enterprises can be made based on its mission.

Jange and Medling [23] tried to address the problem of cost benefit ratio of EA with a qualitative research design. They conducted a series of semi structured interviews with industry experts on enterprise architecture in order to identify classes of EA goals, corresponding EA frameworks adoption to achieve those goals and employed EA benefit assessment approaches. Their findings point to, among others, a fairly stable set of EA goals that shift over time and EA frameworks that lack modularity and adjustment capabilities to easily customize towards these goals.

Zhou and Zhang [37] presented an architecture-centric assessment approach for model evaluation over reference architecture to quantitatively estimate architecture maturity and quality. They selected a nine-layer (S3) SOA solution stack as reference architecture, and introduced the necessary mathematical definitions and formulation. The baseline for such an assessment is a model template composed of S3 solution patterns. A template is the starting point of creating a design model.

There has been interesting research performed on the analysis of the composition of services [6]. The authors observed and grouped common service composition techniques into six solution patterns with distinct characteristics of their integration intermediary. Their effort also can be used as a base to develop better solution templates to include architectural building blocks level interactive patterns into solution template creation.

The application of ontologies to provide a systematic and formal description of architectural decisions was first proposed by Kruchten in [22]. The ontology distinguished several types of decisions that can be applied to software architecture and its development process. Main categories included: Existence, Ban, Property and Executive decisions. The ontology defined also attributes, which were used to describe decisions, including states (Idea, Tentative, Decided, Rejected, etc.). In [9] an ontology supporting ATAM based evaluation was proposed. The ontology specified concepts covering the ATAM model of architecture, quality attributes, architectural styles and decisions, as well as influence relations between elements of architectural style and quality attributes. The effort to structure the knowledge about architectural decisions, was accompanied by works aimed at a development of tools enabling the edition and graphical visualization of design decisions, often in a collaborative mode, e.g. [5, 8, 25].

This short selection of works proves that the problem of documenting and visualizing architectural decisions as a support for software development process and architecture evaluation remains a challenge. In contrast to approaches aimed at providing classification of concepts and their relations (commonly referred as TBox), we attempt to gather in the proposed ontology also facts (commonly referred as ABox) constituting ready to use dictionaries of decisions (properties of architectural design) and the knowledge about their relations reflecting current state of the art for SOA technologies.

3 ATAM Overview

The goal of software architecture evaluation methods is to assess whether a system meets or will meet certain requirements concerning quality characterized as *quality attributes*. A standardized list of quality attributes is published in ISO/IEC 9126-1 norm [14], which enumerates six groups of quality attributes: Functionality, Reliability, Usability, Efficiency, Maintainability and Portability. This set was extended in the superseding norm ISO/IEC 25010[15] to 8 groups by adding Compatibility and Security. Many of the quality attributes were known elsewhere under different names, e.g. Efficiency as Performance, Changeability (sub-attribute of Maintainability) as Modifiability, etc.

Architecture evaluation methods may bring the greatest benefits to software development if applied early in the software lifecycle, as identified flaws in system design can be corrected at a lower cost [26]. Typically, an assessment is conducted based on the specification of the software architecture (architectural views) and use other sources of information, such as interviews with various stakeholders including owners, future users, architects and development teams. At an early development stage it is difficult to give the ultimate answer whether a particular quality attribute can or cannot be assured. Therefore, assessment methods aim at estimating such characteristics as a risk or cost.

Identified high risk to achieve a quality attribute can trigger mitigation actions which consist in revising the design and changing the design decisions. However, it should be emphasized that even after changes and corrections are applied some acceptable residual risks can still be present because the estimated effort required to remove them exceed expected losses.

ATAM (Architecture-based Tradeoff Analysis Method) was developed at the Software Engineering Institute (SEI) in 2000 [18], [7] as a successor of the SAAM method [20].

The method aims at evaluating architectural decisions against specific quality attributes and detecting:

- *risks* – architectural decisions that may cause problems to assure some quality attributes,
- *sensitivity points* – decisions related to components or their connections that are critical for achieving required level of quality attribute,
- *tradeoffs* – decisions of increasing one quality attribute with a negative impact on the others.

ATAM provides evaluations based on the requirements expressed as *scenarios* that are elicited and assessed in a formal process divided into phases and steps.

ATAM uses a quality model called the *utility tree*. At the root of the utility tree, an abstract concept *Utility* is placed. Its child nodes are annotated with general quality attributes, e.g. these specified in the ISO norm (performance, reliability, security, modifiability, etc.); at the next level they can be decomposed into more specific attributes, and finally, scenarios are placed at leaves. Both quality attributes present in the utility tree and scenarios are elicited from various stakeholders and represent their point of view on expected system qualities.

According to ATAM, the architecture assessment process is a group effort of various stakeholders involved in system development. It deploys typical group techniques such as brainstorming, assigning priorities and voting. The course of evaluation is divided logically into four phases including nine steps:

1. *Presentation*: (1) presentation of the ATAM method, (2) business drivers and (3) the assumed software architecture.

2. *Investigation and Analysis*: (4)identification of architectural approaches, (5) generation of quality attribute tree and (6) an analysis of the architectural approaches.

3. *Testing*: (7) brainstorming and the prioritization of scenarios, (8) repeated analysis of the architectural approaches with reference to high priority scenarios.

4. *Reporting*: (9) presenting the results of the analysis: risks, sensitivity points and tradeoffs.

4 The concept of SOAROAD ontology application

ATAM has many obvious benefits: it precisely defines the quality model based on a utility tree, enumerates the expected outcomes, indicates the participants and provides an organizational framework for conducting the evaluation. Nevertheless, due to its generic character, the method can cause problems related to collecting and representing information that can be used for an architecture assessment. The identification of key design decisions (properties) that should be considered is up to experts' knowledge and experience. In the case of inexperienced evaluators, some key architectural decisions strongly influencing the system qualities can be easily overlooked. Gathering knowledge related to leading technologies, e.g. web services, business process execution environments, databases, semantic web as a support to ATAM would be beneficial for the efficiency and reliability of the evaluation.

The proposed approach consists in collecting and formalizing this knowledge as an ontology. SOAROAD (SOA Related Ontology for Architectural Decisions) ontology has four main goals, it should:

1. provide a comprehensive description of architectural views, i.e. components and their connections;

2. gather a domain knowledge providing a unified vocabulary related to SOA and enterprise architecture;

3. help to ask question about various properties of architectural design and decisions;

4. be capable to represent assignments of properties relevant to SOA compliant technologies to elements of system architecture.

It was assumed that the ontology would follow a foundational model (ontology skeleton) described later in the section 5.1 defining various properties corresponding to design decisions that can be attributed to components, connections, interfaces and compositions. If applicable, these design decisions can be supplemented by additional relations. The ontology would also specify design patterns.

Another assumption is related to a distribution of the knowledge between ontology TBox (set of classes, their attributes and relations) and ABox (individuals, values of their attributes and relationships). The types of elements appearing in architectural views are classified in the TBox. Concrete elements, e.g. those appearing in the diagrams of architectural views, are represented as individuals in an ABox. The ontology describes types of design decisions (properties) as classes, whereas their values as individuals that can be directly assigned to elements of architectural views or linked to form trees. Such approach is more flexible, than e.g. a simplistic model of *key-value* pairs assigned to components, where a *key* would correspond to a decision type, and a *value* to a concrete decision.

The concept of the ontology application is presented in the Fig. 1 (thick lines indicate data flows and thin arrows describe import relations among ontologies). The process of building an architecture description starts with eliciting *Architecture views ABox*, i.e. a set of linked components, interfaces and connections. This model can be prepared either manually or with the support of dedicated import tools converting ArchiMate [33, 34] models of Archi editor [2] or UML [27], e.g. from VisualParadigm. For clarity, the figure shows only one import tool that converts the ArchiMate model into *Architecture views ABox* encoded in OWL language.

A web based tool supporting architecture description uses the classes and individuals defined in the *SOAROAD ontology Domain Description TBox* and *SOAROAD Architectural decisions ABox* to generate forms or questionnaires in which software architects or members of development teams can make assignments of property values to elements of architecture views.

These questionnaires are dynamically generated from the ontology content by transforming relevant items to XML representation and then applying XSLT transforms to give them a visual appearance. Users selections in questionnaires after feeding them to a web server are converted into assertions in *Detailed Architecture ABox* ontology stored at the server side. For this purpose we use Jena library and TDB [1] as the storage system. The resulting *Detailed Architecture ABox* refers elements of *Architecture views ABox* and individuals defined in SOAROAD ontology (merging two input ontologies and asserting additional relations). This ontology serves as a detailed architecture documentation within a software development project. It can be examined either manually or with use of automated tools.

It should be mentioned that for large projects realized by multiple teams at distant locations, maintaining a centralized repository documenting software architecture and architectural decisions can be considered as a key factor for project success. In many cases, independent teams make many implicit decisions that may influence interoperability, performance, modifiability and other quality attributes. Collecting detailed information by the suggested in ATAM interviews is more time consuming and less exhaustive than filling in questionnaires driven by an ontology content.

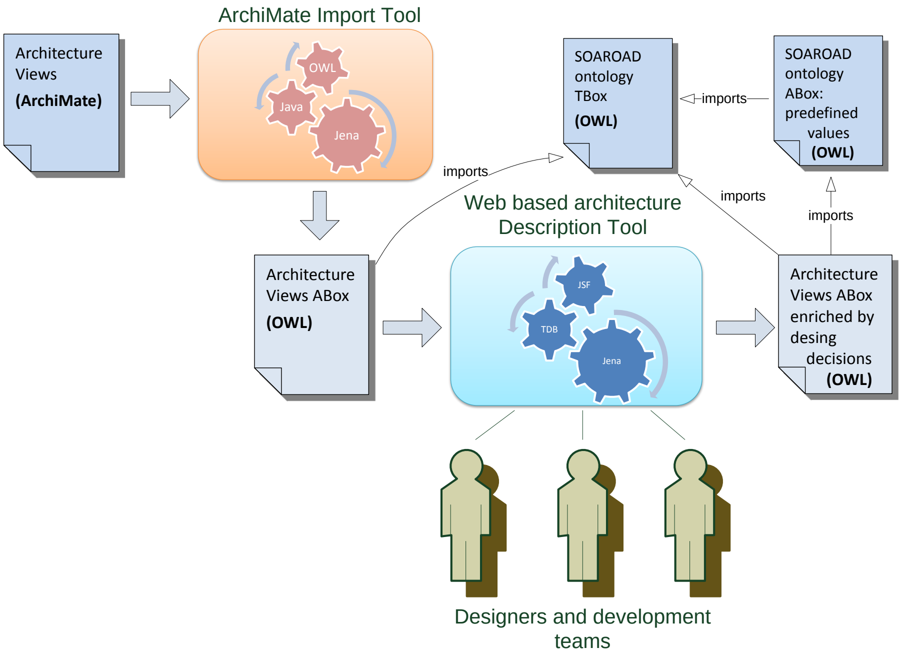

ArchiMate Import Tool
Architecture Views (ArchiMate)
OWL
Java
Jena
SOAROAD ontology TBox (OWL)
imports
SOAROAD ontology ABox: predefined values (OWL)
imports
imports
imports
Web based architecture Description Tool
Architecture Views ABox (OWL)
JSF
TDB
Jena
Architecture Views ABox enriched by desing decisions (OWL)
Designers and development teams

Figure 1: A concept of application of SOAROAD ontology

5 Ontology description

Ontology engineering methodologies [11, 12, 29] usually distinguish the following common steps in the ontology development:

1. Specification, aimed at establishing the domain of the ontology, its scope, usage and competency questions (including preparing motivating examples);

2. Conceptualization. The goal of this step is to identify concepts, arrange them in hierarchies and establish relations;

3. Formalization which consists coding ontology in a formal language, e.g. OWL;

4. Deployment – using the ontology in a software tool.

In this section we will briefly describe the assumptions determined in the specification phase and results of formalization. The main outcome of the specification phase is the foundational model described in section 5.1. During the conceptualization step, we manually gathered and analyzed information related to service oriented architectures, technologies, architectural approaches, design patterns, etc. originating from various sources: books, technical papers, reference manuals and Internet resources.

The ontology was populated with the information during the formalization phase by translating intermediate textual description into OWL constructs. For this purpose a small software tool using Jena [1] library was developed. The resulting ontology content is described in section 5.2.

5.1 Foundational model of software architectures

The basic model of software architecture used in ATAM [3] defines it after [28] as a set of components and linking them connections. We extend this simplistic model by defining *Interfaces* and *Functions* of components as presented in Fig. 2. A connection links a component having the caller role with an interface (calee). Components, connections and interfaces can be attributed with: *ComponentProperties*, *ConnectionProperties* and *InterfaceProperties* respectively. Examples of such properties are: platform, web service type, communication type, queueing and query granularity.

Composition is a coherent set of components and connectors. System architecture is itself a composition. For the purpose of analysis we may focus on a particular subset of components and connectors and describe their properties, e.g. a distribution of queries among several databases building up a composition or realization of a design pattern.

During the ATAM based evaluation the overall system architecture and properties of its parts are analyzed to establish scenario responses and achievements of corresponding quality attributes. It may be, however, observed that some architecture properties or their combinations have known influence on quality attributes, e.g. a use of asynchronous web services or applying MVC design pattern, which increases modifiability and a granularity of queries, has an impact on performance. This kind of knowledge can be expressed as *influences* relations.

Architectural decision is an assignment of a property value to a component, interface, connection or a composition. In this context the terms *property* and *architectural*

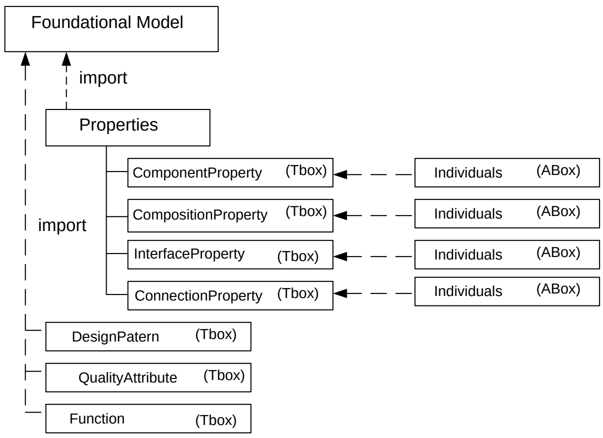


Figure 3: Structure of SOAROAD ontology

decision can be used to some extent interchangeably. However, it may happen that certain decisions or components are dependent on previously assigned properties. An example of such a dependency is the composition type – a property assigned to a set (composition) of web service components. Selecting orchestration as the composition type requires that an orchestration component, e.g. BPEL capable module is be used. The *required* relation or its subproperties in the ontological model express this dependency.

The assumed foundational model adopts a reification strategy while modeling various properties of an architectural design. Properties are defined as classes, whose individuals can be linked by additional relations indicating specific roles. An example of such a property is MVC design pattern – pattern, which requires the identification of a components playing the roles of a Model (typically a database), a Controller (e.g. an EJB) and a View (e.g. a set of HTML pages produced by JSP scripts).

Two types of components are distinguished: *ApplicationComponents* and *InfrastructureComponents*. Application components correspond to software developed modules; infrastructure components provide such supporting functions, as message queuing or service registry.

5.2 The ontology content

SOAROAD ontology, provides a knowledge about software architecture, its structure, components, connections and required properties in the context of the SOA paradigm. It consists of 110 classes, 9 object properties and 105 individuals.

The structure of SOAROAD ontology is shown in (Fig. 3). The Foundational Model presented earlier in Fig. 2 forms the ontology skeleton. In the conceptualization phase, the skeleton was extended by defining subclasses of classes marked in gray: various types of properties (*ComponentProperty*, *ConnectionProperty*, *CompositionProperty* and *InterfaceProperty*), functions, design patterns and quality attributes.

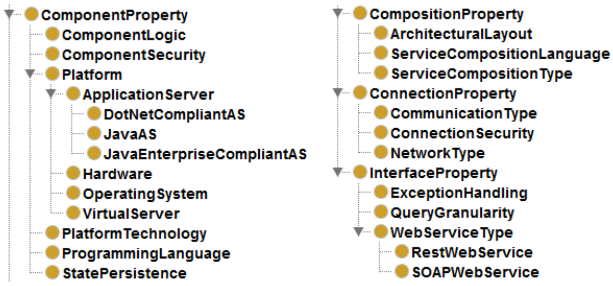


Figure 4: Classes of component properties

For each property, that can be treated as a class of design decision, a number of individuals (corresponding to decision values) is defined. They can be selected in assignments, e.g. *JavaEECompliantAS* (a subclass of *ComponentProperty*) has several predefined individuals: *JBoss*, *Glassfish*, *WebLogic*, *Web-Sphere*, *ColdFusion*, etc.

ComponentProperty class defines various properties and design decisions, which can be assigned to components (Fig. 4). Software architect preparing ATAM evaluation should consider them as an exhaustive list of questions related to important issues in SOA architectures. Examples of such properties are *Platform* (*Hardware*, *OperatingSystem*, *ApplicationServer*), *PlatformTechnology*, *ProgrammingLanguage*, *ComponentLogic* and *ComponentSecurity*.

Example ontology assertions related to component properties are presented in Table 1 and Table 2. A property (an ontology class) is followed by property values (individuals in the ontology) put in parentheses.

ConnectionProperty subsumes the *CommunicationType*, *ConnectionSecurity* and *NetworkType*. The class *CommunicationType* has two individuals: *ComunicationType.asynchronous* and *ComunicationType.synchronous*. *ConnectionSecurity* has individuals representing various security technologies SSL, VPN, WS_Security.

CompositionProperty is a superclass for *ArchitecturalLayout*, *ServiceCompositionLanguage*, *ServiceCompositionType*. *ArchitecturalLayout* defines types of application structure. Its individuals are: *LayeredArchitecture, P2P, ServiceComposition* and *SpokeAndHub*.

ServiceCompositionLanguage defines languages (*BPEL*, *CDL* or *not_defined*) and *ServiceCompositionType* with individuals: *choreography* and *orchestration*.

Class *InterfaceProperty* has subclasses *ExceptionHandling* (defining exception handling method), *QueryGranularity* (granularity level of of interface functions), *WebServiceType* (type of communication protocol: *SOAPWebService* or *RESTWebService*).

Apart from defining design decisions, the ontology specifies functions of components. Their list is rather related to infrastructure components. Class Function contains classes of entities such as: *Routing*, *MessageMapping*, *ProtocolSwitch*, *MediationService*, *MessageValidation*, *AuditFunction*, *DatbaseIntegration*, etc.

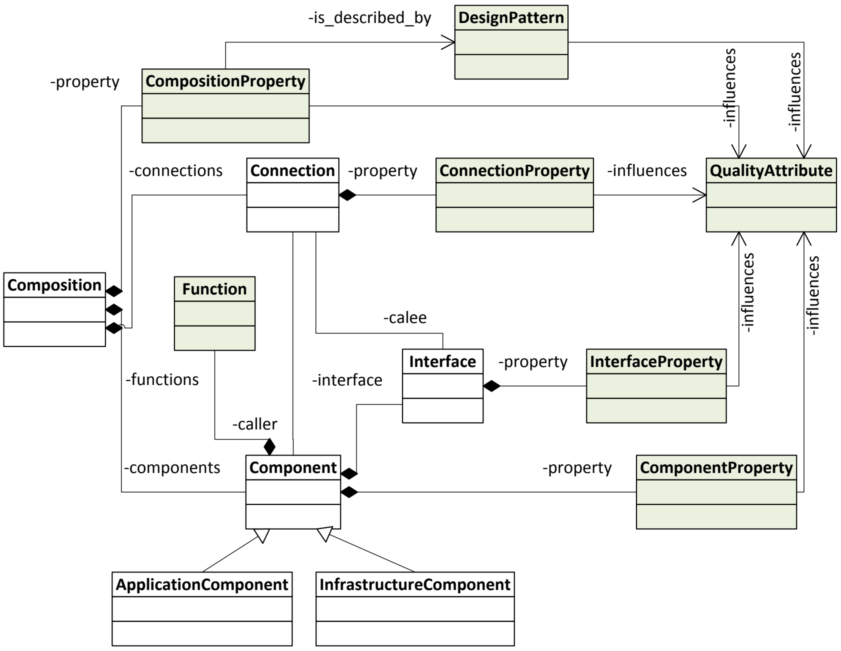
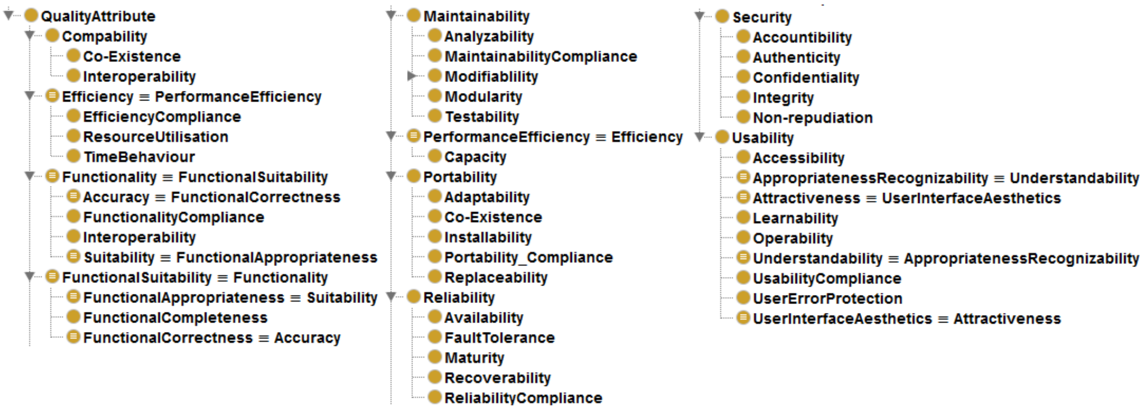


Figure 2: Foundational model of software architecture and its properties



- QualityAttribute
  - Compability
    - Co-Existence
    - Interoperability
  - Efficiency ≡ PerformanceEfficiency
    - EfficiencyCompliance
    - ResourceUtilisation
    - TimeBehaviour
  - Functionality ≡ FunctionalSuitability
    - Accuracy ≡ FunctionalCorrectness
    - FunctionalityCompliance
    - Interoperability
    - Suitability ≡ FunctionalAppropriateness
  - FunctionalSuitability ≡ Functionality
    - FunctionalAppropriateness ≡ Suitability
    - FunctionalCompleteness
    - FunctionalCorrectness ≡ Accuracy
  - Maintainability
    - Analyzability
    - MaintainabilityCompliance
    - Modifiablility
    - Modularity
    - Testability
  - PerformanceEfficiency ≡ Efficiency
    - Capacity
  - Portability
    - Adaptability
    - Co-Existence
    - Installability
    - Portability_Compliance
    - Replaceability
  - Reliability
    - Availability
    - FaultTolerance
    - Maturity
    - Recoverability
    - ReliabilityCompliance
  - Security
    - Accountibility
    - Authenticity
    - Confidentiality
    - Integrity
    - Non-repudiation
  - Usability
    - Accessibility
    - AppropriatenessRecognizability ≡ Understandability
    - Attractiveness ≡ UserInterfaceAesthetics
    - Learnability
    - Operability
    - Understandability ≡ AppropriatenessRecognizability
    - UsabilityCompliance
    - UserErrorProtection
    - UserInterfaceAesthetics ≡ Attractiveness

Figure 5: The tree of quality attributes (according to ISO/IEC 9126 and ISO/IEC 25010)

Table 1: Component properties

Property (values)	Description
PlatformTechnology (CORBA, EJB, JINI, RMI)	Set of technologies used on the platform.
ComponentLogic (flexible, fixed, rulebased)	Specifies an approach the component logic implementation.
Platform	Defines the component platform. Has several subclasses: ApplicationServer, Hardware, OperatingSystem and VirtualServer
ProgrammingLanguage (Cpp, Java, Ruby, PHP, Erlang, Python, C, C_sharp)	Define programming language used to implement a component.
StatePersistence (Stateless, Statefull)	Specifies whether a component saves internal data during and in between calls of operations on the client's behalf.

The ontology provides a taxonomy of quality attributes. Quality attribute is a nonfunctional characteristic of a component or a system. It represents the degree to which software possesses a desired combination of properties, which are defined by means of externally observable features of software systems. Some of the attributes are related to the overall system design, while others are specific to run-time or design time. Quality attributes can be categorized into two broad groups: attributes that can be directly measured (e.g. performance) and attributes that can be indirectly measured (e.g., usability or maintainability). In the latter category, attributes are divided into subcharacteristics.

SOAROAD ontology defines 30 quality attributes including both terms defined in software quality model by the ISO/IEC 9126-1 norm [14] and those arising directly from requirements to architectures formulated in the SOA man-

Table 2: Properties describing platform (subclasses of *Platform*).

Property (values)	Description
ApplicationServer	Subclass of Platform. Defines an application server on which a component is deployed, can have such attributes, as: version (string), vendor (string)
JEECompliantAS (TomEE, Glassfish, JBoss, Interstage, JOnAS, Geronimo, SAPNeatWeaver, WebSphere, Resin, ColdFusion, WebLogic)	Subclass of ApplicationServer dedicated to JEE compliant components.
DotNetCompliant-AS (AppFabric, IIS, TNAPS, Base4, Mono)	Subclass of ApplicationServer; its individuals define products for .NET enviroment
JavaAS (Jetty, Enhydra, iPlanet)	Application servers for Java environment
Hardware	Subclass of Platform. Used to specify a hardware configuration on which the component is deployed. Attributes: memory (double), processor (string), number_of_cores (int)
OperatingSystem (Windows, Unix, Linux, iOS, Android, Bada, Blackberry)	Subclass of Platform. Defines types of operating systems on which a component is executed. Attributes: version (string), vendor (string), product (string)
VirtualServer (no, yes)	Subclass of Platform. Specifies whether a component is deployed on a virtual server

ifesto . Examples of classes belonging to the first group (see Fig. 5) are: *Functionality*, *Reliability*, *Usability*, *Efficiency*, *Maintainability* and *Portability*. The example of classes originating from SOA manifesto are *ServiceAutonomy*, *PlatformIndependency*, *LooseCoupling*, *Modularity*, *OpenStandardAdoptation*, *BusinessAgility* etc.

When designing an applications to meet quality requirements, it is necessary to consider a potential impact of design properties on various quality attributes. SOAROAD ontology defines *influences* object property to this kind of relation.

A design pattern can be seen as a structure build of components of particular types, defining their roles and relations among them together with a set of restrictions on their usage. Design patterns do not change the functionalities of a system but only the organization or structure of those functionalities. One of the most important benefits of using design patterns is that they constitute standardized software building blocks with a well defined influence on quality attributes. In SOAROAD ontology the class *DesignPattern* has 56 subclasses representing patterns dedicated to SOA architecture. The examples of subclasses are: *EnterpriseServiceBus*, *EventDrivenMessaging*, *Orchestration*. The relation *is_described_by* links a particular *CompositionProperty* to one of the defined design patterns.

6 Example

We illustrate the proposed approach on an example of a small system aimed at publishing and browsing of free of charge announces. The diagram in Fig. 6 gives the system architecture specified in ArchiMate language. As it can be noticed, two layers: application and technology are presented. In the application layer several system components are distinguished: *Data Base* with the *SQL interface*, three Java beans: *Announcement JPA* (Java Persistence API), *Announcement Business Logic* and a *Facade* providing *Announcement WS* – web service based interface. The last component of the application layer visible on the diagram is *Announcement JSF Presenter* being responsible for presentation and interaction with end users. It plays here the role of web service consumer. The components are packaged as three artifacts: *ANN_DB* (PostgreSQL), *ann.ear* and *ann_pres.war* and deployed at three separate servers (technology layer nodes) linked with two connections: *JDBC* and *WS Presenter*.

The above specification is an input for *ArchiMate Import* tool (indicated in Fig. 1) that transforms it into *Architecture Views ABox*. Fig. 7 gives and excerpt of this ontology (node marked with boldlines). We focus on three elements application layer: *Announcement Facade*, *WS* interface and accessing it *JSF presenter*. Following the foundational model that encompasses connections and their properties, the *WS Presenter Connection* was also included. The remaining elements of of ArchiMate specification are converted into properties.

The tool supporting architecture description allows to assign various properties (architectural decisions) to ontology individuals corresponding to components and connections of the software architecture. In the presented example:

- *Announcement Facade* is deployed on Intel Xeon 2.13 GHz machine running Ubuntu 10.4 system and GlassFish application server.

- *Announcement WS* is a SOAP web service with low query granularity and exception handling based on soap faults.

- *WS Presenter Connection* is asynchronous, uses SSL based protection mechanism and 10Gb network.

- *Announcement JSF Presnter* is deployed on JBoss application server and is stateless.

The resulting graph of interconnected elements with assigned properties presented in a user-friendly browseable form can be input to ATAM analysis performed in the standard manner.

The SOAROAD ontology specifies additional relations (Fig. 8) that can be used in architecture assessment.

The *supports* relation indicates that particular elements can be used together, e.g. JBoss (ApplicationServer) supports Document.Literal (SOAP web service style).

The *supports* property has two subproperties: *supports_fully* and *supports_partially*, that can be used to indicate possible incompatibility issues. Another way to define potentially conflicting architectural decisions is to use Conflict objects (reified multirole properties) that indicate sets of properties, which should not be used together, provide specification of conflict levels (e.g. *partially_compatible*,

http://www.soa-manifesto.org/

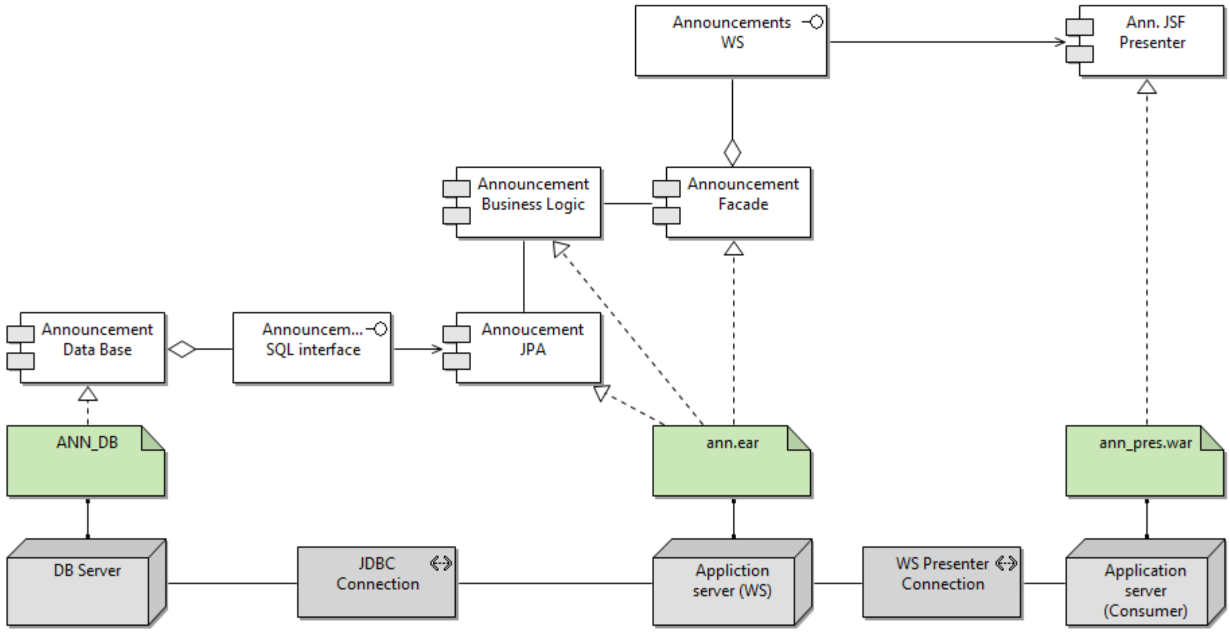


Figure 6: Announcement system expressed in Archimate language.

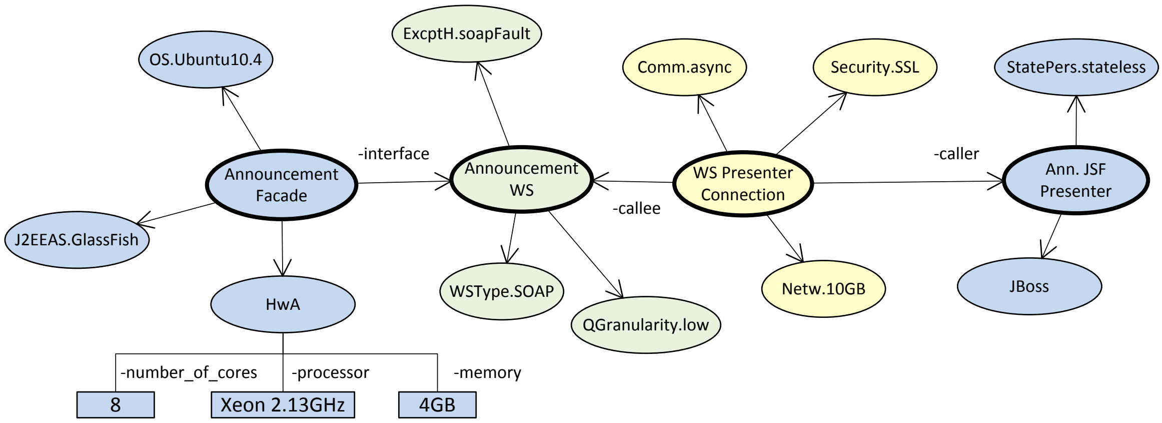


Figure 7: ABox describing the architecture of the announcement system. Elements of an architectural view (marked with boldlines) are assigned with design decisions (individuals of classes defined in the ontology)

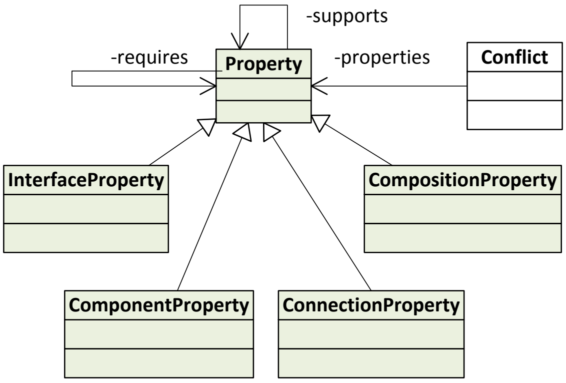


Figure 8: Relations between properties

incompatible, *error_prone*) and textual description (rationales). The required relation can be used to specify that one element requires another. Such assertions can be explored, while reasoning about implicit decisions, i.e. resulting from earlier assignments.

The SOAROAD ontology is formalized in the OWL language. In consequence, it should follow the Open World Assumption (OWA) to be compatible with OWL reasoners, e.g. Pellet, Fact+ or Racer.

According to OWA, the following approach was adopted:

- A lack of the assertion on property of a particular type, means that nothing is known about the assignment. For example in Fig. 7 no information is provided about the hardware or operating system for *Annotations JSF Presenter*.

- A lack of decision is represented explicitly by an individual (constant) of a particular type, e.g. and individual *OperatingSystem.not_decided* can be assigned to *Annotations JSF Presenter*.

- Conflicting decisions of the same type can be attributed to a component, e.g. *Annotations JSF Presenter* can be attributed with Windows and Linux properties. Such conflicts reflect, that in a certain step an alternative is envisaged. During an evaluation process (possibly supported by reasoning with the use of a separately developed set of SWRL rules) such an assertion can be indicated as non valid.

- Negative assertions about properties are represented by a special ban relation, whose object can be an anonymous individual of a selected type. For example an assertion (*Annotations JSF Presenter*, *ban*, *IOS.anonymous*) can be made, where *IOS.anonymous* belongs to the class *IOS* (operating system).

7 Conclusion

This paper describes the SOAROAD ontology and a concept of a tool that supports the documenting architectures of SOA-based systems. The proposed approach addresses the problem that can be encountered during architecture assessment: to be reliable, a reasoning about architecture qualities, must have solid foundations in a knowledge related to a particular domain: architectural styles, design patterns, used technologies and products. The idea behind SOAROAD ontology is to gather experts knowledge to enable even inexperienced users performing ATAM-based architecture evaluation. An advantage of the presented approach is that its result is a joint representation of architecture views and properties attributed to design elements formalized in OWL language.

From a software engineering perspective, such centralized information resource may represent a valuable artifact, which, if maintained during the software lifecycle, can provide reference to design decisions that can be examined later in the integration, testing and deployment phases.

On the other hand, the machine interpretable representation, constituting a graph of interconnected objects (individuals), can be processed automatically to check consistency, detect potential flaws and calculate metrics. An extensive list of metrics related to architectural design was defined in [35]. We plan to adapt them to match the structural relations in the SOAROAD ontology, as well to develop new ones.

Another direction that is at present researched is an application of fuzzy reasoning to evaluate quality attributes. We use fuzzy Mamdani rules encoded in SWRL language defining influence of selected design decisions on quality. The approach taken follows the idea presented in [31].

Further plans are related to the extensions of the currently developed tool. At present its functionality is limited to building the architecture description. Our intention is to fully integrate it with ATAM process allowing specifying scenarios, describing sensitivity points, tradeoffs and risks.

References

[1] Jena - a semantic web framework for java.

[2] Archi, archimate modelling tool, 2011. [Online; accessed 23-June-2012].

[3] P. Bianco, R. Kotermanski, and P. Merson. Evaluating a service-oriented architecture. Technical Report CMU/SEI-2007-TR-015, Carnegie Mellon, September 2007.

[4] N. Bouck'e, D. Weyns, K. Schelfthout, and T. Holvoet. *Applying the ATAM to an Architecture for Decentralized Control of a Transportation System*, volume 4214, pages 180–198. Springer, 2006.

[5] R. Capilla, F. Nava, S. Pérez, and J. C. Dueñas. A web-based tool for managing architectural design decisions. *ACM SIGSOFT Software Engineering Notes*, 31(5), 2006.

[6] Y.-C. Chang, P. Mazzoleni, G. A. Mihaila, and D. Cohn. Solving the service composition puzzle. *IEEE SCC*, 2:387–394, 2008.

[7] P. Clements, R. Kazman, and M. Klein. *Evaluating Software Architectures: Methods and Case Studies*. Addison-Wesley Professional, 2001.

[8] R. C. de Boer, P. Lago, A. Telea, and H. van Vliet. Ontology-driven visualization of architectural design decisions. In *WICSA/ECSA*, pages 51–60. IEEE, 2009.

[9] A. Erfanian and F. S. Aliee. An ontology-driven software architecture evaluation method. In *Proceedings of the 3rd international workshop on Sharing and reusing architectural knowledge*, SHARK '08, pages 79–86, New York, NY, USA, 2008. ACM.

[10] S. Ferber, P. Heidl, and P. Lutz. *Reviewing product line architectures: Experience report of ATAM in an automotive context*, volume 2290, pages 364–382. Springer, 2001.

[11] M. Fernandez-Lopez, A. Gomez-Perez, and N. Juristo. Methontology: from ontological art towards ontological engineering. In *Proceedings of the AAAI97 Spring Symposium*, pages 33–40, Stanford, USA, March 1997.

[12] M. Gruninger and M. S. Fox. Methodology for the design and evaluation of ontologies. In *International Joint Conference on Artificial Inteligence (IJCAI95), Workshop on Basic Ontological Issues in Knowledge Sharing*, 1995.

[13] IEEE. IEEE standard 1471-2000, ieee recommended practice for architectural description of software-intensive systems, 2000.

[14] ISO/IEC. Software engineering – product quality, ISO/IEC 9126-1. Technical report, International Organization for Standardization, 2001.

[15] ISO/IEC. ISO/IEC cd 25010-3: Systems and software engineering – software product quality requirements and evaluation (SQuaRE) – software product quality and system quality in use models. Technical report, International Organization for Standardization, 2009.

[16] M. Javanbakht, M. Pourkamali, and F. M. Derakhshi. A new method for enterprise architecture assessment and decision-making about improvement or redesign. *Proceedings of the Fourth International Multi-Conference on Computing in the Global Information Technology*, pages 69–76, 2009.

[17] L. G. Jones and A. J. Lattanze. Using the architecture tradeoff analysis method to evaluate a wargame simulation system: A case study. *Technical Report CMUSEI2001TN022 Software Engineering Institute Carnegie Mellon University Pittsburgh PA*, (December):33, 2001.

[18] Kazman. Atam:method for architecture evaluation. *CMUSEI2000TR004*, 2000.

[19] R. Kazman, M. Barbacci, M. Klein, J. Carriere, and S. G. Woods. Experience with performing architecture tradeoff analysis. *Proceedings of the 21st international conference on Software engineering ICSE 99*, pages 54–63, 1999.

[20] R. Kazman, L. Bass, G. Abowd, and M. Webb. *SAAM: a method for analyzing the properties of software architectures*, volume 16pp, pages 81–90. IEEE Comput. Soc. Press, 1994.

[21] R. Kazman, L. Bass, and M. Klein. The essential components of software architecture design and analysis. *Journal of Systems and Software*, 79(8):1207–1216, 2006.

[22] P. Kruchten. *An ontology of architectural design decisions in software intensive systems*, pages 54–61. Citeseer, 2004.

[23] M. Lange and M. Jan. An experts' perspective on enterprise architecture goals, framework adoption and benefit assessment. *Proceedings of the 15th IEEE International Enterprise Distributed Object Computing Conference Workshops*, pages 304–313, 2011.

[24] J. Lee, S. Kang, H. Chun, B. Park, and C. Lim. Analysis of VAN-core system architecture- a case study of applying the ATAM. In *Proceedings of the 2009 10th ACIS International Conference on Software Engineering, Artificial Intelligences, Networking and Parallel/Distributed Computing*, SNPD '09, pages 358–363, Washington, DC, USA, 2009. IEEE Computer Society.

[25] L. Lee and P. Kruchten. *Visualizing Software Architectural Design Decisions*, volume 5292, pages 359–362. Springer-Verlag, 2008.

[26] B. Roy and T. C. N. Graham. Methods for evaluating software architecture : A survey. *Computing*, 545(2008-545):82, 2008.

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

[28] M. Shaw and D. Garlan. *Software Architecture: Perspectives on an Emerging Discipline*, volume 123. Prentice Hall, 1996.

[29] J. Sliwa, K. Gleba, W. Chmiel, P. Szwed, and A. Glowacz. IOEM - ontology engineering methodology for large systems. In P. Jedrzejowicz, N. T. Nguyen, and K. Hoang, editors, *ICCCI (1)*, volume 6922 of *Lecture Notes in Computer Science*, pages 602–611. Springer, 2011.

[30] H. Song and Y.-T. Song. Enterprise architecture institutionalization and assessment. *Proceedings of the 9th IEEE/ACIS International Conference on Computer and Information Science*, pages 870–875, 2010.

[31] P. Szwed. Application of fuzzy ontological reasoning in an implementation of medical guidelines. In *Human System Interaction (HSI), 2013 The 6th International Conference on*, pages 342–349, 2013.

[32] P. Szwed, I. Wojnicki, S. Ernst, and A. Glowacz. Application of new ATAM tools to evaluation of the dynamic map architecture. In A. Dziech and A. Czyżewski, editors, *Multimedia Communications, Services and Security*, volume 368 of *Communications in Computer and Information Science*, pages 248–261. Springer Berlin Heidelberg, 2013.

[33] The Open Group. Archimate 1.0 specificattion, 2009.

[34] H. Van Den Berg, H. Bosma, G. Dijk, H. Van Drunen, J. Van Gijsen, F. Langeveld, J. Luijpers, T. Nguyen, R. Oosting, Gerand Slagter, and et al. ArchiMate made practical. *Work*, 2007.

[35] A. Vasconcelos, P. Sousa, and J. Tribolet. Information system architecture metrics: an enterprise engineering evaluation approach. *The Electronic Journal Information Systems Evaluation*, 10(1):91–122, 2007.

[36] P. Wallin, J. Froberg, and J. Axelsson. Making decisions in integration of automotive software and electronics: A method based on ATAM and AHP. *Fourth International Workshop on Software Engineering for Automotive Systems SEAS 07*, pages 5–5, 2007.

[37] N. Zhou and L.-J. Zhang. Analytic architecture assessment in soa solution design and its engineering application. *Proceedings of the IEEE International Conference on Web Services*, pages 807–814, 2009.

Artificial Immune Based Cryptography Optimization Algorithm

Xuanwu Zhou[1, 2], Kaihua Liu[1], Zhigang Jin [1], Shourong Tian[3], Yan Fu[1,3] and Lianmin Qin[3]
[1] School of Electronics and Information Engineering, Tianjin University, Tianjin 300072, China
[2] Command College of the Chinese Armed Police Forces, Tianjin 300250, China
[3] Administrative Centre of Yantai Tax-free Port, Yantai 265400, China
E-mail: schwoodchow@163.com

Keywords: artificial immune, cryptography optimization, blind signcryption, optimization coefficient, ECDSA

Received: November 8, 2012

In the paper, an improved clone selection algorithm for cryptography optimization is proposed, the algorithm integrates genetic algorithm with immune computing and makes use of reproduction and mutation operator to maintain the diversity and optimization of candidate objects. As an experiment of the clone algorithm, a blind signcryption scheme with immune optimized parameter is proposed. In the signcryption scheme, parameters generated with clone selection have relatively higher level of fitness and thus avoids the arbitrary selection of essential parameters. Then we analyze the efficiency and feasibility of immune optimization algorithms with experiment data from the signcryption scheme. The reproduction operator in the algorithm can greatly improve the fitness level of candidate group, while the mutation operator effectively maintains the diversity of candidate individuals. In the experiment, the optimization coefficient (OC) reaches 0.9301 when the clone algorithm is executed just once. Lastly, we make detailed comparison between the optimized signcryption scheme and other typical schemes, including the blind signature of D.Chaum and the ECDSA signature. The data from the experiment and comparison show that the optimization algorithm can effectively improve the efficiency and accuracy of parameter optimization in cryptography systems.

Povzetek: Predstavljen je izviren algoritem za kriptografsko optimizacijo, ki temelji na genetskih imunskih sistemih.

1 Introduction

Artificial immune system is an important branch of computation intelligence; it simulates the architecture and operating pattern of biological immune system and makes full use of the superior bionic mechanisms. In terms of computing ability, biological immune system is a self-adaptive and self-organized system with highly distributed and parallel architecture, and it has prominent capability in learning, recognition, memorizing and property extracting. Artificial immune system is an application-orientated model of biological immune system based on the bionic mechanisms; it also has superb capability in data processing and problem solving. Presently, artificial immune system has been widely applied in pattern recognition, intelligent optimizing, machine learning, data mining and information security, etc [1,2,3].

In traditional cryptography schemes, system parameters are simply generated with pseudo-random generator or the selection process is just overlooked. The arbitrary selection of system parameters makes the cryptography system more vulnerable to malicious attack. In order to reinforce the stability and security of cryptography algorithms, the random parameters can be generated by intelligent optimization algorithm with random selection.

In this paper, we propose an improved clone selection algorithm which integrates genetic algorithm with immune optimization algorithm. Then a signcryption scheme with immune optimized parameter is proposed as an experiment of the clone selection algorithm. Then the optimized signcryption scheme is compared with other typical schemes, including blind signature of D. Chaum and the ECDSA signature scheme. The signcryption scheme and the experiment show that the improved clone algorithm can effectively improve the efficiency and accuracy of parameter optimization in cryptography systems.

2 Artificial immune system and its algorithms

Artificial immune system (AIS) is a series of algorithms and systems based on the superior architecture and operating mechanism in biological immune system. Artificial immune system has a wide application in pattern recognition, intelligent optimizing, machine learning, data mining and information security, etc.

Biological immune system can recognize and clear invading pathogens, toxin, tumour cells from genetic mutation and prostrate cells to achieve immune defending effect and organism homeostasis. One of two immune responses is innate immune response taking rapid defending measures at first, which is fulfilled by skin, mucous membrane, phagocyte cells, natural killer, compliments etc. The other is adaptive immune response that is mainly executed by T lymphocyte cells and B lymphocyte cells. The hierarchical defence structure of biological immune system is demonstrated in Figure 1[4,5,6].

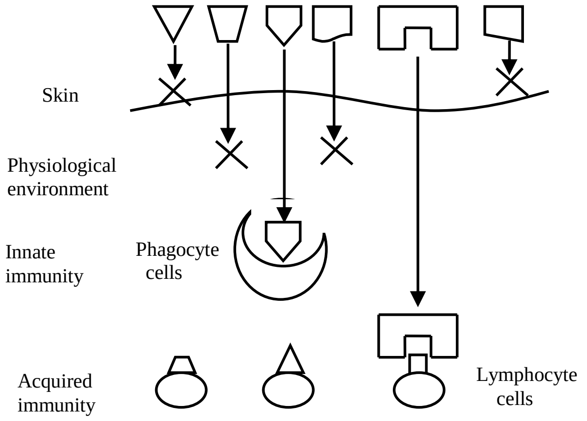


Figure 1: Hierarchical defence structure of biological immune system.

Biological immune system has superior ability to learn, memorize and recognize information, and its operating mechanism is characterized with self-organizing, distribution and diversity. Therefore many researchers have been applying the superior bionic mechanisms of biological immune system to develop corresponding models and algorithms in artificial immune system.

The basic bionic mechanisms of biological immune system can be categorized as: immune learning, immune memorizing, immune recognition, clone selection, diversity, distribution, self-adapting and immune network [7, 8].

Simulating the architecture and operating pattern of biological immune system, artificial immune system has three types of immune algorithms: basic immune algorithm, negative selection algorithm and clone selection algorithm.

Basic immune algorithm: generally, basic immune algorithms have similar searching strategy to genetic algorithms, and they also apply selecting and mutating in optimization.

Negative selection algorithm: this algorithm is based on the principles of negative selection in biological immune system. Negative selection provides protection against mistaken immune response toward normal organisms.

Clone selection algorithm: this algorithm is based on the principles of clone selection in biological immune system. In clone selection algorithms, individual objects will also undergo a process of clone reproduction with the stimulus from corresponding evaluation function (antigen). In the process of clone selection, the objects with higher suitability will be selected for reproduction and the suitability (affinity) and scale of these objects will also be gradually improved [9, 10, 11].

3 Immune optimization in cryptography schemes

In cryptography schemes, the proper selection of certain parameters is essential to the security and feasibility of the whole system. In many schemes, such parameters are simply generated with pseudo-random generator or the selection process is just overlooked. The arbitrary selection of system parameters makes the cryptography system more vulnerable to malicious attack. In the scheme, we introduce clone selection optimization into the design and analyzing of cryptography schemes, and put forward improved cryptography schemes with artificial immune optimization.

3.1 Parameter optimization algorithm

An optimized random parameter is first generated with clone selection algorithm.

(1) Encoding .The candidate parameters should first be encoded as a number string. In our example, we set the length of string as 4 bits.

(2) Initial group generating. The initial group is selected with random. And the number of individuals in the group is set as 5 for the convenience of computing. In our example, they are:

$$x_1= (0, 0, 0, 1),\ x_2= (0, 1, 1, 0),\ x_3= (1, 1, 1, 0),$$

$$x_4= (1, 0, 1, 0),\ x_5= (1, 0, 0, 1).$$

(3) Computing and evaluating of fitness. To evaluate the fitness of parameters, we use an objective function to decide the difference between selected strings. In our example, we use a linear function to compute the maximum function value as the standard of selection.

The objective function is set as:

$$f(x) =-x^2+2x+1. \qquad (1)$$

Then we compute the function value of different strings.

$$f(x_1) = f(0001) =2,\ f(x_2) = f(0110) =-23,\ f(x_3) = f(1110)$$

$$=-167,\ f(x_4) = f(1010) =-79,\ f(x_5) = f(1001) =-62.$$

(4) Reproduction of individuals. Mimicking clone selection of immune system, a certain number of individuals with high level of fitness should be selected for reproduction. In our example, two individuals with the highest level of fitness will be selected for reproduction, and the scale of reproduction is also directly proportional to its level of fitness.

The function value of string x_1, x_2 is the highest, so the two strings should be selected for reproduction to increase their perception in the whole group. In proportion to the level of fitness, string $x_1$will be reproduced twice, and string x_2 once.

Now, the temporary individuals in the group are:

$$x_1= (0, 0, 0, 1),\ x_1= (0, 0, 0, 1),\ x_1= (0, 0, 0, 1),$$

$$x_2= (0, 1, 1, 0),\ x_2= (0, 1, 1, 0),$$

$$x_3= (1, 1, 1, 0),\ x_4= (1, 0, 1, 0),\ x_5= (1, 0, 0, 1).$$

(5) Mutation. There are two mutation operations in the clone selection algorithm: crossover and self-mutation. In crossover mutation, two individuals are selected from the temporary group to develop into two new stings by

exchanging some value of the string. The probability of crossover is connected with the level of fitness in inverse proportion. In our example, string x_3, $x_4$and $x_5$have the lowest level of fitness, so$x_3$and x_5 are selected to exchange the latter two bits. And two new strings are generated [12, 13, 14].

x_6= (1, 1, 0, 1) , x_7= (1, 0, 1, 0) = x_4.

In self-mutation operation, the algorithm will change some bits in some selected strings, that is 0 to 1 or 1 to 0. The probability of mutation is also connected with the level of fitness in inverse proportion. In the example, x_4= (1, 0, 1, 0) has relatively lower level of fitness, and will be selected to change some bits of itself. And the new string is x_8= (0, 0, 1, 0).

(6) Repeating of algorithm. Then we should also compute the function value of the new strings, and make the decision of reproduction, mutation and discarding.

$f(x_1)$ = f(0001) =2, $f(x_2)$ = f(0110) =-23, $f(x_3)$ = f(1110) =-167, $f(x_4)$ = f(1010) =-79, $f(x_5)$ = f(1001) =-62, $f(x_6)$ = f(1101) =-142, $f(x_8)$ = f(0010) =1.

Then the above operating algorithm will be repeated for certain times until the requirement is satisfied. In our example, the algorithm will be executed only once, and the final temporary strings in the group are:

x_1= (0, 0, 0, 1), x_1= (0, 0, 0, 1), x_1= (0, 0, 0, 1),
x_2= (0, 1, 1, 0), x_2= (0, 1, 1, 0),
x_3= (1, 1, 1, 0), x_4= (1, 0, 1, 0), x_5= (1, 0, 0, 1),
x_6= (1, 1, 0, 1), x_8= (0, 0, 1, 0).

After comparing the function values of different strings, the strings with the lowest fitness level will be excluded from the group. In this example, five strings with the lowest level of fitness should be excluded from the group to keep the stability of group scale, they are

x_3= (1, 1, 1, 0), x_6= (1, 1, 0, 1), x_4= (1, 0, 1, 0),
x_5= (1, 0, 0, 1) , x_2= (0, 1, 1, 0).

And the final optimized strings of the group are:

x_1= (0, 0, 0, 1), x_1= (0, 0, 0, 1), x_1= (0, 0, 0, 1),
x_2= (0, 1, 1, 0), x_8= (0, 0, 1, 0).

3.2 Immune optimized blind signcryption

In our scheme, the random parameter is generated with the above optimizing algorithm in advance, and other secret parameters of the scheme can also be generated with clone selection algorithm in advance.

Definition 3.2.1 (Elliptic Curve) an elliptic curve $E(F_q)$ over finite field F_q is a sextuple: T = (q , a , b , P , l , h), where P =(x_P , y_P) is the base point of $E(F_q)$, prime l is the order of P . As to $t \in Z_l^*$, Q and $G \in E(F_q)$, $Q = tG$ denotes multiple double additions on elliptic curve. O is the point at infinity, satisfying lP = O and $G +$ O = G for any point $G \in E(F_q)$ [15,16,17].

Definition 3.2.2 (**ECDLP**, Elliptic Curve Discrete Logarithm Problem). ECDLP is the following computation

$x \leftarrow$ $ECDLP(Q, P)$ (P is a base point and $Q \in \langle P \rangle$, $x \in Z_l^*$, $Q = xP$).

In the scheme, user A entrusts signcryption generator B to generate a signcryption for message $m \in Z_l^*$ without disclosing any information about it.

$$\Phi = (GC , GK , BSC , USC)$$

Common parameters generation:

$$GC\ (1^k) = \text{“On input} (1^k):$$

$$(T, H, (E, D)) \leftarrow GC\ (1^k)\text{.”}$$

T = (q , a , b , P , l , h)where P =(x_P , y_P) is the base point of $E(F_q)$, $ord(P) = l$ is a prime, O is the point at infinity. $H : \{0,1\}^* \rightarrow Z_l^*$, (E , D) is secure symmetric encryption/decryption algorithm.

Key pair generation:

$$GK\ (A, 1^k) = \text{“On input} (A, 1^k):$$

$$sk_A \xleftarrow{\$} Z_l^*, PK_A = sk_A P \neq O,$$

$$(sk_A , PK_A) \leftarrow \text{.”}$$

$$GK\ (B, 1^k) = \text{“On input} (B, 1^k):$$

$$sk_B \xleftarrow{\$} Z_l^*, PK_B = sk_B P \neq O,$$

$$(sk_B , PK_B) \leftarrow \text{.”}$$

Signcryption generating:

$$BSC(sk_A, PK_B, m) = \text{“On input} (sk_B , PK_A , C):$$

$$r \leftarrow_R Z_l^*,\ R = rP \neq O,$$

$$A \xrightarrow{R} Q.$$

$$(u, v, w) \leftarrow_R Z_l^*, U = uPK_B \neq O,$$

$$k = (U)_x \bmod(|E(\cdot)|),$$

$$c = E_k(m), h \leftarrow H(m \| ID_Q),$$

$$F = (h + w)R - vP, e = (h + w) \bmod l,$$

$$A \xleftarrow{e} Q\text{.”}$$

$$t = (sk_A + er) \bmod l, i \leftarrow_R Z_l^*, I = iP \neq O.$$

$$A \xrightarrow{t} Q\text{.”}$$

$$s = u^{-1}(t - v - h) \bmod l,$$

$$A \xleftarrow{(c,h)} Q.$$

$$h' \leftarrow H(c \| (I)_x), s' = (i - sk_A h') \bmod l,$$

$$A \xrightarrow{(h',s')} Q.$$

$$s'P + h'PK_A = iP = I, h'\ ? = H(c \| (I)_x),$$

$$C = (c , h , h' , s , s' , F)\text{.”}$$

Unsigncryption algorithm:

$$USC\ (sk_B , PK_A , C) = \text{“On input} (sk_B , PK_A , C):$$

If $sk_B \notin Z_l^*$ or $PK_A \notin < P >$ return $\perp$,
Parse C into (c , h , s , F , h' , s'),
If $s, s' \notin Z_l^*$ or $c \notin SP_E$ or $F \notin < P >$ return $\perp$, else
$s^{-1} sk_B(PK_A + F - hP) = U$,
$k = (U)_x \bmod(|E(\cdot)|)$, $m = D_k(c)$,
$h \ ?= H(m \| ID_Q)$,
If the equation holds return m , else return $\perp$.”

4 Analysis of the optimization scheme

Artificial immune optimization is the simulation of biological immune system and it is also an improved genetic algorithm with biological inheritance and natural selection mechanism. Clone selection algorithm in artificial immune system is an iteration algorithm. While searching for optimized group, clone selection generates a new improved individual from the original one; and from the improved one to another further improved one. Therefore, clone selection algorithm has much superiority in efficiency and stability compared with other optimization algorithm.

In our scheme, the optimization of random parameters is executed only once, but the fitness level of the strings has been greatly improved. The comparison can be made in the following table.

Initial group	Fitness level	Temporary group	Fitness level	Optimized group	Fitness level
x_1= (0, 0, 0, 1)	2	x_1= (0, 0, 0, 1)	2	x_1= (0, 0, 0, 1)	2
		x_1= (0, 0, 0, 1)	2		
x_2= (0, 1, 1, 0)	-23	x_1= (0, 0, 0, 1)	2	x_1= (0, 0, 0, 1)	2
		x_2= (0, 1, 1, 0)	-23		
x_3= (1, 1, 1, 0)	-167	x_2= (0, 1, 1, 0)	-23	x_1= (0, 0, 0, 1)	2
		x_3= (1, 1, 1, 0)	-167		
x_4= (1, 0, 1, 0)	-79	x_4= (1, 0, 1, 0)	-79	x_2= (0, 1, 1, 0)	-23
		x_5= (1, 0, 0, 1)	-62		
x_5= (1, 0, 0, 1)	-62	x_6= (1, 1, 0, 1)	-142	x_8= (0, 0, 1, 0)	1
		x_8= (0, 0, 1, 0)	1		
Sum of fitness	-229	Sum of fitness	-489	Sum of fitness	-16
Average level	-45.8	Average level	-48.9	Average level	-3.2

Table 1: Comparison of fitness level.

Definition: Let α is the average fitness level of the initial group, and β is the average fitness level of the temporary group or the optimized group, $\delta = \beta - \alpha$ is the difference between α and β , then optimization coefficient(OC) γ can be defined as the following formula.

$$\gamma = \frac{\delta}{|\alpha|} = \frac{\beta - \alpha}{|\alpha|} . \quad (2)$$

According to the definition of optimization coefficient, the smaller the value of γ, the weaker the optimization effect of clone selection algorithm on initial group. The larger the value of γ, the stronger the optimization effect of clone selection algorithm on initial group. When $\gamma > 0$, the algorithm has positive optimization effect on the group, when $\gamma < 0$, the algorithm has negative optimization effect on the group, When $\gamma = 0$, the algorithm has no optimization effect on the average level of the group.

In the above table, the average fitness level of the initial group is -45.8, after clone selection operation, the average fitness level of the optimized group is -3.2. The optimization coefficient γ between the initial group and the optimized group is 0.930131, the average fitness level of the initial group has been greatly improved by93.01%, and thus the optimization effect of clone selection algorithm proves to be remarkable.

Different immune operations render different optimization effect on the group. In reproduction operation, individuals with higher level of fitness will be reproduced to obtain their majority in the group, and thus the scale of the whole group will be improved. The average level of fitness will also be improved with the increase of ideal individuals. In mutation operation, new individuals can not necessarily be those with relatively higher level of fitness, therefore, the average level of fitness can not necessarily be improved. On the contrary, the fitness level will most probably be reduced. Yet, mutation operation in the immune optimization algorithm maintains the diversity of the candidate group.

The comparison of different clone operations can be made in the following table.

	Initial group	Temporary group with reproduction	Temporary group with mutation	Optimized group
Sum of fitness	-229	-348	-489	-16
Average level	-45.8	-43.5	-48.9	-3.2
OC γ		0.0502	-0.1241	0.9346

Table 2: Comparison of different optimization effect.

In the above table, the average fitness level of the initial group is -45.8, after reproduction operation, the fitness level is -43.5, and the optimization coefficient γ is 0.0502, the fitness level is improved by 5.02％. Yet, after mutation operation, the average fitness level is -48.9, the optimization coefficient γ is -0.1241, the average fitness level is reduced by 12.41％.With the discarding process, the scale of the group keeps stable, and the fitness level is also improved with the discarding of improper individuals with low level of fitness. The average fitness level increases from -48.9 to -3.2 with a prominent optimization coefficient γ 0.9346, and the average fitness level is improved by 93.46%.

5 Comparison with other typical schemes

In this section, the proposed artificial immune based optimization algorithm and the optimized signcryption scheme will be compared with other typical schemes, including the famous blind signature put forward by D.Chaum and the ECDSA signature algorithm, which has been accepted as standard elliptic curve algorithm in many international standardization organizations, such as ISO14888-3, ANSI X9.62, IEEE1363-2000, etc.

5.1 Comparison with blind signature of D.Chaum

The signature algorithm for comparison in our scheme is based the original scheme put forward by D.Chaum and the security of the blind signature is based on elliptic curves cryptosystem.

(1)System parameter

F_q is a finite field (q is a prime number of n bits, $n \geq 190$), an elliptic curve on this finite field is defined as the following.

$$E: y^2 = x^3 + ax + b \; (a, b \in F_q,$$
$$4a^3 + 27b^2 (\bmod q) \neq 0). \tag{3}$$

$P \in E(F_q)$ is a base point whose order is a large prime number l . $\#E(F_q)$ denotes the order of the elliptic curve which has a factor of large prime number larger than 160 bits [18, 19, 20].

$(P)_x$ is a function which makes the conversion from a point $P = (x, y)$ on elliptic curve to x . In the blind signature scheme, user A requires B to generate a blind signature of his message $m \in Z_l^*$ for him. ($K_A = k_A P$, k_A), ($K_B = k_B P$, k_B) are the public/private key pairs of A and B. In our scheme, the Hash function in signing algorithm is eliminated for simplicity, which can be easily added without loss of generality.

(2)Message blinding

Before generating signatures, the original user should blind the secret message with blinding parameters.

Step1: As to message $m \in Z_l^*$, User A randomly selects parameter $v \in Z_l^*$ and computes

$$m' = vm (\bmod l) \tag{4}$$

$$V = v^{-1} P \tag{5}$$

Then he sends m' and V to B.

Step2: The blind signature generator B randomly selects $r \in Z_l^*$ and then computes

$$R = rV \neq 0 \tag{6}$$

$$t = m'(R)_x (\bmod l) \tag{7}$$

$$s = r - k_B t (\bmod l) \tag{8}$$

Then he sends (t , s) to user A.

(3)Signature generating

After getting the partial signature (t , s), user A computes the following to get the blind signature.

$$s' = v^{-1} s (\bmod l) \tag{9}$$

$$t' = v^{-1} t (\bmod l) \tag{10}$$

Then (s' , t') is the blind signature for message $m \in Z_l^*$ generated by entrusted signer B.

(4)Blind signature verifying

After getting blind signature (s' , t'), the signature verifier can testify the signature with the public key of the entrusted signer B.

$$R = s'P + t'K_B \tag{11}$$

$$t' ?= m(R)_x (\bmod l) \tag{12}$$

If the formula holds, the verifier will accept (s' , t') as a valid blind signature of message $m \in Z_l^*$ [21, 22].

Remark 1. As a comparison, in the traditional schemes with random parameter selection, the parameters are selected without any optimization, such as in the step of message blind protocol (4) - (8). In these steps, parameters r and v are generated randomly without any optimization or selection standards. Many insecure parameters or weak keys will be selected to insure the security of the scheme, which will make the cryptography system more vulnerable to malicious attack. While with the proposed signcryption optimized algorithm, many insecure parameters or weak keys will be discarded or undergo the mutation process because of their low level of fitness.

5.2 Comparison with ECDSA signature

ECDSA signature scheme is as the following:

(1)System parameter

System parameters are the same as the above scheme, $k_A \in Z_l^*$ is the private key, $K_A = k_A P$ is the corresponding public key, $H: \{0,1\}^* \rightarrow Z_l^*$ is a secure one-way hash function.

(2) Signing algorithm

As to message $m \in Z_l^*$, the signer randomly selects parameter $u \in Z_l^*$ and computes

$$U = uP \neq 0, \tag{13}$$

$$e = H(m), \tag{14}$$

$$s = u^{-1}(e + k(U)_x)(\bmod l). \quad (15)$$

$\sigma = (U, s)$ is the signature text.

(3)Verifying algorithm

After getting signature $\sigma = (U, s)$, the verifier can testify the signature with the public key of the signer.

$$w = s^{-1}, \quad (16)$$

$$u_1 = ew(\bmod l), \quad (17)$$

$$u_2 = (U)_x w(\bmod l), \quad (18)$$

$$(u_1 P + u_2 K)_x ?= (U)_x. \quad (19)$$

If the above formula is correct, the signature verifier will accept $\sigma = (U, s)$ as a valid signature of message $m \in Z_l^*$ [23, 24, 25].

Remark 2. Although ECDSA signature has been accepted as standard signature algorithm in elliptic curves, parameter $u \in Z_l^*$ in signature generating is still generated randomly without any optimization or selection to avoid weak keys and insecure parameters. Compared with the proposed scheme with immune optimization in the paper, ECDSA is more vulnerable to malicious attack, such as signature forgery and attack on the secret key for signing.

5.3 Comparison of performance

In this section, we will make a performance comparison between our immune optimized signcryption scheme and other traditional techniques, including the blind signature of D.Chaum and the ECDSA signature. To fulfil both the functions of encryption and signature as the proposed immune based blind signcryption, the above signature schemes must be improved with a secure symmetric encryption/decryption algorithm, for which the typical ElGamal encryption algorithm is selected with its simplicity and security. ElGamal public key encryption algorithm is as follows.

(1)System parameter

p is a large prime with binary length no less than 1024 such that $p-1$ has a large prime factor. $G = Z_p^*$ is a cyclic group under multiplication modulo p in which the discrete exponentiation function is conjectured to be one-way (meaning the discrete logarithm function is computationally hard) . g is the generator of group G ,meaning $G = \{g^0, g^1, \cdots, g^{l-1}\}$, where $l = |G|$ is the order (size) of G.

Then, as to any $x \in Z_l^*$, the computation of $y = g^x$ via x and g is called discrete exponentiation function, which is computationally feasible; but the computation of x via y and g is called discrete logarithm problem (DLP), which is computationally infeasible. $k \in Z_p^*$ is the private key, and $K = g^k$ is the public key.

(2)Encryption algorithm

As to message $m \in Z_p^*$, the sender randomly selects $r \in Z_p^*$, and computes

$$c_1 = g^r (\bmod p), \quad (20)$$

$$c_2 = mK^r (\bmod p). \quad (21)$$

Then (c_1, c_2) is the cipher text.

(3)Decryption algorithm

$$c_2 (c^k{}_1)^{-1} = mK^r (g^{rk})^{-1}$$
$$= mg^{rk} (g^{rk})^{-1} \equiv m(\bmod p). \quad (22)$$

In these schemes, such computing as modular exponential, modular inverse and elliptic curve addition ,elliptic curve scalar multiplication should be taken into comparison for computing complexity, while computing cost of modular addition, modular multiplication, hash, symmetric encryption/decryption are negligible. To ensure the security of basic cryptographic primitives, the minimum security parameters recommended for current practice are as follows: for DLP, $|p|$=1024bits, $|q|$=160bits. For RSA, $|N|$=1024bits; for ECC, $|q|$=131bits (79, 109 may also be chosen), $|l|$=160bits. The block length of the block cipher is 64bits. The length of secure hash function is 128bits.

Scheme	*GC+ GK*	*Sign*	*VF*	*Sum cost*	*IO*	*Length of C*
Blind signature	1*kP*	2*kP* +3*I*	2*kP*	5*kP* +3*I*	/	2 $\lvert l\rvert$
ECDSA	1*kP*	1*kP* +1*I*	2*kP*+ 1*I*	4*kP*+2*I*	/	$\lvert l\rvert$+ $\lvert q\rvert$
	GC+ GK	*EC*	*DC*			
Elgamal encryption	1*E*	2*E*	1*E*+1 *I*	4*E*+1*I*	/	2 $\lvert p\rvert$
	GC+ GK	*Sign and EC*	*VF and DC*			
Compound scheme 1	1*kP*+ 1*E*	2*kP* +2*E*+ 3*I*	2*kP*+ 1*E*+1 *I*	5*kP*+4 *E*+4*I*	/	2 $\lvert l\rvert$+ 2 $\lvert p\rvert$
Compound scheme 2	1*kP*+ 1*E*	1*kP* +2*E*+ 1*I*	2*kP*+ 1*E*+2 *I*	4*kP* +4*E*+ 3*I*	/	$\lvert l\rvert$+ $\lvert q\rvert$+ 2 $\lvert p\rvert$
	GC+ GK	*SC*	*USC*			
Immune based blind signcryption	2*kP*	4*kP* +1*I*	1*kP*+ 1 *I*	7*kP* +2*I*	*N*	$\lvert E(\cdot)\rvert$+2$\lvert h\rvert$+ 2$\lvert l\rvert$

Table 3: Comparison of computing and communication cost.

Notes of notations: 1. *GC+GK* denotes the common parameters and key generation algorithms; *Sign/VF* denotes the signature/verification algorithms; *IO* denotes immune optimization algorithm; *EC/DC* denotes encryption/decryption algorithm; *SC* denotes the signcryption algorithm; *USC* denotes the unsigncryption algorithm; *Length of C* denotes the length of signcryption text /cipher-text/signature. *Compound scheme 1* is the scheme of blind signature+ Elgamal encryption;

Compound scheme 2 is the scheme of ECDSA+ Elgamal encryption. 2. *E* denotes modular exponential; *I* denotes modular inverse; *kP* denotes scalar multiplication on elliptic curves. / denotes there is no relevant computation. 3. |*E* (·) |denotes the block length of block cipher. 4. *N* denotes negligible.

In the above ECC and Elgamal based schemes, elliptic curve scalar multiplication *kP* and modular exponential $\alpha^k \bmod p$ are the most complex computations, so we will compare these two typical computations with the currently recommended security parameters:

(1) Elliptic curve scalar multiplication *kP*, where $P \in E(F_{2^l})$, *E* is a non-supersingular curve, $l \approx 160$，*k* is a random160-bit integer.

(2) Modular exponential $\alpha^k \bmod p$, where *p* is a 1024-bit prime and *k* is a random160-bit integer.

A field multiplication in F_q takes l^2 ($q= 2^l$) bit operations, then a modular multiplication in (2) takes $(1024/160)^2 \approx 41$times longer than a field multiplication in (1). Computation of *kP* by repeated doubling and adding on the average requires 160 elliptic curve doublings and 80 elliptic curve additions. From the addition formula for non-supersingular elliptic curves, an elliptic curve addition or doubling requires 1 field inversion and 2 field multiplications. The time to perform a field inversion is equivalent to that of 3 field multiplications. Hence, computing *kP* requires the equivalent of 1200 field multiplications, or $1024/41 \approx 29$ 1024-bit modular multiplications. On the other hand, computing $\alpha^k \bmod p$ by repeated squaring and multiplying requires an average of 240 1024-bit modular multiplications. Thus, the operation in (1) can be expected to be about $240/29 \approx 8$ times faster than the operation in (2) [26].

In the following table, the computation costs of the schemes are compared by the equivalence of *kP*, $\alpha^k \bmod p$ and field inversion to field multiplication in F_q ($q= 2^l$,$|q| \approx 160$bits).

Scheme	*GC+GK*	*Sign*	*VF*	*Sum cost*	*IO*	*Length of C*
Blind signature	1200	2409	2400	6009	/	320bits
ECDSA	1200	1203	2403	4806	/	291bits
	GC+GK	*EC*	*DC*			
Elgamal encryption	9840	19680	9881	39401	/	2048bits
	GC+GK	*Sign and EC*	*VF and DC*			
Compound scheme 1	11040	22089	12281	45410	/	2368bits
Compound scheme 2	11040	20883	12284	44207	/	2339bits
	GC+GK	*SC*	*USC*			
Immune based blind signcryption	2400	4803	2403	9606	*N*	640bits

Table 4: Comparison of computing and communication data.

Remark 1. (Comparison with compound scheme 1). Based on the result of Koblitz and Menezes [26], the computing cost in parameter and key generation in our scheme is 2400/11040 ≈ 1/5of that in compound scheme1; signcryption operation in ours is about 4803/22089 ≈ 1/5 of that in scheme1, and unsigncryption is about 2403/12281 ≈ 1/5 of that in scheme1. To sum up, our scheme reduces about 1-9606/45410 ≈ 78.9% commutating cost compared with compound scheme1.

Remark 2. (Comparison with compound scheme 2). As per the result of [26], the computing cost in parameter and key generation in our scheme is 2400/11040 ≈ 1/5of that in compound scheme2; signcryption operation in ours is about 4803/20883 ≈ 1/5 of that in scheme2, and unsigncryption is about 2403/12284 ≈ 1/5 of that in scheme2. To sum up, our scheme reduces about 1-9606/44207 ≈ 78.3% commutating cost compared with compound scheme2.

Remark 3. (Comparison of communication efficiency). The length of signcryption text in our scheme is 640/2368 ≈ 1/4of that in compound scheme1 and 640/2339 ≈ 1/4of that in compound scheme2; our scheme reduces about 1-640/2368 ≈ 73% communication cost compared with compound scheme1and reduces about 1-640/2339 ≈ 72.6% communication cost compared with compound scheme2.

Remark 4. Furthermore, the immune based optimization algorithm in our blind signcryption scheme is an algorithm of polynomial time complexity which can be neglected in the comparison of computation and communication efficiency. For specific application systems, the optimization algorithm can be executed in advance without any influence to the efficiency and designed as a separate computing unite which provide optimization service to other function units, such as encryption, signature, authentication, etc.

Therefore, the proposed cryptography optimization algorithm and the blind signcryption scheme prove to be more efficient and applicable to many security schemes in resource-restricted environment.

6 Conclusions

This paper studies the unique properties of biological immune system and optimization application in cryptography system. In the scheme, we introduce clone selection optimization into the design and analyzing of cryptography schemes, and put forward an improved signcryption scheme with artificial immune optimization. In the scheme, parameters with high level of security and fitness are selected as candidate individuals, and those with security problem or low level of fitness are rejected. On this basis, the final selection of parameters can be made with random mode. Thus the scheme avoids the security problems of other cryptography scheme and reinforces its stability, adaptability and robustness.

Acknowledgement

The authors should thank the anonymous reviewers for their constructive advice and comments to the paper,

with which we can improve our work clerically and academically.

References

[1] Han K H, Park K H. Parallel Quantum-inspired Genetic Algorithm for Combinatorial Optimization Problems, Proceedings of the CEC. Piscataway: IEEE Press, 2001:1442-1429.

[2] Xuanwu Zhou, Ping Wei,etc. Study on Proxy Signature Schemes with Bionic Optimization[C]. Proceedings of FITME’2009, IEEE Press. 2009, (Vol.3)365-368.

[3] D.W. Matolak, and B. Wang. Efficient Statistical Parallel Interference Cancellation for DS-CDMA in Rayleigh Fading Channels. IEEE Transactions On Wireless Communications, vol. 6, no. 2, pp.566-574, February 2007.

[4] Alexandra Boldyreva,Adriana Palacio,Bogdan Warinschi. Secure Proxy Signature Schemes for Delegation of Signing Rights [J]. Journal of Cryptology. 2012, 25(1): 57-115.

[5] Yong Yu,Yi Mu,Willy Susilo,ect. Provably secure proxy signature scheme from factorization[J]. Mathematical and Computer Modelling. 2012, 55(3-4): 1160-1168.

[6] Emura Keita,Miyaji Atsuko,Rahman Mohammad Shahriar. Dynamic attribute-based signcryption without random oracles[J]. International Journal of Applied Cryptography. 2012, 2(32): 199-211.

[7] Seung Hyun Seo;Kyu Young Choi;Jung Yeon Hwang;Seungjoo Kim. Efficient certificateless proxy signature scheme with provable security [J]. Information Sciences. 2012, 188: 322-337.

[8] Degabriele Paul,Paterson Kenny,Watson Gaven. Provable Security in the Real World[J]. IEEE Security & Privacy. 2011, 9(3): 33-41.

[9] Harendra Singh,Girraj Kumar Verma. ID-based proxy signature scheme with message recovery[J].Journal of Systems and Software. 2012, 85(1): 209-214.

[10] Xuanwu Zhou, Zhigang Jin, etc. Short Signcryption Scheme for the Internet of Things [J]. Informatica. Vol.35 (4) 521-530, 2011.

[11] Han Yu Lin;Chien Lung Hsu;Shih Kun Huang. Improved convertible authenticated encryption scheme with provable security[J]. Information Processing Letters. 2011, 111(13): 661-666.

[12] Tzong-Sun Wu;Han-Yu Lin;Pei-Yih Ting. A publicly verifiable PCAE scheme for confidential applications with proxydelegation[J].European Transactions on Telecommunications. 2012, 23(2): 172-185.

[13] Zhang Chuanrong. Zhang Yuqing . Li Fageng and Xiao Hong.: New Signcryption Algorithm for Secure Communication of ad hoc Networks. Journal of Communications, 2010, 31(3): 19-24.

[14] Han K H,Kim J H.Quantum-inspired Evolutionary Algorithms with a New Termination Criterion, Hε Gate, and two-phase Scheme. IEEE Transactions on Evolutionary Computation, 2004, 8(2):156-169.

[15] Gu Jingjing,Chen Songcan,Zhuang Yi. Wireless Sensor Networks-Based Topology Structure for the Internet of Things Location [J].Chinese Journal of Computer . 2010, 33(9): 1548-1556.

[16] Zhu Hongbo,Yang Longxiang,Yu Quan.Investigation of Technical Thought and Application Strategy for the Internet of Things [J]. Journal of Communication. 2010,31(11):2-9.

[17] Haipeng Zhang, Mitsuo Gen. Effective Genetic Approach for Optimizing Advanced Planning and Scheduling in Flexible Manufacturing System.GECCO’06, July 8-12, 2006, Seattle, Washington, USA.

[18] Z Luo, M Zhao, S Liu,ect. Generalized Parallel Interference Cancellation With Near-Optimal Detection Performance[J].IEEE Transactions On Signal Processing. 2008, 56(1): 304-312.

[19] Xuanwu Zhou. Elliptic Curves Cryptosystem Based Electronic Cash Scheme with Parameter Optimization [C]. Proceedings of KESE’2009, IEEE Press. 2009, 182-185.

[20] S Manohar, V Tikiya, R Annavajjala,etc. BER Optimal Linear Parallel Interference Cancellation for Multicarrier DSCDMA in Rayleigh Fading [J]. IEEE Transactions On Communications. 2007, 55(6): 1253-1265.

[21] Blundo C, Desantis A. Perfectly Secure Key Distribution for Dynamic Conferences. Advances in Cryptology-Crypto’92. New York: Springer-Verlag, 1993, 471-486.

[22] Keita Emura,Atsuko Miyaji,Mohammad Shahriar Rahman. Dynamic Attribute–based Signcryption without Random Oracles[J].International Journal of Applied Cryptography,2012,2(3):199-211.

[23] Xu Peng,Cui Guohua,Lei Fengfu, Tang Xueming, Chen Jing. An Efficient and Provably Secure IBE Scheme Under the Standard Model[J].Chinese Journal of Computer . 2010, 33(2): 335-1556.

[24] Xuanwu Zhou. Elliptic Curves Cryptosystem Based Electronic Cash Scheme with Parameter Optimization[C]. Proceedings of KESE’2009, IEEE Press. 2009, 182-185.

[25] Kim Y K,Park K,Ko J.A symbiotic evolutionary algorithm for the integration of process planning and job shop scheduling. Computers and Operations Research.2003, 30:1151- 1171.

[26] Koblitz N, Menezes A and Vanstone S. The State of Elliptic Curve Cryptography [J]. Designs, Codes and Cryptography, 2000, 30(19): 173-193.

Bilinear Grid Search Strategy Based Support Vector Machines Learning Method

Li Lin, Zhang Xiaolong, Zhang Kai and Liu Jun
School of Computer Science and Technology, Wuhan University of Science and Technology, China
Hubei Province Key Laboratory of Intelligent Information Processing and Real-time Industrial System, China
E-mail: lilin@wust.edu.cn

Keywords: support vector machines, model selection, parameters optimization, protein structure prediction

Received: October 21, 2013

Support Vector Machines (SVM) learning can be used to construct classification models of high accuracy. However, the performance of SVM learning should be improved. This paper proposes a bilinear grid search method to achieve higher computation efficiency in choosing kernel parameters (C, γ) of SVM with RBF kernel. Experiments show that the proposed method retains the advantages of a small number of training SVMs of bilinear search and the high prediction accuracy of grid search. It has been proved that bilinear grid search method (BGSM) is an effective way to train SVM with RBF kernel. With the application of BGSM, the protein secondary structure prediction can obtain a better learning accuracy compared with other related algorithms.

Povzetek: Razvita je nova metoda iskanja parametrov za metodo SVM.

1 Introduction

Support Vector Machines (SVM) is a new machine learning method based on statistical learning theory and structural risk minimization [1-3]. The core function of SVM identifies the maximal margin hyperplane and a set of linearly separable data, classifies data correctly, so as to maximize the minimum distance between data and the hyperplane. A number of recent studies on SVM attempt to explore simple and efficient methods to solve the problem of maximal margin hyperplane [4-6]. Many of these works study the performance of SVM learning [7-9]. Several kernel functions can be used in SVM, such as linear function, polynomial function, RBF function, Gaussian function, MLPs with one hidden layer and spline.

SVM is used to construct accurate classification models and has been widely applied, such as in handwritten character recognition, web page/text automatic classification, gene analysis and so on [10]. However, there is still no widely accepted way of selecting kernel function and its parameters in SVM learning. The selection of parameters for SVM algorithms usually depends on large-scale search.

SVM learning is a kind of quadratic programming (QP) problem. Despite its advantages, there are a number of drawbacks in selecting hyperparameters in the size of matrix involved in the QP problem. Therefore, this paper proposes a bilinear grid search method to compute the penalty parameter and the kernel parameter (C, γ) of SVM using RBF kernel. This method is efficient in reducing the training space in QP. Bilinear grid search algorithm has the advantages of both bilinear search and grid search. The proposed algorithm expands the search range of (C, γ) so that it can perform SVM learning with a small size of training samples to construct classification models with high accuracy.

The rest of the paper is structured as follows: Section 2 introduces SVM learning and relevant search strategies; Section 3 proposes bilinear grid search method in SVM learning with RBF kernel; In section 4, we conduct experiments to test the efficiency and applicability of the proposed algorithm; Finally, Section 5 is devoted to concluding remarks and future research recommendations.

2 Search strategy for SVM learning

SVM classification can be described as:

Given:

– A training set of instance-label pairs *(x_i, y_i)*, $i = 1,...,l$, where $x_i \in R^n$ *and* $y \in \{1,-1\}^l$.

Find:

– The solution to the minimum value of $\frac{1}{2}w^T w + C \sum_{i=1}^{l} \xi_i$,

where $y_i(w^T Z_i + b) \ge 1 - \xi_i$, $\xi_i \ge 0, i = 1,...,l.$

Here, the training vector x_i are mapped into a higher - (probably infinite-) dimensional space using function ϕ as $Z_i = \phi(x_i)$; $C(C > 0)$ is the penalty parameter of the error term.

Usually, the formation (1) can be considered as the following dual problem:

- Minimize $F(\alpha) = \frac{1}{2}\alpha^T Q\alpha - e^T \alpha$, subject to $0 \le \alpha \le C, i = 1,...,l,$

- $y^T\alpha = 0$, where *e* is the vector of all ones and *Q* is an l by l positive semidefinite matrix. Element *(i,j)th* of *Q* is given by $Q_{ij} \equiv y_i y_j K(x_i,x_j)$, where $K(x_i,x_j) \equiv \phi(x_i)^T \phi(x_j)$ is the kernel function. Then, the decision function can be given as: $\mathrm{sgn}(w^T\phi(x)+b) = \mathrm{sgn}(\sum_{i=1}^{l}\alpha_i y_i K(x_i,x)+b)$. The above definition is employed to minimize the predictable error in SVM learning. Several kernel functions can be used in SVM learning, including linear kernel, polynomial kernel, sigmoid kernel, radial basis function (RBF) kernel (also called Gaussion kernel) etc. This paper selects RBF kernel as the SVM kernel function, i.e., RBF kernel $K(x,y) = \exp(-\gamma \| x-y \|^2), \gamma > 0$. The RBF kernel nonlinearly maps the training data into a higher dimensional space, so it can handle non linear relation between the class labels and the attributes. Keerthi and Lin [10] prove that a linear kernel with a penalty parameter $\tilde{C}$ has the same performance as the RBF kernel with (C,γ) (*C* is the penalty parameter, γ is the kernel parameter). In addition, the application of sigmoid kernel in SVM learning and the similar parameters to RBF kernel are given by [9].

It is known that the number of hyperparameters influences the complexity of model selection. In RBF kernel, $0 < K_{ij} \le 1$. However, for polynomial kernels, there are two cases: $\gamma x_i^T x_j + \gamma > 1$ means its value is infinite; $0 < \gamma x_i^T x_j + \gamma < 1$ is the opposite. The authors of [12] believe that since there is no inner product of two vectors, the application of the sigmoid kernel has some limitations.

As mentioned above, there are two hyperparameters in model selection for the RBF kernel [11]: the penalty parameter C and the kernel width γ. We can improve SVM learning by optimizing the parameter pair (C,γ). Several methods can be used to compute these two parameters [12]. (C,γ) can be computed in the same way as $(\log C, \log\gamma)$. When searching for a good set of $\log C$ and $\log\gamma$, it is usual to form a two-dimensional uniform grid $(n \times n)$ in the training space to find a set of $(\log C, \log\gamma)$ which has the smallest generalization error in SVM classification. This method is called *grid search method*. This method searches for n^2 pairs of (C,γ).

Keerthi and Lin [11] propose a simple and efficient heuristic method for computing (C,γ) . It forms a unit slope which cuts through the middle part of the good region and searches for a good set of (C,γ) within the good region. Suppose line $\tilde{C}$ is the optimal penalty parameter of linear SVM, follow the procedures below (call it *bilinear search method*) to compute $\tilde{C}$:

·Search for the best $C$of linear SVM and denote it as $\tilde{C}$.

·Fix $\tilde{C}$, and search for the satisfying (C,γ) $\log\gamma = \log C - \log\tilde{C}$ using the RBF kernel.

Keerthi and Lin [11] have difficulty in deciding the range of $\log C$ for the computation of $\tilde{C}$ in the first step. This paper employs an *improved bilinear search method* to solve this problem by searching for $\log\gamma = \log C - \log\tilde{C}$ with $0.5\tilde{C}$, $\tilde{C}$ and $2\tilde{C}$ respectively. The best $\tilde{C}$ is computed from the range of $\log C$.

Grid search is time-consuming. Based on the bilinear search method adopted by Keerthi and Lin [11], we propose an improved bilinear search method to decide (C,γ). First, identify a 'better' region (the range of $\log C$ is larger than that of [11]), and compute a (C_1,γ_1) pair. Then, invoke *an improved grid search* to obtain a better pair (C_2,γ_2) than (C_1,γ_1) for accurate prediction. It is stated that the grid search method can be improved by a improved grid search method, to obtain a better set of (C,γ) and an accurate SVM model.

3 Bilinear grid search algorithm

In SVM learning with RBF kernel, several methods can be applied to compute (C,γ). As aforementioned, (C_1,γ_1) of the (coarse) grid search can be optimized using the improved grid search, to acquire a more suitable set of (C_2,γ_2) for training accurate SVM models. Bilinear search method is used to search for the best parameter $\tilde{C}$ in linear SVM. These parameters, $0.5\tilde{C}$, $\tilde{C}$ and $2\tilde{C}$ are acquired in this paper, and computed with the related $\gamma_{0.5}, \gamma_1, \gamma_2$ respectively. In [13], the advantage of determining (C,γ) with the improved bilinear search method is also presented.

Due to the complexity of search space, grid search method requires n^2 pairs of (C,γ) to be tried, while bilinear search method requires only $2n$. Compared to bilinear search, grid search method usually has a higher accuracy of prediction. The bilinear grid search method proposed in this paper retains the advantages of these two methods: it attempts to search for (C,γ) with less training points while maintaining not the accuracy of SVM models. Details algorithm is presented as follows:

First, compute the best C using bilinear method and denote it as $\tilde{C}$. Then, compare $0.5\tilde{C}$, $\tilde{C}$ and $2\tilde{C}$ to search for the best parameter pair (C_{bt}, γ_{bt}) among (C_j, γ_j), using the improved bilinear search method. According to (C_{bt}, γ_{bt}), invoke a finer search using aimproved grid search smaller grid spacing of $2^{0.25)}$ in the scope of [2^{-2}, 2^{2}] around the best (C_{bt}, γ_{bt}) to obtain $(C_{final}, \gamma_{final})$. Denote $(C_{final}, \gamma_{final})$ as the optimized (C, γ) and use it to train a SVM model with RBF kernel and acquire the objective SVM model with the highest accuracy.

Algorithm: Bilinear grid search algorithm

```
Input: Training examples
Output: Classification model with the best accuracy
Begin
  Map the training data to the SVM space;
  Select a linear kernel SVM;
  Search for the best $C$ of linear SVM and call it $\tilde{C}$;
  for $j = 0.5\tilde{C}, \tilde{C}, 2\tilde{C}$ do
    Compute the $\gamma_j$ according to $\log\gamma = \log C - \log\tilde{C}$ using the RBF kernel;
  Select the best $(C_{bt}, \gamma_{bt})$ from the $(C_j, \gamma_j)$;
  For $(C_{bt}, \gamma_{bt})$, invoke improved grid search to do
    For $k = 2^{-2}$ to $2^{2}$ step $2^{0.25}$;
      Compute their $(C_k, \gamma_k)$;
    Select the best $(C_{final}, \gamma_{final})$ among $(C_k, \gamma_k)$;
  Train the SVM with RBF kernel using $(C_{final}, \gamma_{final})$;
  Obtain the classification model with the best accuracy
End.
```

In the process, evaluate the accuracy of all models with 10-fold cross-validation. For grid search method, we uniformly discretize (C, γ) within a [-10, 16] × [-15, 11] region i.e., $27^2 = 729$ training points. For bilinear search method, we search for $\tilde{C}$ using the value of uniformly spaced $\log C$ in [-10, 16]. Then, discretize [-15, 11] as values of $\log\gamma$ and check all points that matches $\log\gamma = \log C - \log\tilde{C}$ (compared with bilinear search method, the improved bilinear search method takes all three values of $0.5\tilde{C}$, $\tilde{C}$, and $2\tilde{C}$ to satisfy the bilinear equation).

4 Experimental results

The proposed bilinear grid algorithm has been evaluated and compared to existing algorithms. This section presents the experimental results.

Classification accuracies of grid search, bilinear search, improved bilinear search, and bilinear grid search are compared in this section. The experiments employ 10 sets of data chosen from UCI database [15]. These data are trained on LIBSVM [16] with four methods respectively, namely the grid search method, bilinear search method, improved bilinear search method and bilinear grid search method.

Table 1 presents the basic information of the 10 data sets. For example, the Breast-cancer (BC) data set includes 9 attributes, 683 examples, and 2 classes. Table 2 demonstrates the model errors of these 4 different search algorithms. Figures inside the parentheses indicate set $(C_{final}, \gamma_{final})$, which is computed by our proposed bilinear grid search. It shows that bilinear grid algorithm is very competitive compared with grid search in terms of testing error. Among these 10 data sets, both bilinear grid search and grid search obtain the same accuracy on 6 data sets (i.e., Breast-cancer, Iris, Vowel, Wine, Wpbc, Zoo); Bilinear grid search trains more accurate models than grid search on 2 data sets (Credit-screening, Letter-recognition). On Diabetes and Wdbc, bilinear grid search obtains higher accuracy compared with grid search, even though the latter obtains higher accuracy during training. On all the 10 data sets, bilinear grid search learns more accurate models than bilinear search and improved bilinear search.

Data set	attribute	example	class
Breast-Cancer (BC)	9	683	2
Credit-screening (CS)	15	690	2
Diabetes (DIAB)	8	768	2
Iris(IR)	4	150	3
Letter-recognition (LR)	16	20000	26
Vowel(VO)	10	528	11
Wdbc	10	569	2
Wine	13	768	3
Wpbc	33	194	2
Zoo	16	101	7

Table 1: Training data set.

Table 3 shows the number of training SVMs required by these 4 different algorithms. For all the 10 data sets, grid search needs to run the same

Data	Grid search	Bilinear search	IB search	BG search
BC	0.027(-3,-3)	0.030(-4,-2)	0.030(-4,-2)	0.027(2.8,-3)
CS	0.130(2,-1)	0.139(3,-1)	0.130(2,-1)	0.128(2.5,-1.5)
DIAR	0.225(0,-4)	0.244(-3,0)	0.234(-3,-1)	0.226(-1.8,-2.5)
IR	0.026(2,-3)	0.046(-2,-2)	0.026(0,-1)	0.026(0,-1)
LR	0.020(10,2)	0.019(5,1)	0.019(6,1)	0.019(6,1)
VO	0.003(3,2)	0.003(6,1)	0.003(6,1)	0.003(6,1)
Wdbc	0.019(3,-5)	0.040(-3,-1)	0.031(-2,-1)	0.021(-0.5,-2.3)
Wine	0.005(0,-2)	0.028(-2,0)	0.016(-2,-1)	0.005(0,-2)
Wpbc	0.164(6,-5)	0.201(1,-3)	0.190(2,-3)	0.164(3.8,-3.5)
Zoo	0.039(10,-9)	0.138(-2,-3)	0.049(0,-2)	0.039(1.3.-3)

IB search: Improved Bilinear search. BG search: Bilinear Grid search.

Table 2: Model error comparison of bilinear grid search with other search methods.

training SVMs for 729 times because it trains SVMs with the same grid 27^2. Both bilinear search and the improved bilinear search require a smaller number of training SVMs. The number of training SVMs of bilinear grid search algorithm is much smaller compare with grid search algorithm.

From Tables 2 and 3, we can see that bilinear grid search algorithm has the best performance in terms of accuracy and the number of training SVMs. For large data sets, bilinear grid search algorithm is preferable over grid search algorithm, since the former checks fewer points on the (log C, log γ) two-dimension plane, thus saves computing time. The experimental results show that, with the largest training SVMs, grid search method generates higher accuracy of prediction than bilinear search method because the latter searches a smaller number of training SVMs. Bilinear grid search method retains the advantages of both bilinear search and grid search, thus reducing the number of training SVMs (compared with bilinear search and grid search method), while obtaining a competitive accuracy of prediction. Therefore, it is preferable over grid search.

Data set	Grid search	Bilinear search	IB search	BG search
BC	729	47	87	376
CS	729	53	105	394
DIAB	729	46	84	373
IR	729	49	93	382
LR	729	53	105	394
Vowel	729	54	106	395
Wdbc	729	47	87	376
Wine	729	44	83	372
Wpbc	729	53	105	394
Zoo	729	50	96	385

IB search: Improved Bilinear search. BG search: Bilinear Grid search.

Table 3: Comparison of SVM training times.

5 BGSM on protein secondary structure prediction

Due to potential homology between proteins in the training and testing set, the selection of protein database for secondary structure prediction is complicated. Homologous proteins in the database may generate misleading results. This is because in some cases the learning method memorizes the training set. Therefore protein chains without significant pairwise homology are used for developing our prediction model. To have a fair comparison, we train and test the same 130 protein sequences used by Rost and Sander [17] and Jung-Ying Wang [18]. These proteins, taken from the HSSP (Homology-derived Structures and Sequences alignments of Proteins) database [19], all have less than 25% of the pairwise similarity and more than 80 residues. Meanwhile, we also train and test the same seven-fold cross-validation are used in Rost and Sander [17] and Jung-Ying Wang[18]. Table 4 lists the 130 protein sequences used for seven-fold cross-validation.

The secondary structure assignment was done using the DSSP (Dictionary of Secondary Structures of Proteins) algorithm [20], which distinguishes between the eight secondary structure classes. The eight classes are reclassified into the following three classes: H (α -helix), I (π -helix), and G (310-helix) are classified as helix (α), E (extended strand) as β -strand (β), and all others as coil (c). Table 5 lists the reclassification process. Note that different classification methods influence the prediction accuracy to some extent, as discussed by Cuff and Barton [21]. For an amino acid sequence, the objective of secondary structure prediction is to predict a secondary structure state (α , β , coil) for each residue in the sequence.

Set A	256b_A 2aat 8abp 6acn 1acx 8adh 3ait 2ak3_A 2alp 9api_A 9api_B 1azu 1cyo 1bbp_A 1bds 1bmv_1 1bmv_2 3blm 4bp2
Set B	2cab 7cat_A 1cbh 1cc5 2ccy_A 1cdh 1cdt_A 3cla 3cln 4cms 4cpa_I 6cpa 6cpp 4cpv 1crn 1cse_I 6cts 2cyp 5cyt_R
Set C	1eca 6dfr 3ebx 5er2_E 1etu 1fc2_C fdl_H 1dur 1fkf 1fnd 2fxb 1fxi_A 2fox 1g6n_A 2gbp 1a45 1gd1_O 2gls_A 2gn5
Set D	1gp1_A 4gr1 1hip 6hir 3hmg_A 3hmg_B 2hmz_A 5hvp_A 2i1b 3icb 7icd 1il8_A 9ins_B 1l58 1lap 5ldh 1gdj 2lhb 1lmb_3
Set E	2ltn_A 2ltn_B 5lyz 1mcp_L 2mev_4 2or1_L 1ovo_A 1paz 9pap 2pcy 4pfk 3pgm 2phh 1pyp 1r09_2 2pab_A 2mhu 1mrt 1ppt
Set F	1rbp 1rhd 4rhv_1 4rhv_3 4rhv_4 3rnt 7rsa 2rsp_A 4rxn 1s01 3sdh_A 4sgb_I 1sh1 2sns 2sod_B 2stv 2tgp_I 1tgs_I 3tim_A
Set G	6tmn_E 2tmv_P 1tnf_A 4ts1_A 1ubq 2utg_A 9wga_A 2wrp_R 1bks_A 1bks_B 4xia_A 2tsc_A 1prc_C 1prc_H 1prc_L 1prc_M

Table 4: 130 Protein sequences name used in experiments

Structural character	Structural name
H	α -helix
G	310-helix
I	π -helix
E	Extended strand
B	β -bridge
T	Turn
S	Bend
C	The rest

(a)

Structural character	Structural name
H	helix
E	strand
C	The rest

(b)

Structural character before conversion	Structural character after conversion
H	H
I	
G	
E	E
B	C
T	
S	
C	

(c)

Table 5: (a) Eight types structural character and name (b) Three classes structural character and name (c) Reclassification between eight types and three classes.

For fair comparison, we train and test same 130 protein sequences used by Rost, Sander and Jung-Ying Wang. These proteins are taken from the HSSP database. The secondary structure assignment was done according to the DSSP algorithms, which are distinguished by eight secondary structures classes, and then three classes.

Moving window and multiple alignment methods are used for encoding. We apply the moving window method for the 17 neighbouring residues in our study. Each window has 21 possible values, including 20 amino acids and a null input. Therefore, the number of data points is the same as the number of residues when each data point has 21×17=357 values. Before testing these proteins, we employ multiple alignment method to acquire more evolutionary information and protein family information. Having replaced single sequence orthogonal coding, input vector is obtained by aligning the similarity between unknown sequences and known sequences. Then, we can obtain evolutionary information by finding out whether these sequences are homologous.

Figure 1 is an example of using evolutionary information for encoding. we align four proteins. In the gray column， the based sequence has residue 'N' while the multiple alignments in this position are 'N', 'A', 'S' and 'E' (indicating point of deletion in this sequence). Finally, we treat frequencies as the values of output coding. Therefore, the coding scheme in this position is as follows: A = 0.2, S = 0.2, E = 0.2, N=0.4.

Prediction is conducted for the central residue in the windows. In order to allow the moving window to overlap the amino- or carboxyl-terminal end of the protein, a null input was added to each residue. Therefore, each data point has 21×17 = 357 values and each data can be represented as a vector. Note that data set RS130 consists of 24,387 data points in three classes where 47% are coil, 32% are helix, and 21% are strand.

An important fact about prediction is that training errors are not significant; only test errors (i.e. accuracy for predicting new sequences) count. Therefore, it is important to estimate the overall performance of a learning method. Previous research proposed different methods to evaluate accuracy. The most common method applied in secondary structure prediction is the overall three-state accuracy (Q_3). It is defined as the ratio of correctly predicted residues to the total number of residues in the database under consideration.

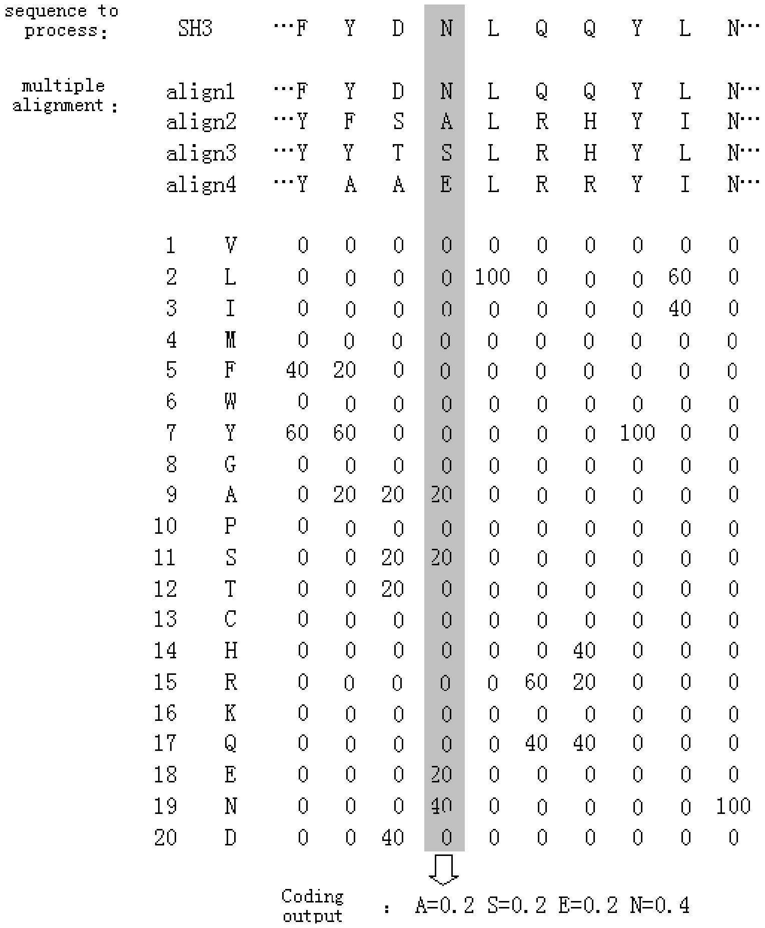
Q_3 is calculated by

$$Q_3 = \frac{q_\alpha + q_\beta + q_{coil}}{N} \times 100 \quad ,$$

where N is the total number of residues in the test data sets, and q_s is the number of residues of secondary structure type *s* that are predicted correctly.

We carry out several experiments to optimiza hyperparameters using bilinear grid search method. The ranges of C and γ are both [2^{-8}, 2^{-7},. . . , 2^{8}], and cross-validation fold is 7.

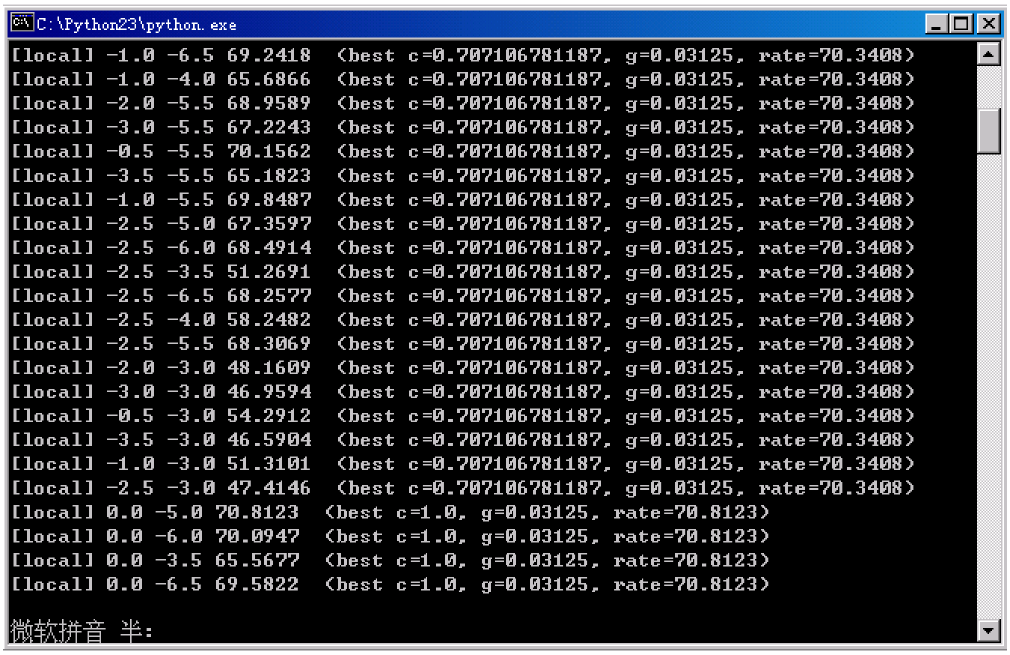
Fig.2 and Fig.3 are the running result charts of command-line and contour. In the chart of command-line, <best c=1.0 g=0.03125, rate=70.8123> the best parameter (C,γ) = (1.0,0.03125), and its classification accuracy is 70.8123%.



sequence to process:	SH3	…F	Y	D	N	L	Q	Q	Y	L	N…
multiple alignment :	align1	…F	Y	D	N	L	Q	Q	Y	L	N…
	align2	…Y	F	S	A	L	R	H	Y	I	N…
	align3	…Y	Y	T	S	L	R	H	Y	L	N…
	align4	…Y	A	A	E	L	R	R	Y	I	N…
1	V	0	0	0	0	0	0	0	0	0	0
2	L	0	0	0	0	100	0	0	0	60	0
3	I	0	0	0	0	0	0	0	0	40	0
4	M	0	0	0	0	0	0	0	0	0	0
5	F	40	20	0	0	0	0	0	0	0	0
6	W	0	0	0	0	0	0	0	0	0	0
7	Y	60	60	0	0	0	0	0	100	0	0
8	G	0	0	0	0	0	0	0	0	0	0
9	A	0	20	20	20	0	0	0	0	0	0
10	P	0	0	0	0	0	0	0	0	0	0
11	S	0	0	20	20	0	0	0	0	0	0
12	T	0	0	20	0	0	0	0	0	0	0
13	C	0	0	0	0	0	0	0	0	0	0
14	H	0	0	0	0	0	0	40	0	0	0
15	R	0	0	0	0	0	60	20	0	0	0
16	K	0	0	0	0	0	0	0	0	0	0
17	Q	0	0	0	0	0	40	40	0	0	0
18	E	0	0	0	20	0	0	0	0	0	0
19	N	0	0	0	40	0	0	0	0	0	100
20	D	0	0	40	0	0	0	0	0	0	0

Coding output : A=0.2 S=0.2 E=0.2 N=0.4

Figure 1: An example of using evolutionary information for coding secondary structure.



```
C:\Python23\python.exe
[local] -1.0 -6.5 69.2418  (best c=0.707106781187, g=0.03125, rate=70.3408)
[local] -1.0 -4.0 65.6866  (best c=0.707106781187, g=0.03125, rate=70.3408)
[local] -2.0 -5.5 68.9589  (best c=0.707106781187, g=0.03125, rate=70.3408)
[local] -3.0 -5.5 67.2243  (best c=0.707106781187, g=0.03125, rate=70.3408)
[local] -0.5 -5.5 70.1562  (best c=0.707106781187, g=0.03125, rate=70.3408)
[local] -3.5 -5.5 65.1823  (best c=0.707106781187, g=0.03125, rate=70.3408)
[local] -1.0 -5.5 69.8487  (best c=0.707106781187, g=0.03125, rate=70.3408)
[local] -2.5 -5.0 67.3597  (best c=0.707106781187, g=0.03125, rate=70.3408)
[local] -2.5 -6.0 68.4914  (best c=0.707106781187, g=0.03125, rate=70.3408)
[local] -2.5 -3.5 51.2691  (best c=0.707106781187, g=0.03125, rate=70.3408)
[local] -2.5 -6.5 68.2577  (best c=0.707106781187, g=0.03125, rate=70.3408)
[local] -2.5 -4.0 58.2482  (best c=0.707106781187, g=0.03125, rate=70.3408)
[local] -2.5 -5.5 68.3069  (best c=0.707106781187, g=0.03125, rate=70.3408)
[local] -2.0 -3.0 48.1609  (best c=0.707106781187, g=0.03125, rate=70.3408)
[local] -3.0 -3.0 46.9594  (best c=0.707106781187, g=0.03125, rate=70.3408)
[local] -0.5 -3.0 54.2912  (best c=0.707106781187, g=0.03125, rate=70.3408)
[local] -3.5 -3.0 46.5904  (best c=0.707106781187, g=0.03125, rate=70.3408)
[local] -1.0 -3.0 51.3101  (best c=0.707106781187, g=0.03125, rate=70.3408)
[local] -2.5 -3.0 47.4146  (best c=0.707106781187, g=0.03125, rate=70.3408)
[local] 0.0 -5.0 70.8123  (best c=1.0, g=0.03125, rate=70.8123)
[local] 0.0 -6.0 70.0947  (best c=1.0, g=0.03125, rate=70.8123)
[local] 0.0 -3.5 65.5677  (best c=1.0, g=0.03125, rate=70.8123)
[local] 0.0 -6.5 69.5822  (best c=1.0, g=0.03125, rate=70.8123)

微软拼音 半:
```

Figure 2: the result chart of command-line.

Table 6 lists the accuracy of different methods on RS130 data set. The average accuracy for bilinear grid search method is 70.8%, which is competitive compared with those methods proposed by Rost, Sander and Jung-Ying Wang. The average accuracy for the method of Rost and Sander [17] is 68.2% , which employs neural networks for encoding. Other techniques must be incorporated in order to increase accuracy to 70.0%. Jung-Ying Wang utilizes basic SVM to [18] obtain 70.5% of the accuracy .

The experiment used the same data set (including the type of alignment profiles) and secondary structure

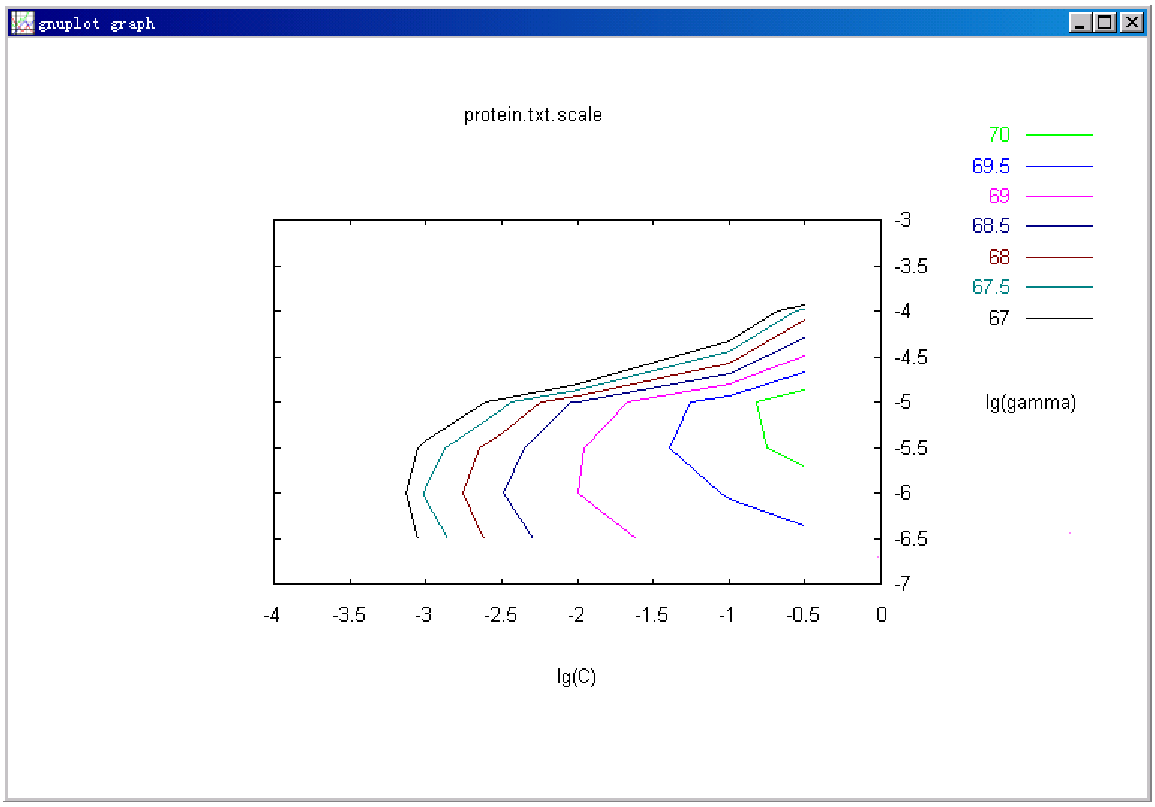


Figure 3: The result chart of contour.

definition (reduction from eight to three secondary structures) as those employed by Rost and Sander [17], and Jung-Ying Wang [18]. The same accuracy assessment of the prior ones is used as well, so as to ensure the fairness of comparison.

Different methods	Secondary structure prediction accuracy %
Neural networks	68.2
Neural networks incorporated With other techniques	70.0
SVM	70.5
SVM with bilinear grid Search method	70.8

Table 6: Comparison of different methods' accuracy on RS130 data set.

6 Conclusion

In this paper, we demonstrate an approach of optimization in SVM learning. The proposed bilinear grid search method can effectively improve learning performance and enhance the accuracy of prediction. A comparison has been made between grid search method, bilinear search method and bilinear grid search method when selecting optimal parameters for RBF kernel. Experiment results prove that the proposed algorithm retains the advantages of both bilinear search method and grid search method. With the application of BGSM, the protein secondary structure prediction also obtains better learning accuracy compared with other algorithms.

Acknowledgement

This work was supported in part by the National Natural Science Foundation of China (NSFC) under Grant No.61273225, No.61100055 and No.31201121, the Natural Science Foundation of Hubei Province under Grant No. 2011CDB233.

Reference

[1] Vladimir N. Vapnik (1998). Statistical learning theory. J. Wiley & Sons, New York.

[2] C. Cortes, Vladimir N. Vapnik (1995). Support vector networks. Machine Learning, Vol.20, No.3, pp.273-297.

[3] Vladimir N. Vapnik (2000). The Nature of Statistical Learning Theory (Second Edition). Springer Press.

[4] B Schölkopf, AJ Smola (2002). Learning with kernels. MIT Press.

[5] Kai Zhang, Tsang, I.W. , Kwok, J.T.(2009). Maximum margin clustering made practical. IEEE Transactions on Neural Networks, Vol.20 , No.4, pp. 583 - 596.

[6] E Blanzieri, F Melgani (2008). Nearest neighbor classification of remote sensing images with the maximal margin principle. IEEE transaction on Geoscience and Remote Sensing, Vol.46, No.6, pp.1804-1811.

[7] GB Huang, QY Zhu, CK Siew (2006). Extreme learning machine: theory and applications. Neurocomputing.

[8] S. Fine (2001). Efficient SVM training using low-rank kernel representations. Journal of Machine Learning Research, Vol.2, pp.243-264.

[9] K. M. Lin and C. J. Lin(2003), A Study on reduced support machines. IEEE Trans. on Neural Computation, Vol.14, No.6, pp. 1449-1559.

[10] Li. Lin and Zhang Xiaolong (2005). Optimization of SVM with RBF Kernel. Computer Engineering and Applications(in Chinese), Vol.29, No. 10, pp.190-193.

[11] S. S. Keerthi, C. J. Lin(2003). Asymptotic behaviours of support vector machines with Gaussian kernel. Neural Computation, No.5:1667–1689.

[12] O. Chapelle, V. Vapnik et al(2002). Choosing multiple parameters for support vector machines. Machine Learning, Vol.46, pp.131–159.

[13] P. Wang, X. Zhu(2003). Model Selection of SVM with RBF Kernel and its Application. Computer Engineering and Applications(in Chinese), Vol.24, pp.72–73.

[14] H. T. Lin, C. J. Lin(2003), A Study on Sigmoid kernels for SVM and the training of Non-PSD kernels by SMO-type methods. Technical Report, National Taiwan University.

[15] Blake C., Merz C.(2013), UCI Repository of Machine Learning Databases. http://www.ics.uci.edu /mlearn/MLRepository.html, Dept. of Information and Computer Science, University of California.

[16] C. C. Chang, C. J. Lin (2013). LIBSVM: A library for support vector machines. Software Available on-line at: http:// www.csie.ntu.edu.tw/~cjlin /libsvm /index.html .

[17] B. Rost and C. Sander(1993). Prediction of protein secondary structure at better than 70% accuracy. Journal of Molecular Biology, Vol.23, No.2, pp.584–599.

[18] Jung-Ying Wang(2002). Application of Support Vector Machines in Bioinformatics. Taipei: Department of Computer Science and Information Engineering, National Taiwan University.

[19] http://www.cmbi.kun.nl/gv/hssp.

[20] W. Kabsch and C. Sander(1983). Dictionary of protein secondary structure: Pat-tern recognition of hydrogen-bonded and geometrical features. Biopolymers, Vol.22, No.12, pp.2577–2637.

[21] J. A. Cuff and G. J. Barton(1999). Evaluation and improvement of multiple sequence methods for protein secondary structure prediction. Proteins: Struct. Funct. Genet., Vol.34, pp.508–519.

An Imperative of a Poorly Recognized Existential Risk: Early Socialization of Smart Young Generation in Information Society

Vladimir A. Fomichov
Faculty of Business Informatics, National Research University Higher School of Economics
Kirpichnaya str. 33, 105187 Moscow, Russia;
E-mail: vfomichov@hse.ru; Web: http://www.hse.ru/en/org/persons/67739

Olga S. Fomichova
State Educational Centre “Dialogue of Sciences”, Universitetsky prospect 5, 119296 Moscow, Russia
E-mail: vfomichov@gmail.com

Keywords: misuse of information and communication technologies, early socialization of children, positive psychology movement, informational-aesthetic conception of developing cognitive-emotional sphere, theory of dynamic conceptual mappings, methods of emotional-imaginative teaching, levels of consciousness expanded model, system of values, cognitive engagement

Received: January 28, 2014

The birth of this paper was motivated by the foundation in July 2013 of the Centre for the Study of Existential Risk (CSER) at the University of Cambridge (UK). The first aim of the paper is to attract the attention of the researchers and educators throughout the world to a poorly recognized kind of existential risk: it is the dangerous consequences of misusing the information and communication technologies by adolescents in the context of permanently increasing intelligent capabilities of computers and an easy access to Internet. The second, principal aim of this paper is to propose a constructive way out: it is earlier than usually socialization of children in order to inscribe a deep awareness of social responsibility into the conceptual picture of the child and adolescent and to enable him/her to analyse the consequences of the fulfilled actions. The proposed way out is elaborated under the framework of a new scientific discipline called cognitonics. This way is provided by the System of the Methods of Emotional-Imaginative Teaching (the EIT-system). The constructive core of this paper consists of two parts. The first part considerably expands the Level of Consciousness model proposed by P.D. Zelazo in 2004. It considers fours levels of the development of conscious control of thought, emotion, and action and covers the child’s age from one to four years. Our model is based on the EIT-system and introduces three additional levels, where the seventh level is called the level of enhanced awareness of social agreements and social responsibility. Our model covers the ages from five – six to 13 – 14 years. The second part of this paper’s constructive core presents a new look at the process of education when the values of the student act like a lighthouse for the teacher at the moment of presenting material and arranging the process of education, the process of acquiring knowledge.

Povzetek: Kako socializirati mlado generacijo v informacijski družbi?

1 Introduction

The technical characteristics of computers have been improving since their birth and their intelligent capabilities since the end of the 1960s, when the scientific-technical field “artificial intelligence” was born. Now computer systems are able to understand a broad range of texts in natural language (English, Russian, Chinese, Japanese, etc.), to recognize visual images, and to do a lot of other things.

The considerations of the kind became the reason for introducing in 1993 the notion of singularity by V. Vinge [30], a scientific fiction writer. It is the moment when intelligent capabilities of computer systems will exceed the capabilities of human beings, and the further development of computer systems will be determined by the needs of the global family of computers but not by the humans. The starting point for V. Vinge was the ideas initially formulated by John von Neumann and I.J. Good [18]. Later the term “the singularity” was popularized by R. Kurzweil [22], an inventor and futurist.

The achievements obtained during last decades on the way of designing intelligent computer systems and in several other fields, in particular, in genetics, biotechnology, nanotechnology, became the reasons for introducing the notion of existential risk and for founding in July 2013 the Centre for the Study of Existential Risk (CSER) [3] as a research centre at the University of Cambridge (UK). Now CSER is hosted within the Cambridge’s Centre for Research in the Arts, Social Sciences, and Humanities. The goal of CSER is to study possible catastrophic threats caused by the existing or future technology. The co-founders of CSER are Dr. Huw Price (Bertrand Russell Professor of Philosophy, Cambridge), Dr. Marteen Rees (an Emeritus Professor of Cosmology and Astrophysics, Cambridge and former President of the Royal Society), and Jaan Tallinn (a computer programmer and a co-founder of Skype) [3, 20, 23].

Professor Price expressed the opinion that "sometime in this or the next century intelligence will escape from the constraints of biology". In this case "we're no longer the smartest things around, and may be at the mercy of "the machines that are not malicious, but machines whose interests don't include us" [20]. According to Price, many people consider his concerns as far-fetched. However, since the risks are very serious and we don't know the time parameters, it is necessary to put the problem into the focus of attention of international scientific community [20].

We completely agree with this opinion. However, we believe that there exists at least one additional kind of existential risk, and it is poorly recognized by international scientific community. This practically unnoticed global problem is the expanding negative consequences of misusing information and communication technologies (ICT) by a part of smart young generation. During last decade, one has been able to observe numerous cases when the hackers-teenagers managed to cause a very considerable damage to significant social objects and even military objects.

Let's consider a risk that can emerge many years before the singularity. We will take into account the well known fact: the technical characteristics of computers double every two years. Imagine that one or several decades later a person with high computer skills (may be, an adolescent or a group of adolescents) will pose a socially dangerous task to a multi-agent system consisting of the future computers with very high intelligent capabilities. Then the damage from achieving the formulated goal may be very high, even immensely high.

In other words, it is easy to assume that a person (regardless the age, spiritual maturity which includes the developed feeling of responsibility, intelligence maturity, which suggests the improved cognitive mechanisms of information processing) will be able to take power and influence the life of community, society, people all over the world due to his/her well-improved skills of using various ICT. It may happen even with school children, because every new generation born in the information society (IS) is much more skilful than the previous one, and they have much more time to improve their skills, because since the early childhood it is as usual as walk and talk for all the children.

On the other hand, the curiosity and strong aspiration to discover the digital world are underpinned by the common (for their age) desire to emulate grown-ups and become as smart and powerful as grown-ups are, or even much smarter and much more powerful in comparison with the people belonging to previous generations.

Even nowadays the teachers in various countries complain that school children are smarter and more skilful as they are. It discourages the teachers and makes the relationships with school children of the kind much more complicated.

In various countries throughout the world, there is an age requirement for allowing an adolescent to drive a car. Obviously, the reason is dramatic consequences both for other people and for the young person in case a socially immature or a technically insufficiently qualified person will drive a car.

In the modern IS, a considerable part of teenagers possesses very high computer skills, and they have access to Internet and its immense technical possibilities. But very often these teenagers are socially rather immature.

For instance, UK is the country where the term "screenager" (instead of "teenager") was born [25]. It means that very many teenagers in UK spend much more time for the communication with computers than with people. This fact allows us to conjecture that very many teenagers in UK possess high computer skills. However, the psychologists discovered in 2013 that many boys and girls at the age from 18 until approximately 25 years are rather socially immature (see Figure 1) and should be treated by the psychologists as teenagers. A part of university students living in one home with their parents considerably increased. The parents were recommended to increase the socialization of their children – students by means of asking them to wash their dresses, to pay various receipts, etc. [31]. Taking this situation into account, we may imagine the cases of misusing ICT not due to any bad intention but because the consequences have not been thought over in detail.

The problem looks like an iceberg, and the humans in general way may become the passengers of "Titannik", because they don't expert an iceberg on the way.

This paper continues the line of the articles [9, 10, 15]. Metaphorically speaking, the aim of this series of publications and of this paper is to propose the kind and the parameters of a manoeuvre preventing the collision of our information society with the iceberg of described sort. This manoeuvre is much earlier socialization of children than it is done now throughout the world; that is, it is a way of early and deep inscription of the notion "responsibility" into the child's conceptual picture.

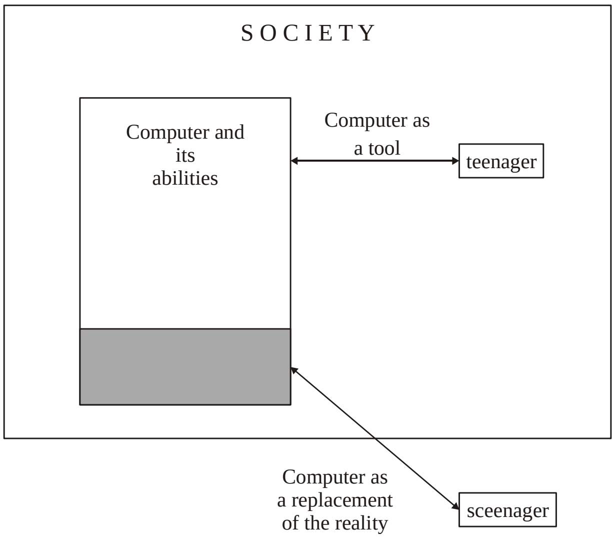

SOCIETY
Computer and its abilities
Computer as a tool
teenager
Computer as a replacement of the reality
sceenager

Figure 1: Screenagers and society

Our key idea is to inscribe into the world's conceptual picture of the child a deep awareness of social agreements and the feeling of social responsibility *before* the transition age 11 – 12 years (before the age of conflicts and evoking sexuality). The proposed way of early children's socialization has been elaborated under the framework of Cognitonics [8-12]. It is a new scientific discipline, its first aim is explicating the distortions in the development of the personality and national cultures caused by the peculiarities of information society and globalization. The second (principal) aim of Cognitonics is coping with these distortions in different fields by means of elaborating systemic solutions for compensating the negative implications for the personality and society of the stormy development of ICT and globalization processes, in particular, for creating cognitive-cultural preconditions of the harmonic development of the personality in the information society and for ensuring the successive development of national cultures and national languages.

The goals and ideas of Cognitonics have evoked a vivid interest of the scholars from over 20 countries located in Africa, Asia, Europe, North America and South America. Due to this interest, three successful international scientific conferences on Cognitonics were organized in the years 2009, 2011, 2013 under the framework of the international scientific multiconference "Information Society" (Slovenia, Ljubljana, Jozef Stefan Institute) [1, 2, 16].

The constructive core of this paper consists of two parts. The first part (Sections 2 – 6) considerably expands the Level of Consciousness model proposed by P. D. Zelazo in 2004 [32]. It considers fours levels of the development of conscious control of thought, emotion, and action and covers the child's age from one to four years. Our model introduces three additional levels, when the seventh level is called the level of enhanced awareness of social agreements and social responsibility. Our model covers the ages from five – six to 13 – 14 years.

The second part (Sections 7 – 11) of this paper's constructive core presents a new look at the process of education when the values of the student act like a lighthouse for the teacher at the moment of presenting material and arranging the process of education, the process of acquiring knowledge. Four discoveries underpinning the proposed complex method of early socialization of children in modern IS are described. *The first discovery* is the fundamental conclusion that young children and adolescents can be attributed to one of two groups (children with preponderance of material values and children with preponderance of sublime values), and different methods of teaching should be developed and used for achieving educational success for each of these two groups.

The second discovery is an original method of splitting young children in two groups of mentioned kinds. *The third discovery* is two developed different practical approaches to teaching allowing to achieve educational success for each of two groups. *The fourth discovery* is the proposed notion of cognitive engagement and original methods enabling a teacher to successfully reach the goals of teaching in each of two groups by means of realizing cognitive engagement of students at lessons. As a result, a new psychological and educational paradigm is presented.

2 Central ideas of positive psychology movement

Let's consider a broad scientific context being most appropriate for stating and assessing our original, many-staged method of early children's socialization. During the 1990s, it was possible to observe the steady growth of the number of children at school age in the developed countries encountering various social, emotional, and behavioral problems. Numerous observations provide the possibility to conjecture that, to a large extent, it was a consequence of more intensive interaction with computers at lessons and at home and of stormy Internet's expansion. Besides, the criminal films and horror films continued to negatively influence the mental state of very many children and adolescents, in particular, causing anxiety and aggression. These negative shifts became sufficiently noticeable by the beginning of the 2000s. According to [28], approximately one fifth of children and adolescents experienced problems showing their need for mental health services.

One of the consequences of this conclusion was the increased attention of the scholars to clarifying the extent of exposure to and use of media and electronic technology by very young children. A large-scale study described in [29] showed, in particular, the following alarming facts: (a) 27% of 5-6-year-old children used a computer during 50 minutes on average on a typical day; (b) more than one third of 3- to 6-year olds also have a television in their bedroom; 54% adults said that it frees up other TV in the house, that is why other family members can watch their own shows, 38% of adults indicated that it keeps the child occupied, so the parents can do things around the house.

As a principal way out in the current situation with mental health of the young generation, many psychologists indicated the importance of promoting children's social and emotional experience in schools. As a consequence, a new paradigmatic shift was observed in psychology: a shift from the accent on repairing weakness to the enhancement of positive qualities and preventing the problems before the moment when these problems arise [27]. As a result, the positive psychology movement was born, the principal objective of this movement is studying the positive features of humans development, in particular, investigating such significant traits of the person as "subjective well-being, optimism, happiness, and self-determination" [27, p. 9].

As a logical consequence, the task of promoting positive emotions in children and adolescents was posed [19]. The evidence obtained in the 2000s shows that a critical role in the success of children in school and in their social and emotional competence is played by self-regulation, in particular, by controlling attention and inhibiting aggressive reactions.

The publications on positive psychology allow to distinguish a factor being beneficial to well-being, this factor is called *mindfulness* [21]. According to the definition given in [24], it is a way of directing attention. Generalizing a number of available definitions of this concept, mindfulness can be characterized as the ability to maximally proceed from the context while taking decisions in any situations. It is the ability of paying attention to many details while elaborating a decision but not only "mechanically" following a number of prescribed rules, etc.

3 The key role of broad beauty appreciation

The analysis of scientific literature provides weighty grounds for concluding that the first educational system satisfying the criteria of a mindfulness-based program was born and well tested several years before the emergence of the term "mindfulness-based educational program". Such criteria are satisfied by the system of the methods of emotional-imaginative teaching (the EIT-system). The core of the EIT-system was elaborated by O.S. Fomichova in the first half of the 1990s and has been expanded in the second half of the 1990s and in the 2000s. This system is underpinned by our Theory of Dynamic Conceptual Mappings (the DCM-theory). This theory is stated in numerous publications both in English and Russian, in particular, in [5 - 15]. Both the DCM-theory and the EIT-methods form a principal part of the cognitonics constructive core.

The main component of the DCM-theory is an original informational-aesthetic conception of developing the cognitive-emotional sphere of the learners: young children, adolescents, and university students [6, 7, 9, 10].

On the one hand, this conception says that it is important to actively develop a broad spectrum of the learners' information processing skills. On the other hand, our conception has a number of original features. First of all, it is the idea of the necessity of inscribing, in a systemic way, the feeling of beauty into the world's conceptual picture of the child. Proceeding from our experience accumulated during 23 years, we consider the following educational processes as the principal instruments of achieving this goal: (a) early support and development of figurative (or metaphoric) reasoning; (b) teaching young children (at the age of 5 – 6) very beautiful language constructions for expressing the impressions from the nature; (c) a unified symbolic approach to teaching natural language (mother tongue and a foreign language), the language of painting, and the language of dance [6 - 15].

The next central idea is the conclusion about the necessity of passing ahead the development of soul in comparison with the development of reasoning skills. A well-developed feeling of beauty plays an especially significant role in the realization of this idea. Besides, it is very important to be aware of the fact that children should have enough time for the development of soul: the time for contemplation, for imbibing the beauty of the nature, etc., i.e. children should have time for self-paced activity [7, 14].

Much more information about our informational-aesthetic conception of developing the cognitive-emotional sphere of the learners can be found in [6, 7, 9, 10, 13, 14].

For the realization of these ideas, an interdisciplinary educational program has been developed by O.S. Fomichova. The elaborated program is intended for teaching children during twelve years, where the starting age is five to six years. The program has been personally tested in Moscow with permanent success by O.S. Fomichova over a period of 24 years. The total number of successfully taught students (young children and adolescents) exceeds nine hundred.

The program is implemented at extra-scholastic lessons of a foreign language (English), literature and poetry in mother tongue and second language, symbolic language of painting, communication culture, and classical dance. All these lessons are the links of one twelve-year-long educational chain. More details about the composition of the program can be found in [9 - 15].

A considerable role in the success of the educational program play regular (every semester) performances – a form of demonstrating knowledge and skills acquired during the current semester. The scripts for the performances are original and take into account both the learned materials and the individuality of each student. All personages in the performances are positive, no one negative. During each semester, all young students and teenagers master new elements of classical dance and train the known elements. Each performance includes singing world known songs in English, e.g., "White Christmas" and "Let It Snow" in case of winter performances. The highest form for demonstrating the acquired communication culture and culture of classical dance is the Christmas and Easter Big Balls, including the most part of students at the age from seven to twenty.

4 An environment of conceptual learning

One of the distinguishing features of our approach to this problem is that it is realized at lessons of a foreign language (FL) – English, where the mother tongue of children is Russian. The use of original analogies (being the parts of fairy-tales and thrilling stories) for teaching the English alphabet, the rules of reading, and the basic rules of English grammar contributes to developing associative abilities of children at the age of 5 – 6. The EIT-system provides *an environment of conceptual learning instead of a memorization-based one*. In particular, it is the principal distinguished feature of the developed original approach to teaching FL as an instrument of thinking.

The interesting stories about the life of verbs and other words (see, in particular, [5, 7, 8]) establish in the consciousness of the young child a mapping from the objects and situations of the real life or fairy-tale life to the domain of language entities (letters, sounds, verbs, nouns, pronouns, etc.). That is why the consciousness of the young child receives a considerable impulse to developing the ability to establish diverse analogies.

The other reason for using the lessons of FL is that (as a 24-year-long experience has shown) young children easier learn beautiful language constructions for describing the impressions from the nature than the equivalent constructions in mother tongue (see [6, 7]). The explanation of this phenomenon is that in the first case children don't feel any contradiction with the every-day use of language.

Example. Let's consider a fragment from the home composition "The Winter Day", it was written in English by an eight-year-old Russian speaking student Polina of the third year of studies in experimental groups:

THE KINGDOM OF THE WINTER

One winter day I was sitting near the window looking at the street covered with fresh clean snow. At first time, there was nothing so remarkable in that. Nor did I think it so very much out of the way to see that falling snowflakes, snow storm, the grey cloudy sky and the noisy crows. But when afterwards in the evening going to sleep I thought it over, it occurred to me that I ought to have wondered at this. I thought that the snow storm might be a wicked magician Winter, the grey sky with running clouds – his kingdom. Every beautiful princess that refused to be his wife because he was very angry and cruel was turned by him into a crow. And then their tears he turned into the falling snowflakes. And only the coming of the kind Fairy Spring can destroy this magic.

The realization of the teaching objectives mentioned above in this section is an important part of the first stage of supporting and developing the reasoning skills and creativity of the child. A map of cognitive transformations realized at this stage and the maps reflecting the next cognitive transformations can be found in [9, 10].

5 A known four-level model of consciousness development

It seems that the model proposed by Zelazo [32] can be considered as a good working instrument for studying the development of conscious control during the first – fourth years of childhood. This model, called the Levels of consciousness (LOC) model, emerged as a result of reflecting the experimentally discovered regularities of the development of conscious control of thought, action, and emotion. The model describes four transitions from one LOC to another, higher LOC, these transitions depend on age. Let us say about the zero LOC in case of newborn babies and very young children at the age less 11 – 12 months. Zelazo [32] characterizes the consciousness of this period as minimal consciousness; it is responsible for approach and avoidance behaviour based on pleasure and pain and is present-oriented, unreflective and doesn't operate with the Self-concept.

The principal distinguished feature of LOC1 is the emergence of concepts and of the connections between the perceived objects and concepts (playing the role of labels of experienced objects). LOC1 is called by Zelazo [32] as the *level of recursive consciousness*. LOC2 emerges at the end of the second year, the essence of the transition from LOC1 to LOC2 consists in the emergence of symbolic thinking, in children's awareness of Self. The signs of LOC2 are the first use of personal pronouns by children, their self-recognition in mirrors. Besides, children feel first self-conscious emotions, first of all, shame.

LOC3 is called by Zelazo as *reflective consciousness 1*, usually this level characterizes the consciousness of three-year-olds. The manifestation of this level is the ability of children to systematically use a pair of arbitrary rules (for instance, the object of big size and of small size) for sorting the pictures representing these objects. However, the executive function of three-year-olds is still limited, it was shown by the experiments with Dimensional Change Card Sort. For being successful in this game, children must integrate two incomparable pairs of rules into a single structure. This ability characterizes the LOC4, called by Zelazo [32] as *reflective consciousness 2*. Usually, LOC4 emerges by the end of the forth year, this level is also characterized by a spectrum of meta-cognitive skills.

6 Expansion of the levels of consciousness basic model in cognitonics

It seems that the broadly felt necessity of promoting children's emotional and social competence in schools and the lack in the scientific literature of rather simple solutions to this problem are the grounds for putting forward the following conjecture: the levels of consciousness model proposed by Zelazo [32] indicates only some basic stages of consciousness development. The goal of creating appropriate theoretical foundations of promoting children's emotional and social competence will lead to discovering additional, higher stages of the child's consciousness development corresponding to mature emotional and social competence of the child.

Realizing this idea, let's give a new interpretation of the methods of developing conscious control of thought, action, and emotion described in [9, 10, 14] and belonging to the System of the Methods of Emotional-Imaginative Teaching. We'll suppose that these methods underpin the transition from the level of consciousness 4 (LOC 4) to LOC 5, from LOC 5 to LOC 6, and from LOC 6 to LOC 7. The new levels LOC 5, LOC 6, and LOC 7 will be respectively called *the level of broad beauty appreciation, the level of appreciating the value of thought, and the level of enhanced awareness of social agreements and social responsibility* [11, 12].

A very short, preliminary description of these levels is as follows. Reaching LOC 5 by the person means that this person possesses a well-developed feeling of beauty in various manifestations: the beauty of a thing, of an idea, of an expression, of a picture or sculpture, of the interpersonal relationships, etc. [10, 14].

The successful transition from LOC 5 to LOC 6 means that (a) a child is aware of the fact that his/her ideas may be socially significant, i.e. the child may be appraised by the friends or adults for the originality and beauty of his/her idea; (b) a child appreciates the value of the thoughts of other persons [8, 13]. Reaching LOC 7 by a person means that this person is sufficiently mature in the social sense, i.e. possesses an enhanced awareness of social agreements and social responsibility [9, 10].

It should be underlined that modern preschool and school educational systems in various countries encourage only a rather small proportion of children to reach the 5th - 7th levels of conscious control. But to considerably increase this proportion is vitally important for successful socialization of children in information society. Happily, at least one broadly applicable way of solving this problem has been available since the 1990s, it is given by the EIT-system.

7 Two kinds of values imply different methods of teaching

7.1 Two kinds of values

The human being is brought up in the own culture and imbibes the spirit of the culture he/she is brought up. On the level of the every-day communication and acting, the culture is revealed in the answers to the following questions: *what you value, what you believe, and how you act.*

It is well known: "For where your treasure is there will your heart be also". It means that main values influence greatly the way a person perceives and processes the information, acquires knowledge, because the values emotionally colour every cognitive process.

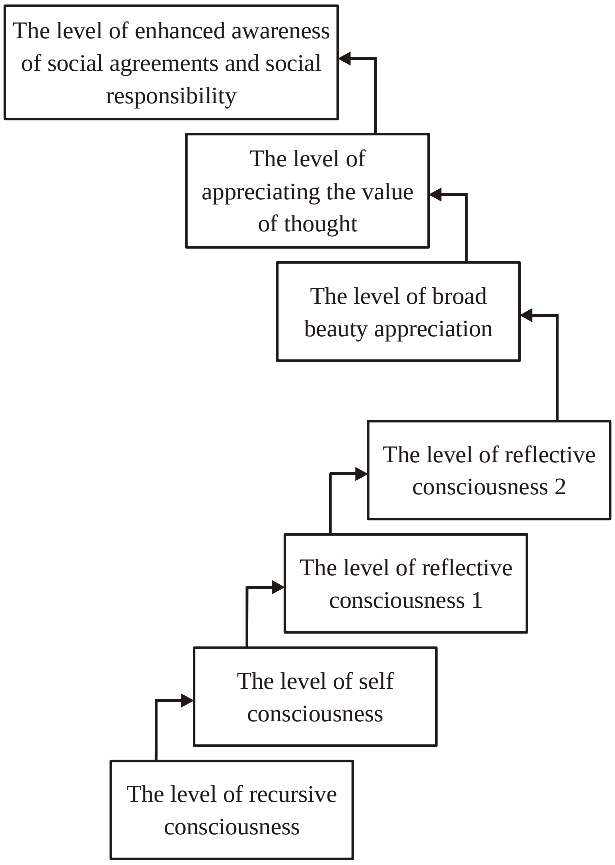

The level of enhanced awareness of social agreements and social responsibility
The level of appreciating the value of thought
The level of broad beauty appreciation
The level of reflective consciousness 2
The level of reflective consciousness 1
The level of self consciousness
The level of recursive consciousness

Figure 2: The updated configuration of the levels of consciousness model. The levels 1 – 4: P.D. Zelazo, 2004; The levels 5 – 7: V. Fomichov and O. Fomichova, 2013.

A cognitive process includes analysis, estimation, forecast, decision making, and it is underpinned by a system of values. An educational process under the frame of Cognitonics takes into account the values of students in order to create an inspiring and creative atmosphere at the lessons. If the students share lofty ideas and sublime values, have aspiration to think and act in terms of public good and benefit to the society, then it is advisable to show, for example, the beauty of mathematical solutions and equations, the beauty and value of a thought, a metaphor, to show how one and the same idea is expressed by the language of painting ("Twilight. Moon" by I. Levitan) and natural language (the moment when Alice is dozing off in the book by Lewis Carroll "Alice in Wonderland").

If the students seek for pleasure and share the commercialized values, then their motivation is different: they take a decision here and right now without awareness of their responsibility for next generations and without gratitude to previous generations. It means that they don't consider themselves as a link between generations.

In this case it is advisable to be logical, give clear solutions to the equations, do not give the so called "additional information", do not quote poetry. E.g., while explaining mathematics, try to avoid establishing the links between various languages and natural language.

The atmosphere of a lesson and the way of presenting information will meet the expectations of the audience, and the process of information processing will be successful and arise curiosity [15].

7.2 The main parameters of the values assessment

The process of assessment is very delicate and can't be called a precise one. The main question the students have to answer to let teachers guess the direction of their way of thinking is as follows: whether it is my cup of tea. If Yes then whether it is good for me; if Yes then it evokes emotions and becomes thought and interest provoking. In case with the young, 6-8-year old children it is helpful to listen to their answers and considerations, paying special attention to the way they put the ideas, answering the following questions:

(a) where did you spend your summer holidays; (b) what is your favourite dish cooked by your Mam or Great Mam for you; (c) what do you do when it is raining outside; (d) do you remember the gift Santa Claus presented you with last Christmas? (e) Do you have free time; (f) what is your favourite book: (g) can you give an example of your brightest impression; (h) what is beauty for you? (i) when do you feel yourself happy; (j) what you like to draw?

The given answers, the way they are considering, the language they use reveal the atmosphere in which they brought up, the way they view the world around, the point of their interests, the things they are impressed by (remember the song “My favourite things” from the film “Sounds of Music”).

While analysing the answers to questions, it is important to pay attention to the following: (a) whether they like dishes cooked by the mother or take away dishes? (b) if they spend summer in one and the same place, whether they are impressed by something? (c) whether children notice the change in the weather, whether they see only dirt (for example, in early spring) or notice dripping roofs, soaked roads, bluish-grey snow, and lots of “mirroirs” scattered everywhere by the spring to make the trees prepare for the spring blooming? (d) what kind of life situations do they appreciate, what makes them think, laugh, cry, feel compassion; (e) what impressed them and what makes them excited and expired; (f) what makes them happy? [15].

8 How to split children into groups and let them shift from one group to another

Let us start with an example. We have received two descriptions of the late autumn. The first one: “It is the time when the weather is getting colder, the day – shorter, the night – darker and longer, but there is no snow”. The second one: “It is the time when the water is getting tired, and it means that the snow is near. “What is up?” – “The snow is up or perhaps down”.

The first child enumerates the signs which help him to understand that the winter will come soon. He acts as a observer, as a researcher, discovering the changes and establishing the links between a cause and a consequence.

The second child reveals a poetic way of observing the nature, he uses the metaphor “tired water” in case he knows nothing about metaphors. It is just his way of viewing the world and establishing another kind of links, endowing everything with feelings.

The way children perceive the world influences the type of material presentation: so called poetical or scientific. In both cases the curiosity is aroused, information processing ability and sound creativity are improved. Both cases aim at paying a special attention to improving the language skills.

It is possible for children to shift from one group to another if the changes in the world perception are revealed.

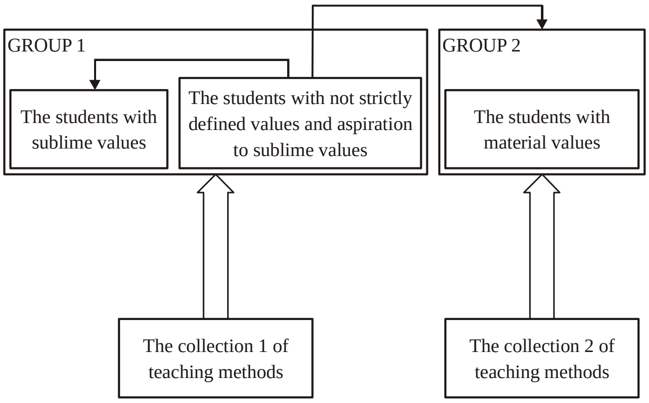

GROUP 1
The students with sublime values
The students with not strictly defined values and aspiration to sublime values
GROUP 2
The students with material values
The collection 1 of teaching methods
The collection 2 of teaching methods

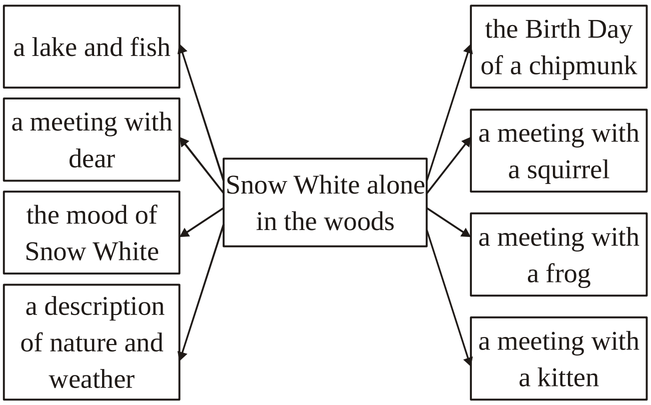
Figure 3: The impact of value system on teaching methods and on a shift (concerning a part of students) from one group to another group.

9 Methods of assessing educational progress

The swiftness of establishing the conceptual links between different thematic domains reflects the maturity of a cognitive mechanism. The process of studying and socialization aims, in particular, at constructing a great number of thematic subspaces in the world's conceptual picture of the child [15].

If the conceptual links are not activated while discussing various books, stories, while analysing information, taking a decision, then the child can't use his/her background knowledge. As a result, the processes of information processing, of taking a decision, of socialization become more complicated and very often mislead the child.

The examples of constructing conceptual links between different thematic domains at the lesson during a 20-minutes active creative work are given on Figure 4 and Figure 5.


Snow White alone in the woods
a lake and fish
a meeting with dear
the mood of Snow White
a description of nature and weather
the Birth Day of a chipmunk
a meeting with a squirrel
a meeting with a frog
a meeting with a kitten

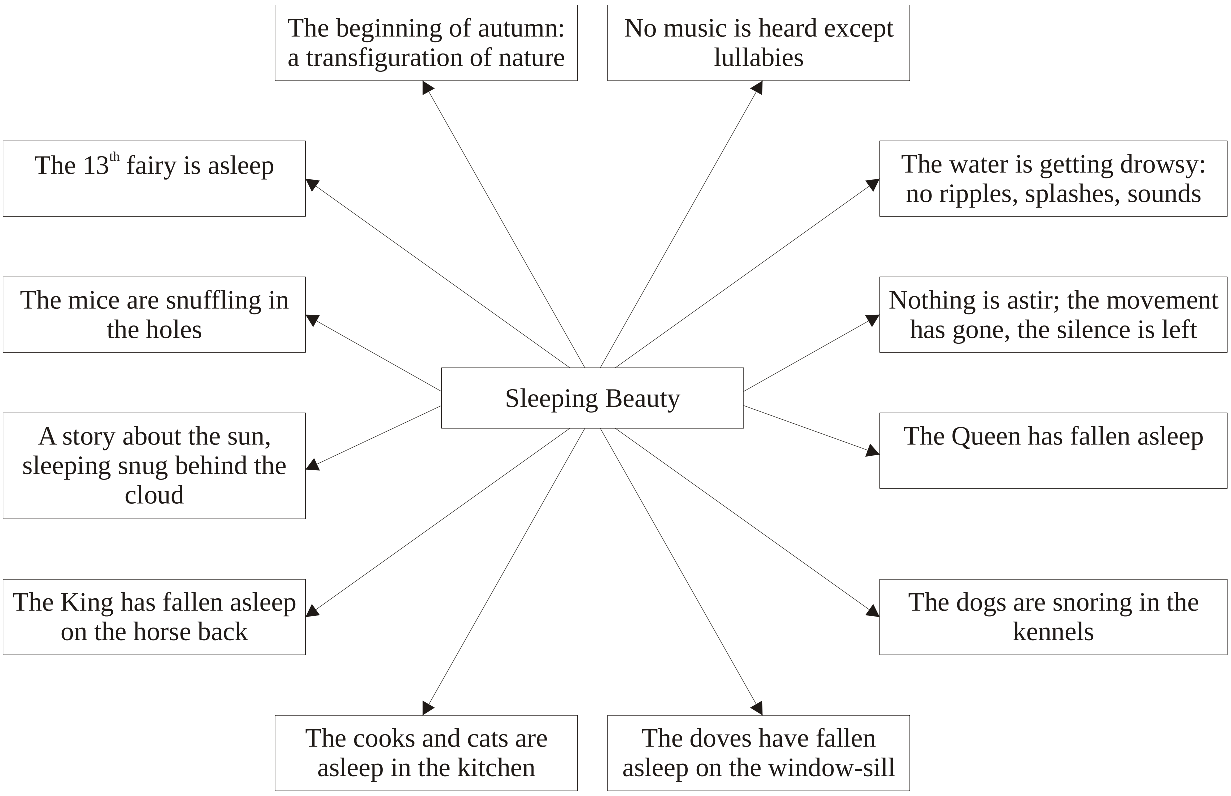
V = 8/20 min = 0.40

Figure 4: The speed of forming conceptual links during one lesson (the first year of studies, the age of children is 6 years).

After the fifth year of studies in experimental groups, the speed of forming conceptual links during one lesson exceeds the value 1. It means that during 20 minutes a group consisting of 21 - 23 students generates over 20 links between different thematic domains.

In order to better understand the difference between computer-dependent and computer-independent thinking, we’ll consider the essence of creative thinking with the help of a scheme of constructing creative cognitive pinnacles.

Creative thinking suggests the ability of the student to create a new reality or transfigure the existing one. Computer dependent thinking means following the logic of the computer. In case of establishing the conceptual links between various application domains, the qualitative characteristic is defined by the quantity of the application domains linked together, on the one hand, and the remoteness of these application domains from one another (that is, the lack of the evident ties between the domains), on the other hand. A study of metaphoric thinking was carried out according to the logic described



The beginning of autumn: a transfiguration of nature

No music is heard except lullabies

The 13th fairy is asleep

The water is getting drowsy: no ripples, splashes, sounds

The mice are snuffling in the holes

Nothing is astir; the movement has gone, the silence is left

A story about the sun, sleeping snug behind the cloud

The Queen has fallen asleep

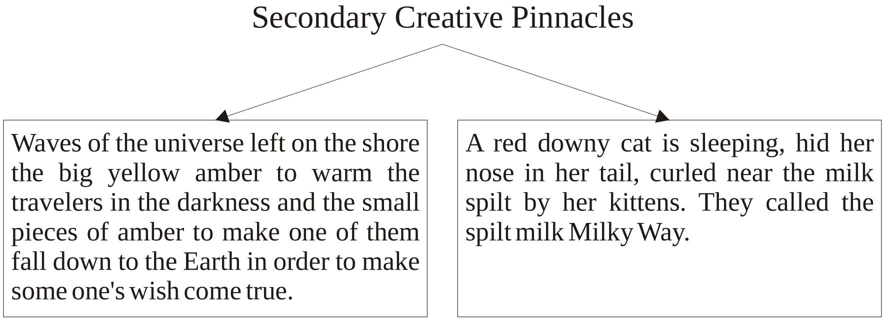
The King has fallen asleep on the horse back

The dogs are snoring in the kennels

The cooks and cats are asleep in the kitchen

The doves have fallen asleep on the window-sill

Figure 5: The speed of forming conceptual links during one lesson of the second year of studies; the age of children is 7 years; V = 12/20 minutes = 0.60.



Waves of the universe left on the shore the big yellow amber to warm the travelers in the darkness and the small pieces of amber to make one of them fall down to the Earth in order to make some one's wish come true.

A red downy cat is sleeping, hid her nose in her tail, curled near the milk spilt by her kittens. They called the spilt milk Milky Way.

Figure 6: The examples of secondary creative pinnacles

below.

Step 1. Taking into account the initial metaphor and the number of the metaphors created in accordance with the initial model, the students reach *the first creative pinnacle*. It corresponds to a new metaphor being very different from the initial one. It is a result of the unexpected coincidence of the phenomena from two application domains.

Example. Suppose that the initial metaphor is "The moon is a piece of cheese for the mice". Following this model, the young students generate a lot of metaphors, for instance, "The moon is a big round ice cream", "The moon is a pancake with a sour cream", "The moon is a piece of melon". Then one student reaches the following *first creative pinnacle*: "The moon is the silvery ball under the circus cupola. In the circus everyone is awaiting for his/her turn to appear on arena lit up with the millions of the sparkling starts scattered from that silvery ball. In the morning the moon will disappear, the stars will fade, and everyone will go for a work. The miracle happens only night".

Step 2. The *secondary creative pinnacles* designate the appearance of a principally new metaphor based on the independent creative pinnacles (see Figure 6). The initial metaphor usually is a response to the request of a teacher. Then the process of creating metaphors goes on until a principally new conceptual metaphor is created (a creative pinnacle). For the researchers, the creation of secondary creative pinnacles is much more interesting. The existence of the tendency of the emergence of the secondary creative pinnacles and the development of the process of the creation reveal the speed and the quality of the development of the cognitive mechanisms. The maturity of the cognitive mechanisms is revealed in the ability of using metaknowledge.

Unfortunately, computer dependent thinking reveals only the initial metaphor suggested by the computer and the process of creating metaphors according to a model. But it doesn't reveal creative pinnacles of any other levels, because of the lack of a vivid, lively, inspired atmosphere of discussion without computer support. Computer dependency blocks the ability of creating a new reality as a result of considering this activity as an excessive activity.

The digital reality makes the computer and ICT overwhelming in numerous spheres of human's activity. It creates the illusion of a new step on the way of civilization. But the development of the civilization without spiritual development is the greatest distortion that diminishes the creative ability of the mind or transfers it into another form, a form of adjusting but not a kind of breakthrough.

There should be two clear, well-balanced main subjects of the educational process of any level: (a) computer literacy, because ICT can directly contribute to human capabilities; computers and the Internet have a crucial influence on individual economic achievements and carrier development in the information society; (b) the development of the cognitive mechanisms of information processing and the improvement of the ability of metaphoric thinking, it leads to improving the serendipity.

10 How to achieve cognitive engagement of the students

Cognitive engagement can be defined as the process of highly motivated intellectual activity when the interest towards the subject under discussion is so strong that the students loose the track of time and, as a result, they are not tired. The students' interest determines the level of involvement. The emotional response is very close to inspiration, because they are making their own discoveries, and their mental efforts are appreciated. It helps to provide a conceptual learning environment instead of a memorization based one and enhances the motivation [15].

Cognitive engagement is characterized by the following things:

- *focused attention*; it means that within the first five minutes of a lesson the students have come to the conclusion: it is my cup of tea;

- *positive effect* (how do you feel about it); it means that the second conclusion is as follows: "it is good for me";

- *aesthetics*; it means that the way the material is presented meets the expectations of the students, it can be compared with various communicative styles: while communicating, it is better to stick to one style; in this case, it won't disappoint the partner of communication and make the conversation an easy and pleasant business; if the values of the students are clear and they are split into the groups according to their values, then it is easier to arrange the presentation either in a more pragmatic or a more poetical way (metaphorical way);

- *endurability*; it means that a student remembers a good experience and wants to repeat it;

- *novelty*; it is present at every lesson and provides intellectual and spiritual nourishment;

- *reputation, trust, and expectation*; the reputation of a teacher (his/her personal reputation and the professional one) suggests the situation when the students trust the teacher, appreciate his/her time and knowledge and act as the colleagues in the process of co-creation, still being aware of the distance between the teacher and the students, they respect this distance due to reputation of the teacher; in this case, the actions of both sides of the educational process meet the expectations of each other;

- *motivation*; the motivation of the students is closely connected with their values; the human being can be called a biological anticipatory system; everyone answers the questions: "What is good for me and how to achieve the state of complete happiness?"; but everyone defines happiness in his/her own way according to his/her understanding of values; some students are happy if they receive excellent marks; others need not only excellent marks but the awareness of intellectual and spiritual maturity, broad outlook (unconsciously, they are searching for their calling); and only in this case their level of happiness is changed [15].

To achieve cognitive engagement is very important. On the one hand, it is a marvel, because the teacher and the students become colleagues in the process of co-creation and making decision and keep the distance between the students and the teacher which is underpinned by trust, respect, and appreciation. On the other hand, it is a well managed process of knowledge acquisition. This process is underpinned by the described above mechanism of starting up the creative process in the heads of the students and creating at a lesson a special, thought-provoking atmosphere providing an opportunity for the most effective knowledge acquisition and information processing.

We have discovered the conditions under which this mechanism works well. The main condition is splitting students into different groups according to their values. The values are taken into account for creating an inspiring atmosphere, it is the most comfortable for knowledge acquisition. The students step by step receive serendipitous information: it is not expected but desirable and conduces to making their own discoveries.

11 Related approaches

One of the central ideas of our approach to early socialization of children by means of introducing them, in an original way, to the humanities is to teach young children and adolescents to appreciate beauty in all its manifestations, in particular, the beauty of nature, painting, poetry, music, classical dance. The first additional level of consciousness (LOC) development introduced by us in comparison with the LOC model by P.D. Zelazo [32] is called the level of broad beauty appreciation (see Figure 2 in Section 6).

This idea and very positive educational results obtained during 24-years-long study excellently correlate with the conclusions of a three-year study carried out by cognitive neuroscientists from seven leading universities in USA [17]. The latter study was led by Dr. M.S. Gazzaniga from the University of California at Santa Barbara. This study included, in particular, the following conclusions [4, 17]:

An interest in a performing art contributes to the development of the sustained attention, and it is necessary both for improving performance and for the training of attention as a precondition of the improvement in other conceptual domains.

There are special links (extending far beyond the music) between high levels of music training and the ability of processing information in both working and long-term memory. In particular, as concerns children, the success at the lessons of music develops the skills of solving geometrical problems.

There are positive correlations between the regular lessons of music and both reading acquisition and learning of sequences.

The training of complex actions improves the memory due to the learning of general skills for manipulating semantic information.

The process of learning to dance by means of attentive and effective observation is close to the process of learning with the help of physical practice. The effective observational learning may contribute to the development of other cognitive skills.

The educational program realizing the ideas of our informational-aesthetic conception of developing cognitive-emotional sphere of the learners includes, in particular, every-semester performance in each group. The significant components of each performance are classical dances, musical pieces, and songs. The observations accumulated during 24 years confirm the listed conclusions of the three-year long study described in [17].

However, some our theoretical ideas and obtained results considerably expand the conclusions formulated in [17]. First of all, it applies to the following discovery in cognitive biology, cognitive psychology, and cognitive linguistics done in the end of the 1990s: the consciousness of normal, average child at the age of five – six physiologically needs a rich language (much richer than it is broadly accepted to believe) for expressing the impressions from the beauty of nature (see [6, 7] and Section 4 of this paper).

The next significant result is an original method of supporting and developing metaphoric thinking of the child (at lessons devoted to foreign language and to studying the symbolic languages of poetry and painting) as a basic tool for supporting and developing creativity of the child, for the realization of the child's Thought-Producing Self [8, 13, 14].

The method of reaching LOC7 (the level of enhanced awareness of social agreements and social responsibility) proceeds from the central idea of J.R. Searle [26] about natural language as the primary means of constructing social reality and considerably expands and works out in detail this idea, inscribes it into educational practice.

12 Conclusions

This paper grounds the necessity of much earlier socialization of children in modern information society than it is usually done throughout the world. The paper sets forth the deep connections of cognitonics with the positive psychology movement. It is shown that cognitonics suggests a system of original, mindfulness-based educational methods supporting well-balanced cognitive-emotional development of the personality in modern information society, it is called the system of the methods of emotional-imaginative teaching (the EIT-system). The analysis of central ideas of the EIT-system provided the possibility to enrich developmental psychology: the basic model proposed by P.D. Zelazo (2004) considers 4 levels of consciousness development (corresponding to the age from one to four years), and this paper introduces three new levels (they cover the ages from five – six to 13 – 14 years).

The paper presents a new look at the process of education when the values of the student act like a lighthouse for the teacher at the moment of presenting material and arranging the process of education, the process of acquiring knowledge. Four discoveries underpinning the proposed way of solving this problem are shortly described. This way is provided by the EIT-system belonging to the constructive core of Cognitonics. The described methods have been successful tested in the course of a longitude study covering 24 years of introducing young children and adolescents to the humanities.

The EIT-system has been mainly realized at lessons of English as a foreign language for Russian-speaking children and at the lessons of poetry and literature in English, at lessons devoted to explaining the symbolic language of painting, the culture of communication, and the symbolic language of classical dance. These kinds of lessons are considered in numerous countries as highly appropriate for young children and teenagers. The carefully selected collection of texts used at lessons is provided by a number of classical, world-known fairy-tales and novels, in particular, “Snow White”, “Cinderella”, “Sleeping Beauty”, “Pinocchio”, “Pollyanna”, “The Life and Adventures of Santa Claus“ by L. Frank Baum, “Alice in Wonder Land” by Lewis Carroll, “The Wind in the Willows” by Kenneth Grahame, “The Hundred and One Dalmatians” by Dodie Smith, etc. That is why the EIT-system may be used (after a certain adaptation requiring a small time) in English-speaking countries and in numerous countries where the English language is learned as a second language.

Acknowledgement

We are grateful to our friend, Senior Software Researcher Alexander Artyomov for his generous help concerning the preparation of the electronic versions of the figures in this paper.

References

[1] Bohanec, M., Gams, M., Rajkovič, V. et al., Eds. (2009). Proceedings of the 12th International Multiconference Information Society – IS 2009, Vol. A, Slovenia, Ljubljana, 12 – 16 October 2009. The Conference Kognitonika/Cognitonics. Jozef Stefan Institute; http://is.ijs.si/is/is2009/zborniki.asp?lang=eng; pp. 427-470; retrieved 15.12.2013

[2] Bohanec, M., Gams, M., Mladenić, D. et al, Eds. (2011). Proceedings of the 14th International Multiconference Information Society – IS 2011, Vol. A, Slovenia, Ljubljana, 10 – 14 October 2011. The Conference Kognitonika/Cognitonics. Jozef Stefan Institute; , http://is.ijs.si/is/is2011/zborniki.asp?lang=eng; pp. 347-430; retrieved 15.12.2013

[3] Centre for the Study of Existential Risk (2013); http://cser.org; retrieved 24.09.2013.

[4] Christofidou, A. (2013). Remembrance of things past, a research hypothesis. In Proceedings of the 16th International Multiconference Information Society – IS 2013, Slovenia, Ljubljana, 7 – 11 October 2013. The Conference Kognitonika/Cognitonics. Jozef Stefan Institute; pp. 409-412; http://is.ijs.si/is/is2013/zborniki.asp?lang=eng; retrieved 15.12.2013

[5] Fomichov, V.A., Fomichova, O.S. (1994). The Theory of Dynamic Conceptual Mappings and its Significance for Education, Cognitive Science, and Artificial Intelligence. Informatica. An International Journal of Computing and Informatics (Slovenia), 1994, Vol. 8, No. 2, pp. 31-148.

[6] Fomichov, V.A., Fomichova, O.S. (1997). An Informational Conception of Developing the Consciousness of the Child. Informatica. An International Journal of Computing and Informatics (Slovenia), Vol. 21, pp. 371-390.

[7] Fomichov, V.A., Fomichova, O.S. (2001). A Many-Staged, Humanities-Based Method of Realizing the Thought-Producing Self of the Child. Consciousness, Literature and the Arts, 2001, Vol. 2, No. 1; http://blackboard.lincoln.ac.uk/bbcswebdav/users/dmeyerdinkgrafe/archive/fomichov.htm; retrieved 09.12.2013

[8] Fomichov, V.A., Fomichova, O.S. (2006). Cognitonics as a New Science and Its Significance for Informatics and Information Society; Informatica. An International Journal of Computing and Informatics (Slovenia), Vol. 30, pp. 387-398.

[9] Fomichov, V.A., Fomichova, O.S. (2011). A Map of Cognitive Transformations Realized for Early Socialization of Children in the Internet Age, in M. Bohanec et al (eds.). Proceedings of the 14th International Multiconference Information Society – IS 2011, Ljubljana, pp. 353-357; http://is.ijs.si/is/is2011/zborniki.asp?lang=eng; retrieved 14.12.2013

[10] Fomichov, V.A., Fomichova, O.S. (2012). A Contribution of Cognitonics to Secure Living in Information Society; Informatica. An International Journal of Computing and Informatics (Slovenia), Vol. 36, pp. 121-130; www.informatica.si/vol36.htm#No2; retrieved 17.11.2013

[11] Fomichov, V.A., Fomichova, O.S. (2013a). The Peculiarities of the Mindfulness-Based Development of the Personality under the Frame of Cognitonics. In George E. Lasker and Kensei Hiwaki (Eds.) Personal and Spiritual Development in the World of Cultural Diversity, Vol. X. The International Institute for Advanced Studies in Systems Research and Cybernetics (IIAS), Tecumseh, Ontario, Canada, 2013, pp. 37-41.

[12] Fomichov, V.A., Fomichova, O.S. (2013b). The Significance of Mindfullnes-Based Educational Methods Provided by Cognitonics for Positive Psychology Movement. M. Gams, R. Piltaver, D. Mladenić et al (Eds.), Proceedings of the 16th International Multiconference Information Society – IS 2013, Slovenia, Ljubljana, 7 – 11 October 2013. Vol. A. The Conference Kognitonika/Cognitonics. Jozef Stefan Institute, pp.

425-429; http://is.ijs.si/is/is2013/zborniki.asp?lang=eng; retrieved 14.12.2013

[13] Fomichova, O.S., Fomichov, V.A. (2000). Computers and the Thought-Producing Self of the Young Child; The British Journal of Educational Technology, Vol. 31, pp. 213-220.

[14] Fomichova, O.S., Fomichov, V.A. (2009). Cognitonics as an Answer to the Challenge of Time; Proceedings of the 12th International Multiconference Information Society - IS 2009, Slovenia, Ljubljana, 12 – 16 October 2009. The Conference Kognitonika/Cognitonics. Jozef Stefan Institute, 2009, pp. 431-434; available online at http://is.ijs.si/is/is2009/zborniki.asp?lang=eng; retrieved 10.12.2013

[15] Fomichova, O.S., Fomichov, V.A. The Risk of Postponing Early Socialization of Smart Young Generation in Modern Information Society. M. Gams, R. Piltaver, D. Mladenić et al (Eds.), Proceedings of the 16th International Multiconference Information Society – IS 2013, Slovenia, Ljubljana, 7 – 11 October 2013. Vol. A. The Conference Kognitonika/Cognitonics. Jozef Stefan Institute. 2013, pp. 430-434; http://is.ijs.si/is/is2013/zborniki.asp?lang=eng; retrieved 11.12.2013

[16] Gams, M., Piltaver, R., Mladenić, D. et al., Eds. (2013). Proceedings of the 16th International Multiconference Information Society – IS 2013, Slovenia, Ljubljana, 7 – 11 October 2013. The Conference Kognitonika/Cognitonics. Jozef Stefan Institute; pp. 403-482; http://is.ijs.si/is/is2013/zborniki.asp?lang=eng; retrieved 15.12.2013

[17] Gazzaniga, M.S. (2008). Learning, Arts, and the Brain: The Dana Consortium Report on Arts and Cognition. NY, Washington, DC., Dana Press; http://www.wjh.harvard.edu/~lds/pdfs/DanaSpelke.pdf; retrieved 10.12.2013

[18] Good, I.J. (1965). Speculations concerning the first ultraintelligent machine. In F.L. Alt and M. Rubinoff (Eds.), Advances in Computers, Vol. 6, Academic Press, pp. 31-88.

[19] Huebner, E.S., Gilman, R. (2003). Toward a Focus on Positive Psychology in School Psychology; School Psychology Quarterly; V. 18, pp. 99-102.

[20] Hui, S. (2012). Cambridge to study technology's risk to humans. Associated Press, November 25, 2012; http://bigstory.ap.org/article/cambridge-study-technologys-risk-humans; retrieved 15.07.2013.

[21] Kabat-Sinn, J. (2003). Mindfulness-based Interventions in the Context: Past, Present, and Future; Clinical Psychology: Science and Practice, V. 10, pp. 144-156.

[22] Kurzweil, R. (2005). The Singularity Is Near: When Humans Transcend Biology, 2005.

[23] Naughton, J. (2012). Could robots soon add to mankind's existential threats?". The Observer. 02 December 2012. Retrieved 12 March 2013; http://www.theguardian.com/technology/2012/dec/02/ai-robots-google-car-humans-john-naughton

[24] Schonert-Reichl, K.A., Lawlor, M.S. (2010). The Effects of a Mindfulness-based Education Program on Pre- and Early Adolescents' Well-being and Social and Emotional Competence; Mindfulness, Vol. 1, pp. 137-151.

[25] Screenager (2013). In Oxford Dictionaries. Language matters. English; http://www.oxforddictionaries.com/definition/american_english/screenager; retrieved 05.12.2013

[26] Searle, J.R. (1995). The Construction of Social Reality. The Penguine Press.

[27] Seligman, M.E.P., Csikszentimihalyi, M. (2000). Positive Psychology: an Introduction; American Psychologist, Vol. 55, pp. 5-14.

[28] US Public Health Service. Report on the Surgeon's General's Conference on Children's Mental Health: a national action agenda (2000). Washington, DC., Department of Health and Human Services.

[29] Vandewater, E.A., Rideout, V.J., Wartella, E.A., Huang, X., Lee, J.H., M.-S. Shim, M.-S. (2007). Digital Childhood: Electronic Media and Technology Use Among Infants, Toddlers, and Preschoolers. Pediatrics. Official Journal of the American Academy of Pediatrics, Vol. 119, No. 5, pp. 1006-1015.

[30] Vinge, V. (1993). The Coming Technological Singularity: How to Survive in the Post-Human Era; http://www-rohan.sdsu.edu/faculty/vinge/misc/singularity.html; retrieved 24.11.2013

[31] Wallis, L. (2013). Is 25 the new cut-off point for adulthood? BBC News, 25 Sept. 2013; http://www.bbc.co.uk/news/magazine-24173194; retrieved 27.09.2013.

[32] Zelazo, P.D. (2004). The Development of Conscious Control in Childhood. Trends in Cognitive Sciences, Vol. 8, pp. 12-17.

Semantic Searching of Biological Documents Using Gene Ontology

Marwa Mostafa[1], Enas M.F. El Houby[2] and Akram Salah[1]

[1]Faculty of computers and information, Cairo University, computer science department
5 Dr. Ahmed Zewail, Orman, Giza, Egypt
E-mail: m.mostafa_fci@yahoo.com, akram.salah@fci-cu.edu.eg

[2]Systems & Information Department, Engineering Division, National Research Centre, Dokki, Giza, Egypt
E-mail: em.fahmy@nrc.sci.eg

Keywords: information retrieval, biological ontology, ranking methodology, precision, recall

Received: September 21, 2013

Semantic information retrieval of biological documents is an information retrieval approach that utilizes semantics to improve the search recall and precision. This research presents a framework for a semantic biological retrieval system that effectively searches and retrieves meaningful results using Gene Ontology. The system takes two related biological terms as an input and retrieves relevant documents which contain these inputs. Since the user searches for the documents that contain two related biological terms, the system helps the user to know the hierarchical relationship between these two terms using Gene Ontology. The system utilizes the Gene Ontology to infer semantically related terms to the inputs. The inferred words may include synonyms, parents and grandparents of the input terms entered in the search query. The system uses these related inferred terms in expanding the user query to produce meaningful results since it retrieves the documents that contain the input terms and these inferred related terms. The system uses a ranking methodology to help in ordering the retrieved documents based on the rank values. The proposed technique improves the precision of the retrieved documents as well as the recall which saves researcher time and focus.

Povzetek: Razvita je metoda iskanja bioloških dokumentov z uporabo genskih ontologij.

1 Introduction

The biological repositories contain hundreds of thousands of electronic collections that often contain high quality information [1]. During the past years, the increase in scientific knowledge and the massive data production have caused an exponential growth in the number and size of biological databases and repositories. However, data size, which can reach hundreds of gigabytes, involves serious problems of data access through data storage in local disks. Other challenging issues associated to biological data are that much relevant information is spread out in different databases or repositories [2]. So the biological data is still locked in a large number of resources; remaining not computer-readable. In the current search engines when the user enters two terms it returns a lot of documents including unhelpful ones.

Keyword-based search is currently the most commonly employed search strategy in biomedical digital libraries. When users search by a few keywords, a large number of matched results could be returned. Users spend a significant amount of time to browse these results to find out those documents they are truly interested in because the publications returned may not be organized based on the user needs, forcing users to browse thousands of publications. In most cases, it is impossible for users to manually read every returned entry thus leads to loss of many truly relevant publications [3].

The goal of an information retrieval (IR) system is to rank documents optimally given a query to rate the relevance of documents. In order to achieve this goal, the system must be able to score documents so that the relevant document would ideally have a higher score than the irrelevant one [4].

Most of the current forms of web content are designed to be presented to humans; they are not understandable by computers. The semantic web aims at enhancing existing web content with semantic structure in order to make it meaningful to computers as well as to humans. Ontology plays a key role in the semantic web [5], [6] which offers an advanced approach for managing, retrieving information and processing it.

Ontology is a formal conceptualization of a particular domain into a human understandable, machine-readable format [7].One of the most important bio-ontology is Gene Ontology [8]. It organizes terms in a parent-child hierarchy.

Our first publication about this framework was "Ontology based Biological Information Retrieval System" (OBIRS) [9] which shows how we improved the efficiency of the method used in the system algorithm.

The proposed system presented in this paper uses Gene Ontology to infer semantically related terms to the input terms. The inferred terms may include synonyms which are useful in retrieving documents by authors who use different wording in reference to the same concept.

The system also infers related terms through parent-child relationship up to 2 levels (parents and grandparents) for each term of the input terms to expand the search query.

The proposed system helps the researchers to get more relevant and accurate retrieval of the documents. It allows the researchers to enter two related terms to get the documents that contain both of them. Also the system semantically retrieves the documents if they contain synonyms of the input terms inferred from the Gene Ontology even if these documents do not contain the exact phrase of the input terms. Also the system retrieves the documents that contain the input terms and/or synonyms with any combination of the other inferred terms (parents and grandparents).

The system searches for two related terms because its main idea is to retrieve documents that contain relation among related biological terms and we found that the least number of possible terms to find a relation between is two.

The system uses a ranking methodology that helps in ranking the retrieved documents to achieve the researchers satisfaction and save time and effort consumed by the researchers to rate the relevance of documents manually. The system groups the retrieved documents into five classes to save the time of the researchers. The system also extracts the relation between the input terms from the gene ontology to give the researchers the hierarchical relation between them.

The remainder of the paper is organized as follows: in section 2, an overview for the previous work related to our subject is presented. In section 3, the architecture of the proposed system is described. In section 4, ranking issues are explained. In section 5, an example is introduced to illustrate the proposed system. In section 6 relationship extraction is explained. In section 7, testing the system and the results are produced, before drawing conclusions and future work in section 8.

2 Related work

A lot of previous work was studied for the subject of semantic web and biological information retrieval. Sumithiradevi et al.[10]proposed one such tool called BIOMINING that is designed to eliminate anomalous and redundancy in biological web content. The authors use indexing and mining technology on biological databases to summarize the information of biological data in the document. Zhou et al. [11] designed a biological information retrieval and analysis system (BIRAS) based on the Internet. The system could send and receive information from the Entrez search and retrieval system maintained by the National Centre for Biotechnology Information (NCBI) in USA. Marta Bleda et al.[2]proposed the "CellBase" that provides a solution to the growing necessity of integration by easing the access to biological data. CellBase implemented a set of RESTful web services that query a centralized database containing the most relevant biological data sources. Minlie Huang et al.[12]proposed Ontology-based biological relation extraction system to automatically extract biological relations from a huge number of online MEDLINE abstracts. Authors then made Ontology-based semantic annotation of online biological documents. Anália Lourenço et al. [13] present BioDR which is a novel approach that allows the semantic indexing of the results of a query by identifying relevant terms in the documents. This system makes it possible to navigate semantically between documents and relevant terms, taking advantage of the rich contents of full-text.

Many other researchers [1], [14], [15], [16], [17], [18], [19] used ontologies, inverted list (different tech.) and query expansion to assist biological information retrieval search.

After reviewing several researches that support the retrieval of biomedical information it is our conclusion that the most similar to our system is[12]. However the previous reviewed researches aim to study the design of a biological information retrieval and analysis systems using the Internet. These systems are designed to eliminate anomalies and redundancy in biological web content, integrate biological database, retrieve biological information and extract relations. Our proposed system is a biological semantic retrieval system that tries to improve the recall and precision of the retrieved documents and helps the researchers to get the relevant documents that contain information and relationships between two related biological terms. The system retrieves documents that contain the terms as well as other semantically related terms inferred from the Gene Ontology. The system also ranks the retrieved documents based on their relevance to the input terms. The system retrieves the content of the document, not its address, unlike other retrieval systems. It is our assumption that retrieving relevant documents that contain information about two related biological terms entered from the researchers and ranking them should save the researchers time and effort.

3 Proposed system

Testing our previous system presented in [9] shows that there are many documents that satisfy the researcher's needs and have not been retrieved. Many biological terms have synonyms and it is possible to have a document that contains synonyms of the two terms entered in the search query or that contains one term and the synonym of the other term during the searching process. These documents have not been retrieved to the researchers in spite of its relevancy to the query. Also the retrieved documents have not been ranked appropriately. That was a motivation for improving the effectiveness of the previous system since there are many documents that semantically may satisfy the researchers needs and have not been retrieved.

The system (EOBIRS) presented in this paper is the Enhanced Biological Information Retrieval System which is the updated version of the previous system that highlights the importance of the synonyms of the searched terms and retrieves documents from corpus even if they have the synonyms only and doesn't contain the same wording of the terms entered in the search query because semantically they are reference to the

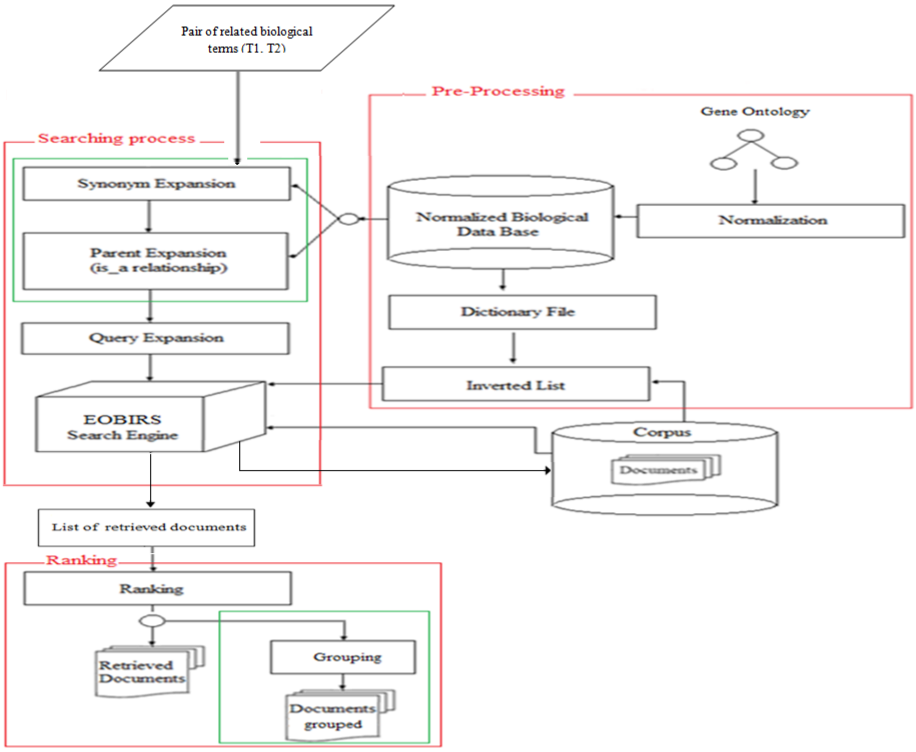


Figure 1: System architecture.

same concept. The system presented in this paper retrieves documents that contain the two entered terms, their synonyms and their parents up to two levels (parents and grandparents).

The system presented in this paper has the same system design like the previous system. It also has the same pre-processing instructions like [9] and differs in the searching process instructions, ranking criteria, grouping criteria and reordering the classes based on the balance value that adds another facility to minimize the time and effort of the researchers. The system architecture is shown in Figure 1.

3.1 Pre-processing

i. The system normalizes the Gene Ontology to a database named "DBGenes". The database "DBGenes" contains all genes that exist in the Gene Ontology with their attributes such as name, id, definition, synonymous, is_a and part_of.

ii. The system builds a dictionary file that contains all the biological terms exist in the normalized database.

iii. The system builds inverted list based on the biological terms only that exist in the corpus's documents. The system compares all terms exist in the corpus's documents with the biological terms exist in dictionary file, so a term added to the inverted list if it was found in the dictionary file. The terms were added to the inverted list with a list of the documents that contains these terms and the positions of the terms and the frequencies of their appearance in each document.

3.2 Searching process

a. The researcher enters the two related biological terms that he/she wants to search for. Where the system searches for unique identifiers for biological terms, the system begins to check if the "DBGenes" contains these terms or not. The search process starts if the two terms exist in the "DBGenes".

b. The system gets all the synonymous for both terms from the normalized database "DBGenes".

c. The system gets all the parents up to two levels (parents and grandparents) for both terms from the normalized database "DBGenes".

d. The system expands the query "term1 AND term2" using synonymous provided from the Ontology as well as parents and grandparents using "is_a"relation that describes the parent-child relationship. The query will expanded as follow:

If we assume that the two related biological terms entered to the retrieval system are G_1 and G_2. The set of synonyms are later called a synset. If the two synsets are $GS_1=\{gs_{11}, gs_{12},\ldots,gs_{1m}\}$ and$GS_2= \{gs_{21},gs_{22},\ldots,gs_{2n}\}$, and if the gene parents are $GP_1= \{gp_{11}, gp_{12},\ldots,gp_{1i}\}$ and $GP_2 =\{gp_{21},gp_{22},\ldots,gp_{2j}\}$, and if the gene grandparents are $GGP_1= \{ggp_{11}, ggp_{12},\ldots,ggp_{1k}\}$ and $GGP_2=\{ggp_{21},ggp_{22},\ldots,ggp_{2l}\}$, the query will expanded into these queries:

Q$_1$ retrieves all the documents that contain the two related biological terms and/or the synonyms and their parents and their grandparents.

$\mathbf{Q_1}=[((G_1)\text{OR}(gs_{11}\text{OR}gs_{12}\text{OR}\ldots gs_{1m}))\text{AND}(gp_{11}\text{OR}\ gp_{12}\text{OR}\ldots,gp_{1i})\ \text{AND}(ggp_{11}\text{OR}\ ggp_{12},\text{OR}\ldots,ggp_{1k})]\ \underline{\text{AND}}\ [((G_2)\ \text{OR}\ (gs_{21}\text{OR}gs_{22}\text{OR}\ldots gs_{2n}))\text{AND}(gp_{21}\text{OR}\ gp_{22}\text{OR}\ldots,gp_{2j})\text{AND}\ (ggp_{21}\text{OR}ggp_{22}\text{OR}\ldots,ggp_{2l})]$

Q$_2$ retrieves all the documents that contain the two biological terms and/or the synonyms with their parents or grandparents.

$\mathbf{Q_2}=[((G_1)\text{OR}(gs_{11}\text{OR}gs_{12}\text{OR}\ldots gs_{1m}))\text{AND}((gp_{11}\text{OR}gp_{12}\text{OR}\ldots,gp_{1i})\ \text{OR}\ (ggp_{11}\ \text{OR}ggp_{12}\ \text{OR}\ldots,ggp_{1k}))]$

AND $[((G_2)$ OR $(gs_{21}ORgs_{22}OR\ldots gs_{2n}))AND((gp_{21}OR\ gp_{22}OR\ldots gp_{2j})$ OR $(ggp_{21}$ OR ggp_{22} OR $\ldots,ggp_{2l}))]$

Q_3 retrieves all the documents that contain the two terms or their synonyms or one term and the synonym of other term.

Q_3= $[(G_1)$ OR $(gs_{11}ORgs_{12}OR\ldots gs_{1m})]$ AND $[(G_2)$ OR $(gs_{21}ORgs_{22}OR\ldots gs_{2n})]$

The expanded query will be:

$Q = Q_1$ OR Q_2 OR Q_3

e. The system uses the inverted list to get the list of the documents that satisfy the query Q. This list will contain the document's names that contain the two terms (G1 and G2) and/or any combination of the related terms inferred from the Gene Ontology.

f. The system calculates the rank value of each document which used to order the retrieved documents. The system ranks the documents under a certain criteria:

- The initial value of ranking of the document is the count of occurrence of the two terms multiplied by weight W_1.
- Finding synonyms of any of the two terms increases the value of ranking by adding W_2 of the number of their occurrence.
- Finding a parent or grandparent of any of the two terms increases the value of ranking by adding W_3 of the number of their occurrence.

The rank value will be calculated as follow:

Rank value =

$$[(F(T_1)+F(T_2))*W_1]+[(F(ST_1)+F(ST_2))*W_2]+ [(F(PT_1)+F(PT_2)+ F(GPT_1)+F(GPT_2)) * W_3] \quad (1)$$

Where T_1 and T_2 are the input terms ST_1, ST_2 are the synonyms of the input terms, PT_1, PT_2 are the parents of the input terms and GPT_1 and GPT_2 are the grandparents of the input terms. F is to count the number of occurrence of the terms.

Supposed that: $W_1 > W_2 > W_3$

We supposed that W_1 to be greater than W_2 because we assume that the weight of input terms must be greater than that of synonyms this is due to that we should give a strong concern to the input terms entered by the researcher. We think that the researcher is more concern about the retrieved documents that contain the exact wording of the input terms than the documents that contain the synonyms of the input terms. We choose W_2 to be greater than W_3 because the existence of the synonyms means the existence of the input terms so we assume that the weight of finding a parent or grandparent must be less than the weight of finding a synonym. The existence of a parent or grandparent adds another prove that the retrieved document talks about the input terms but in the same time it still doesn't represent the same meaning of the input terms so we cannot give it a weight equal to the synonyms.

g. The system retrieves from corpus the documents resulted from the query Q ranked by the system ranking values.

h. The system calculates the value of the precision and recall of the retrieved documents.

$$\text{Precision} = \frac{\text{Number of relevant documents}}{\text{Number of retrieved documents}} \quad (2)$$

$$\text{Recall} = \frac{\text{Number of relevant documents retrieved}}{\text{Number of relevant documents}} \quad (3)$$

i. The system extracts the relation between the two related terms from the Gene Ontology and presents it to the user as additional information about hierarchical relation of the two terms in addition to that mentioned in the documents.

j. The user can open any of the retrieved documents and notice the two terms that he/she searches for are highlighted.

4 Ranking Issues

During testing the system issue has been released "what about if the user wants to get specific documents as the first outcome in the list of the retrieved documents for example the documents that contain terms and their parents only". Because of this issue we have added a grouping option that the system provides to the user in addition to a list of the whole documents. The list is grouped into the following classes:

Class one: Provides all the documents that each one contains the two related biological terms and/or their synonyms and their parents and grandparents.

Class two: Provides all the documents that each one contains the two related biological terms and/or their synonyms and their parents.

Class three: Provides all the documents that each one contains the two related biological terms and/or their synonyms and their grandparents.

Class four: Provides all the documents that each one contains the two related biological terms only.

Class five: Provides all the documents that each one contains the synonyms of the two related biological terms only or one term and the synonym of other term.

Each class can be ordered based on the frequencies of the two related biological terms under search with concern of the balancing between the frequency of term 1 and the frequency of term 2 to ensure that the documents contain material that tackle the relation between the two terms. The system calculates the absolute value of the difference between the frequency of term1 and the frequency of the term2 in the document. It orders the documents based on the balance value, the document is ordered first if it has low balance value. If there are two or more documents having the same balance value then they will be ranked based on the summation of the "term frequency" value of term 1 and the "term frequency"

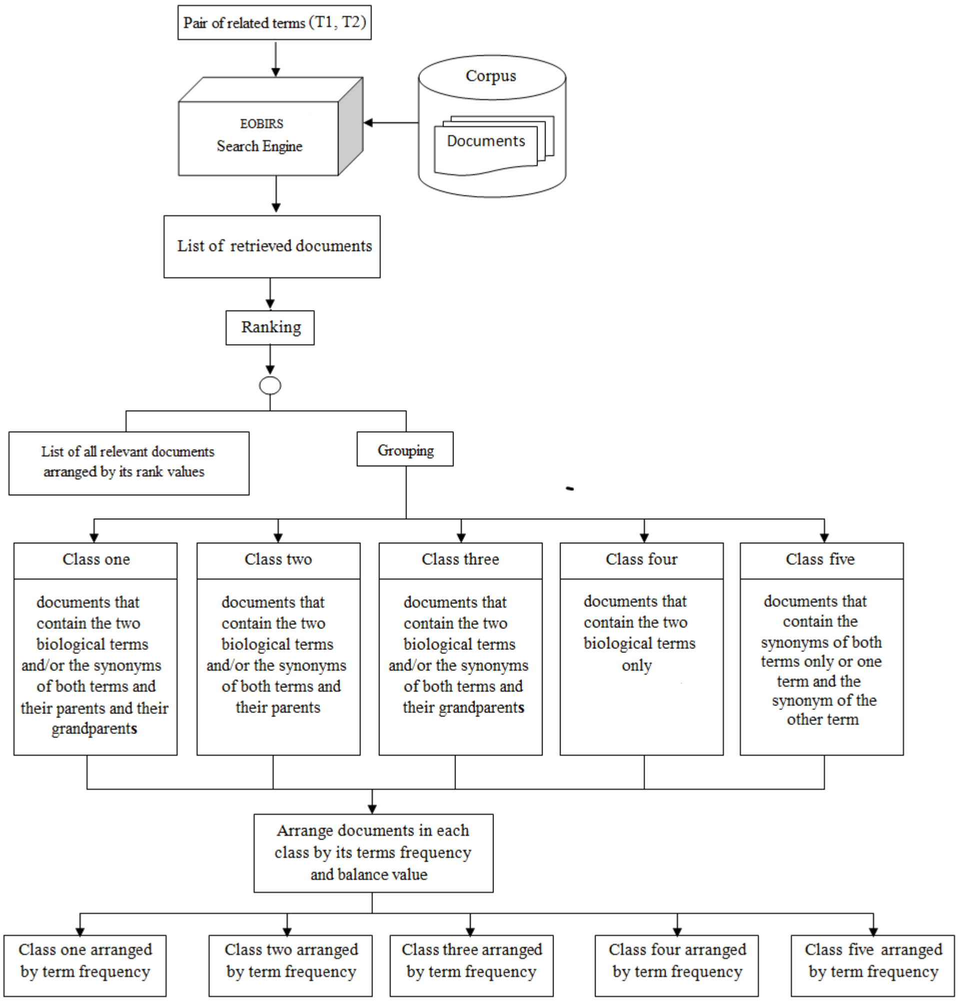

Pair of related terms (T1, T2)
Corpus
Documents
EOBIRS
Search Engine
List of retrieved documents
Ranking
List of all relevant documents arranged by its rank values
Grouping
Class one
documents that contain the two biological terms and/or the synonyms of both terms and their parents and their grandparents
Class two
documents that contain the two biological terms and/or the synonyms of both terms and their parents
Class three
documents that contain the two biological terms and/or the synonyms of both terms and their grandparents
Class four
documents that contain the two biological terms only
Class five
documents that contain the synonyms of both terms only or one term and the synonym of the other term
Arrange documents in each class by its terms frequency and balance value
Class one arranged by term frequency
Class two arranged by term frequency
Class three arranged by term frequency
Class four arranged by term frequency
Class five arranged by term frequency

Figure 2: Ranking workflow.

value of term2. The ranking workflow is shown in Figure 2.

For illustration:
If we have a list of relevant retrieved documents from class four that contains Term1 and Term 2 as shown in Table 1.

Table 1: "term frequency" values of the two terms and the calculation of the balance value for each document.

Document number	Term 1	Term 2	Balance value (absolute value of the difference)
D1	2	2	0
D2	4	4	0
D3	2	1	1
D4	6	5	1
D5	4	1	3
D6	5	19	14

The system will calculate the balance for this class as shown in Table1. The list of the documents will be ordered based on balance value as shown in Figure 3.

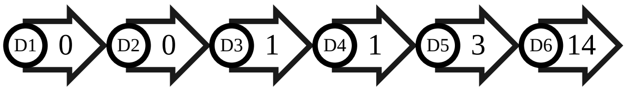


Figure 3: List of documents ordered based on balance values.

Then the system reorders the documents that have the same balance values based on the summation of the "term frequency" values of both term1 and term2.

So for D1 and D2:

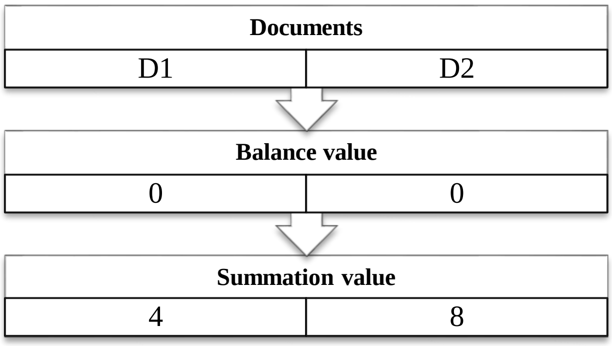


Figure 4: The comparison between term frequency values of D1 and D2.

Based on the calculation in Figure 4 the system will rank D2 higher than D1 because the summation of the "term frequency" values of term 1 and term2 in D2 is greater than their summation in D1.

For D3 and D4:

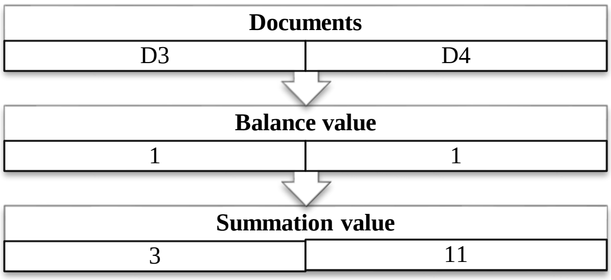


Figure 5: Comparing term frequency values of D3 and D4.

Based on the calculation in Figure5 the system will rank D4 higher than D3 because the summation of the "term frequency" values of term1 and term2 in D4 is greater than their summation in D3.

The system will present the documents for the user as shown in Figure 6.

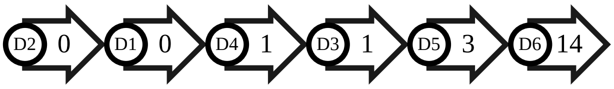


Figure 6: List of documents ordered by term's frequency values.

5 An example

To show how the system searches and orders the documents we present the following example, supposing that W_1=1, W_2= 0.8, W_3=0.25.

If the researcher searches for two terms, Term1:"regulation of DNA recombination"and Term 2:"mitochondrion inheritance"and the corpus contains a list of the following documents:

D1: mitochondrion inheritance and regulation of DNA recombination are biological_process in the Gene Ontology.

D2: mitochondrion inheritance has synonyms and regulation of DNA recombination doesn't have synonyms. Mitochondrial inheritance is a synonym of mitochondrion inheritance and organelle inheritance is a parent for it.

D3:mitochondrion inheritance and regulation of DNA recombination have parents. Regulation of DNA metabolic process is a parent of regulation of DNA recombination. Mitochondrial inheritance is a synonym of mitochondrion inheritance and organelle organization is a grandparent for it.

D4: regulation of DNA recombination is a biological process. Mitochondrion inheritance and regulation of DNA recombination have parents. Organelle organization is a grandparent of mitochondrion inheritance.

D5: Gene Ontology contains regulation of DNA recombination and mitochondrion inheritance. Regulation of DNA recombination is a biological_process. Regulation of DNA recombination is any process that modulates the frequency, rate or extent of DNA recombination. Regulation of DNA recombination is a subset of gosubset_prok. Regulation of DNA recombination has only one parent. Recombination regulates has a relationship with regulation of DNA recombination. Regulation of DNA recombination has "intersection_of" relation with biological regulation and DNA recombination regulates. We can find regulation of DNA recombination in Gene Ontology version1.2.The id of regulation of DNA recombination is GO:0000018 in the Gene Ontology. Mitochondrion inheritance is a biological_process. Mitochondrion inheritance is the distribution of mitochondria, including the mitochondrial genome, into daughter cells after mitosis or meiosis, mediated by interactions between mitochondria and the cytoskeleton.

D6: this document talks about regulation of DNA recombination and mitochondrion inheritance. Regulation of DNA recombination is any process that modulates the frequency, rate or extent of DNA recombination.

D7: Gene Ontology contains genes.

D8: mitochondrial inheritance is a biological_process in the Gene Ontology.

D9: Organelle inheritance is a biological_process in the Gene Ontology.

Table2: The calculations of documents rank values for the presented example.

Document number	Term frequency			Total	Parent	Grandparent	Document Rank value	Balance value (absolute value of the difference)
	Term1	Term2	Synonymous					
D1	1	1	0	2	-	-	2	0
D2	1	2	1	3.8	1	-	4.05	1
D3	2	2	1	4.8	1	1	5.3	0
D4	2	2	0	4	-	1	4.25	0
D5	9	3	0	12	-	-	12	6
D6	2	1	0	3	-	-	3	1
D7	0	0	0	0	-	-	0	0
D8	0	0	1	0.8	-	-	0	0
D9	0	0	0	0	1	-	0	0

Table2 presents the rank value of each document in the corpus. As shown document number 7 does not contain any of the two terms or their synonyms so it has a rank value equal to 0. Also the document number 8 has a value equal to 0 because it contains the synonymous of one term only. A document that contains synonyms of both terms will be ordered based on the total number of synonyms found in it. Also the table shows that document number 9 has a rank value equal to 0 because it contains parents only and does not contain any of the two terms or their synonyms so it will not be retrieved for the user.

The system calculates the balance values to order the documents within the class. In Figure 7 we show how the system retrieves the relevant documents based on our example.

6 Relation Extraction

Since our system objective is to retrieve the documents that relate two biological terms the system extracts the hierarchical relationship between the two terms from the Gene Ontology as additional information for the researcher in addition to that mentioned in the documents. The relationship shows the kinship between term 1 and term 2. The system determines four relationships between terms; these relations are sibling, cousin, child and uncle.

Sibling relationship:

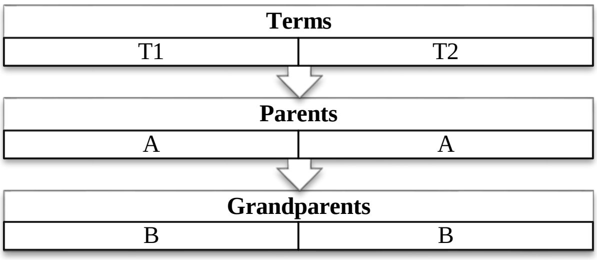


Figure 8: Sibling relationship.

As shown in Figure 8 if the parent of T1 is the same as T2 and the grandparent of T1 is the same as T2 then T1 and T2 are sibling.

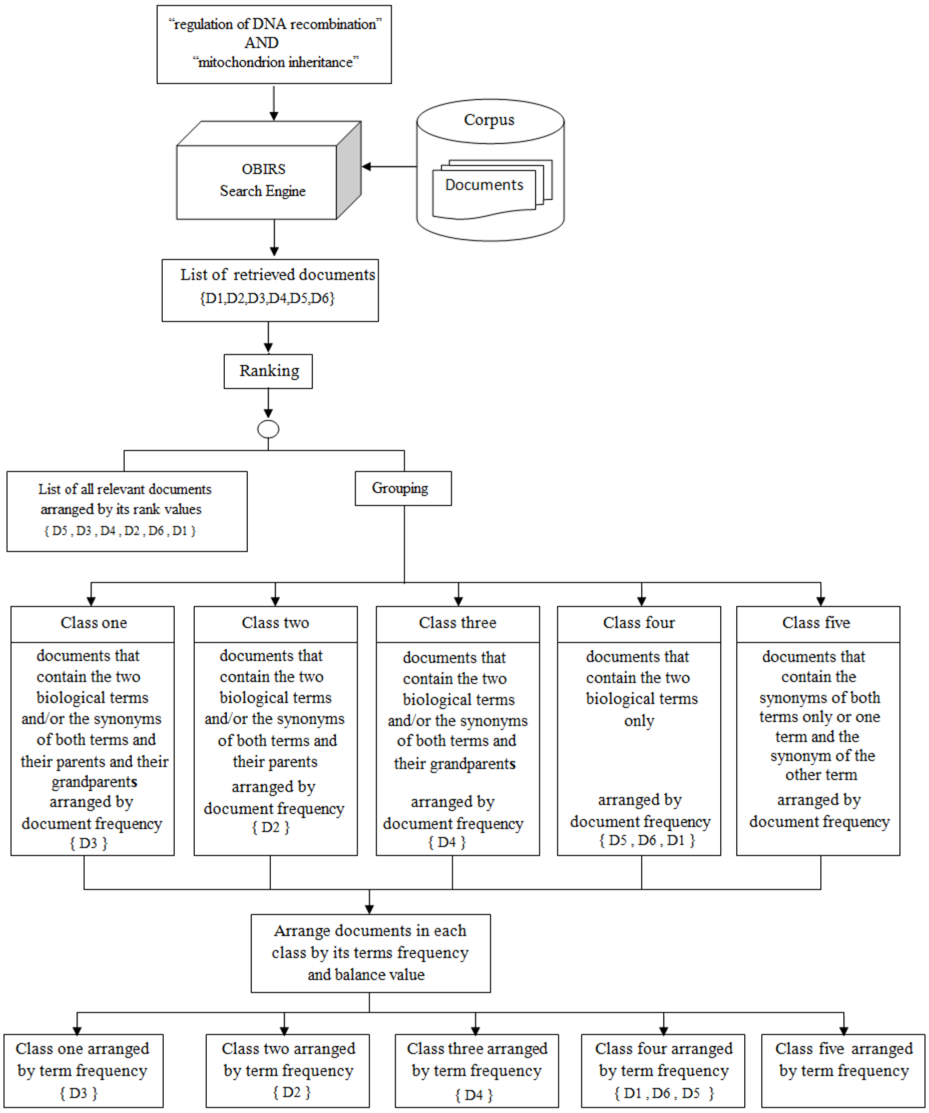
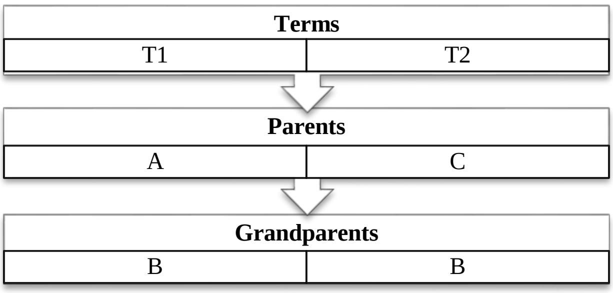

"regulation of DNA recombination"
AND
"mitochondrion inheritance"
OBIRS
Search Engine
Corpus
Documents
List of retrieved documents
{D1,D2,D3,D4,D5,D6}
Ranking
List of all relevant documents
arranged by its rank values
{ D5 , D3 , D4 , D2 , D6 , D1 }
Grouping
Class one
documents that contain the two biological terms and/or the synonyms of both terms and their parents and their grandparents arranged by document frequency
{ D3 }
Class two
documents that contain the two biological terms and/or the synonyms of both terms and their parents arranged by document frequency
{ D2 }
Class three
documents that contain the two biological terms and/or the synonyms of both terms and their grandparents arranged by document frequency
{ D4 }
Class four
documents that contain the two biological terms only arranged by document frequency
{ D5 , D6 , D1 }
Class five
documents that contain the synonyms of both terms only or one term and the synonym of the other term arranged by document frequency
Arrange documents in each class by its terms frequency and balance value
Class one arranged by term frequency
{ D3 }
Class two arranged by term frequency
{ D2 }
Class three arranged by term frequency
{ D4 }
Class four arranged by term frequency
{ D1 , D6 , D5 }
Class five arranged by term frequency

Figure 7: System workflow based on our example.

"Cousin" relationship:



Terms

T1 | T2

Parents

A | C

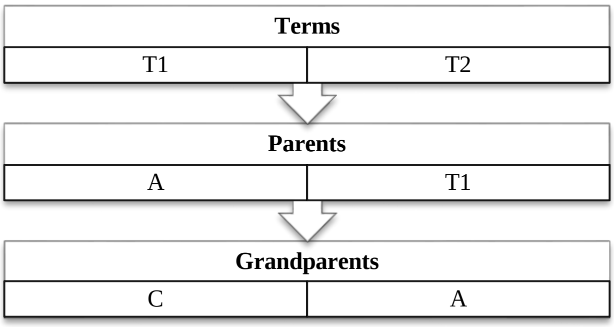
Grandparents

B | B

Figure 9: "Cousin" relationship.

As shown in Figure 9 if the parent of T1 is not the same as the parent of T2 and the grandparent of T1 is the same as T2 then T1 and T2 are cousins.

Child relationship:



Terms

T1 | T2

Parents

A | T1

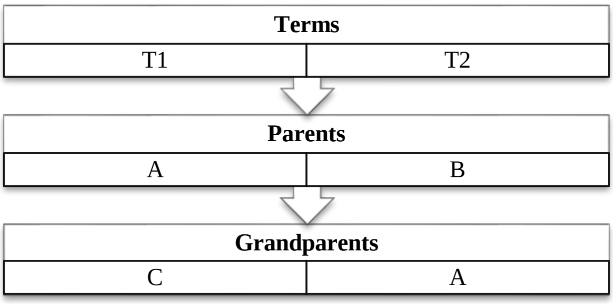
Grandparents

C | A

Figure 10: Child relationship.

As shown in Figure 10 if the parent of T1 is a grandparent of T2 and the parent of T2 is T1 then T2 is the child of T1.

"Uncle" relationship:



Terms

T1 | T2

Parents

A | B

Grandparents

C | A

Figure 11: "Uncle" relationship.

As shown in Figure 11 if the parent of T1 is a grandparent of T2 and the parent of T2 is not T1 then T1 is uncle of T2.

7 System Evaluation

Extensive experiments are preformed to study the effectiveness of our algorithm. The system was tested using corpus named craft [20] and the Gene Ontology version 1.2 [8].

The performance of the system is improved since we retrieve the documents that contain the two related terms and the related inferred terms (synonyms, parents and grandparents). Our system retrieves the documents with a certain criteria of ranking that helps the research to find the document that he/she searches for. The following are screenshots from the system that represent how does the system work.

Pre processing:

The two steps represented in Figure 12 invoked once at the beginning of the system

In step 1, the system builds the dictionary file from the normalized database "DBGenes". In step 2, the system builds the inverted list that helps in retrieving the desired documents.

Searching process:

After building the inverted list the user can make any number of the search queries he/she wants. Figure 13 shows the search query request from user, Figure 14 shows the retrieved relevant documents in two alternative methods for ranking.

System testing:

"DNA" and "RNA"have been entered as two

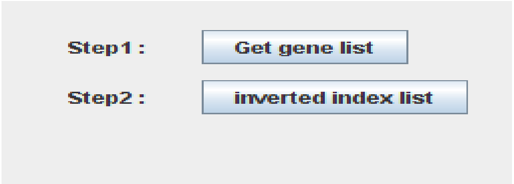
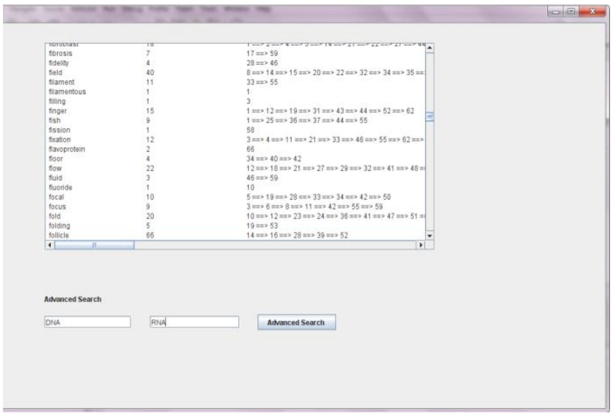


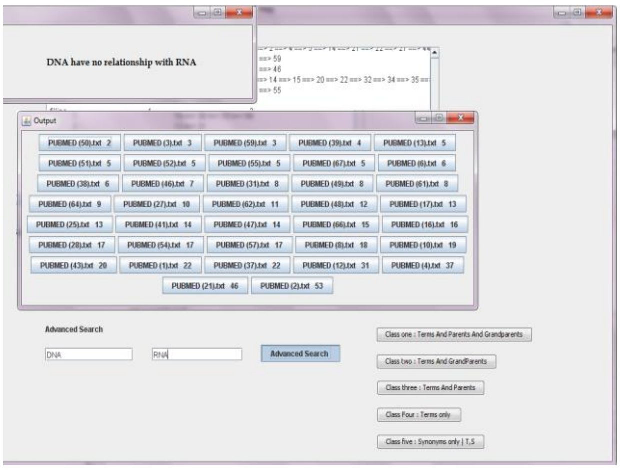
Figure 12: Screen shot of preprocessing.



```
fibrosis       7     17 ==> 59
fidelity       4     28 ==> 46
field          40    8 ==> 14 ==> 15 ==> 20 ==> 22 ==> 32 ==> 34 ==> 35 ==
filament       11    33 ==> 55
filamentous    1     1
filling        1     3
finger         15    1 ==> 12 ==> 19 ==> 31 ==> 43 ==> 44 ==> 52 ==> 62
fish           9     1 ==> 25 ==> 36 ==> 37 ==> 44 ==> 55
fission        1     58
fixation       12    3 ==> 4 ==> 11 ==> 21 ==> 33 ==> 46 ==> 55 ==> 62 ==>
flavoprotein   2     66
floor          4     34 ==> 40 ==> 42
flow           22    12 ==> 18 ==> 21 ==> 27 ==> 29 ==> 32 ==> 41 ==> 48 ==
fluid          3     46 ==> 59
fluoride       1     10
focal          10    5 ==> 19 ==> 28 ==> 33 ==> 34 ==> 42 ==> 50
focus          9     3 ==> 6 ==> 8 ==> 11 ==> 42 ==> 55 ==> 59
fold           20    10 ==> 12 ==> 23 ==> 24 ==> 36 ==> 41 ==> 47 ==> 51 ==
folding        5     19 ==> 53
follicle       66    14 ==> 16 ==> 28 ==> 39 ==> 52
```

Advanced Search: DNA | RNA

Figure 13: Search query request.



DNA have no relationship with RNA

```
Output
PUBMED (50).txt  2    PUBMED (3).txt  3    PUBMED (59).txt  3    PUBMED (39).txt  4    PUBMED (13).txt  5
PUBMED (51).txt  5    PUBMED (52).txt  5   PUBMED (55).txt  5    PUBMED (67).txt  5    PUBMED (6).txt  6
PUBMED (38).txt  6    PUBMED (46).txt  7   PUBMED (31).txt  8    PUBMED (49).txt  8    PUBMED (61).txt  8
PUBMED (64).txt  9    PUBMED (27).txt  10  PUBMED (62).txt  11   PUBMED (48).txt  12   PUBMED (17).txt  13
PUBMED (25).txt  13   PUBMED (41).txt  14  PUBMED (47).txt  14   PUBMED (66).txt  15   PUBMED (16).txt  16
PUBMED (28).txt  17   PUBMED (54).txt  17  PUBMED (57).txt  17   PUBMED (8).txt  18    PUBMED (10).txt  19
PUBMED (43).txt  20   PUBMED (1).txt  22   PUBMED (37).txt  22   PUBMED (12).txt  31   PUBMED (4).txt  37
                  PUBMED (21).txt  46   PUBMED (2).txt  53
```

Advanced Search: DNA | RNA

Advanced Search
Class one : Terms And Parents And Grandparents
Class two : Terms And GrandParents
Class three : Terms And Parents
Class Four : Terms only
Class five : Synonyms only | T,S

Figure 14: The set of retrieved relevant documents.

biological terms to the system and wanted to get all relevant documents that contain both terms from the craft corpus. The terms "DNA" and "RNA" have been added to the database although they have been removed from Gene Ontology since 2003 and they have been chosen to be searched for because they are very common in corpus and important in the search. The two terms found in the corpus as shown in Table 3.

Table 3: DNA and RNA statistics in craft 1.0.

Terms	Number of Documents that contain the term	Number of Documents that does not contain the term	Number of documents that contain one term and the synonym of the second term
DNA	55	12	18
RNA	44	23	7
Number of relevant documents (contain both term)	37		

The documents to be retrieved must have two terms entered by the user. In the previous experiment the system retrieves all the documents that contain both terms which are 37 documents.

Our assumptions insure that all the retrieved documents will be relevant documents as they contain the two biological terms entered by the user. After several experiments we calculated the precision and recall of the system and got a precision equal to 100% and recall equal to 100%.

The system gives these results because of the following points:

- The advantage of the "exact matching" for the query keywords and non-existence of the concept "partial matching" in the standard Boolean model. So documents can be retrieved if it contains the entered keywords otherwise it will not be retrieved.
- Biological keywords are unique. The "Polysemy" problem is absent, so there is no chance to have multiple words with the same meaning.

8 Conclusion and future work

This paper presents a semantic retrieval system that retrieves relevant documents with high performance. The system improves the performance of semantic information retrieving method since we use the Gene Ontology to infer related biological terms such as synonyms, parents and grandparents of the two related terms entered in the query to retrieve all relevant documents that contain these terms with any combination of the inferred terms. The system extracts the relations between the two related terms entered in the search query to give the researchers additional information about these terms.

In the system we used a ranking methodology to help in ordering the retrieved documents based on the rank values. The system groups the retrieved documents into five classes, each class can be ordered based on the frequencies of the input terms with concern of the balancing between the frequencies of input terms.

The system shows improvements in the percentage of the precision and recalls since it retrieves documents that actually contain needed information so all the retrieved documents are relevant ones.

The proposed system can be generalized to other domain specific fields. The authors use JAVA as a programming language to implement the system. JAVA has a limitation that affects the building of inverted list since it allows reading only 750 documents from the corpus. As a future work we aim to increase the number of documents read from the corpus by enhancing the index built in the system to be a multi-index that allows the system to read and store more terms from the documents and organizes the terms by other way. The presented system semantically expands the user query by parents and grandparents up to two levels in the Gene Ontology. As an improvement the system can use more than two levels from the Gene Ontology to enhance the semantic acting of the system. The system ranking issues can be changed and another ranking methodology can be used to get much close to the researchers' needs. The system ranking issues can be enhanced based on the researchers' feedback. The system grouping criteria's can be differed based on the application domain and can be decomposed based on domain requirements. The system extraction process can be enhanced to extract the relations between biological terms from the documents instead of the Ontology. Also other additional relations that are not mentioned in this research can be extracted. The system is based on two related terms and can be enhanced to use more than two terms.

References

[1] Parul Gupta and Dr. A.K.Sharma, *"Context based Indexing in Search Engines using Ontology"*, International Journal of Computer Applications (0975 – 8887), Volume 1 – No. 14, 2010.

[2] Marta Bleda, Joaquin Tarraga, Alejandro de Maria, Francisco Salavert, Luz Garcia-Alonso, MatildeCelma, Ainoha Martin, Joaquin Dopazo and Ignacio Medina, *"CellBase : a comprehensive collection of RESTful web services for retrieving relevant biological information from heterogeneous sources"*, Department of Bioinformatics and Genomics, Centro de Investigación Príncipe Felipe (CIPF), 46012 Valencia, Spain, Nucleic Acids Res,2012 Jul;40(Web Server issue):W609-14. doi: 10.1093/nar/gks575. Epub 2012 Jun 12, 2012.

[3] Meng Hu and Jiong Yang, *"A System of User-Guided Biological Literature Search Engine"*, EECS, Case Western Reserve University, 2005.

[4] ChengXiangZhai, *"Statistical Language Models for Information Retrieval A Critical Review"*, University of Illinois at Urbana-Champaign, 201 N. Goodwin, Urbana, IL, Foundations and Trends in Information Retrieval, Vol. 2, No. 3, 137–213, 2008.

[5] Mohammad Mustafa Taye, *"Understanding Semantic Web and Ontologies: Theory and Application, Journal of Computing"*, Volume 2, Issue 6, June Mohammad Mustafa Taye, ISSN 2151-9617, 2010.

[6] Thomas Eiter, GiovambattistaIanni, Thomas Krennwallner and Axel Polleres (2008) *"Rules and Ontologies for the Semantic Web"*, Reasoning Web, pp. 1-53.

[7] David Jin and Sally Lin, *"Advances in Computer Science"*, Intelligent Systems and Environment: Vol.1, Advances in Intelligent and Soft Computing, Springer - 1st Edition 1:134, 2011.

[8] The Gene Ontology Consortium, *"Gene Ontology: tool for theunification of biology"*. Nat. Genet., 25, 25–29, 2000.

[9] http://www.gene ontology.org/.

[10] MarwaMostafaMostafa,Enas M.F. El Houby and Akram Salah, *"Ontology-based Biological Information Retrieval System"*, Australian Journal of Basic and Applied Sciences, No-9177-AJBAS, 540-545 August 2012.

[11] C.Sumithiradevi, Dr.M.Punithavalli and S.Suresh, *"Biomining:-An Efficient Data Retrieval Tool for Bioinformatics to Avoid Redundant and Irrelevant Data Retrieval from Biological Databases"*, Global Journal of Computer Science and Technology, Volume XI Issue I Version I, 2011.

[12] Qi Zhou, Hong Zhang, Meiying Geng and Chenggang Zhang, *"A Real-Time and Dynamic Biological Information Retrieval and Analysis System (BIRAS)"*, Beijing Polytechnic University, Beijing, China, 2003.

[13] Minlie Huang, Xiaoyan Zhu, Shilin Ding, Hao Yu and Ming Li, *"ONBIRES:Ontology-based Biological Relation Extraction System"*, In Proceedings of the Fourth Asia Pacific Bioinformatics Conference, 2006.

[14] AnáliaLourenço, Rafael Carreira, Daniel Glez-Peña, José R. Méndez, SóniaCarneiro, Luis M. Rocha, Fernando Díaz, Eugénio C. Ferreira, Isabel Rocha, FlorentinoFdez-Riverola, Miguel Rocha, *"BioDR: Semantic indexing networks for biomedical document retrieval Expert Systems with Applications"*, 37(4), 3444-3453, 2010.

[15] Cui Tao, *"Information Extraction and Integration from Heterogeneous Biological Data Sources"*, Department of Computer Science, Brigham Young University, Provo, Utah 84602, U.S.A, 2006.

[16] Jiewen Wu, IhabIlyas and Grant Weddell, *"A Study of Ontology-based Query Expansion"*, Cheriton School of Computer Science, University of Waterloo, CS-2011-04, 2011.

[17] Philippe Thomas, Johannes Starlinger, Alexander Vowinkel, Sebastian Arzt and Ulf Leser, *"GeneView: a comprehensive semantic search engine for PubMed"*, Nucleic Acids Research, 40(Web Server issue):W585-91, 2012.

[18] M.C. Díaz-Galiano, M.T Martín-Valdivia and L.A. Ureña-López, *"Query expansion with a medical ontology to improve a multimodal information retrieval system"*, journal of Computers in Biology and Medicine, Elsevier Science, 2007.

[19] C. Pasquier, *"Biological data integration using Semantic Web technologies"*, Biochimie, 90(4), 584-594, 2008.

[20] Yungang Xu, MaozuGuo, Wenli Shi, Xiaoyan Liu, Chunyu Wang, *"A novel insight into Gene Ontology semantic similarity"*, Genomics, 101(6), 368-375, 2013.

[21] Craft1.0 http://bionlp-corpora.sourceforge.net/CRAFT/index.shtml

Tricorder: Consumer Medical Device for Discovering Common Medical Conditions

Maja Somrak[1], Mitja Luštrek[1], Jakob Šušterič[2], Tomo Krivc[2], Ana Mlinar[2], Tilen Travnik[3], Luka Stepan[4], Mitja Mavsar[3] and Matjaž Gams[1]
[1] Jozef Stefan Institute, Jamova 39, Ljubljana, E-mail: maja.somrak@ijs.si
[2] MESI, razvoj medicinskih naprav, d.o.o., Leskoškova cesta 9d, Ljubljana, E-mail: ana.mlinar@mesi.si
[3] Domenca Labs, d.o.o., Celovška 32, Ljubljana, E-mail: tilen.travnik@dlabs.si
[4] Gigodesign, oblikovanje in komunikacije, d.o.o., Kamniška ulica 49, Ljubljana, luka.stepan@gigodesign.com

Technical Position Paper

Keywords: diagnostic device, Tricorder, medical application, intelligent assistant

Received: January 12, 2014

The Qualcomm Tricorder XPRIZE $10 million competition will open the doors for health and wireless technology. The aim is to design a portable, wireless device that monitors and diagnoses health conditions of residents without medical knowledge. The radical innovation in healthcare will give individuals far greater choices in when, where, and how they receive care. In this paper we present the competition itself and in particular the research prototype of the MESI Simplifying diagnostics team. The deadline of the first part of competition is approaching, after which the ten chosen teams will compete in the second, final round. Our approach builds upon the previous research and applications of several interdisciplinary partners constituting the team. The modular prototype fulfills the Tricorder demands by 24/7 monitoring the user and asking for medical help when needed. In the first stage it enables medical analyses of 15 prescribed medical conditions without demanding any particular medical knowledge. The additional novelties include: a modular structure enabling inclusion of any medical or ambiental sensor or device through web of things, additional previous medical applications not set out in the competition such as the detection of diseases manifesting in movement or activity recognition, and designing an intelligent medical assistant to take care of the user.

Povzetek: Opisano je tekmovanje Qualcomm Tricorder XPRIZE s skupnimi nagradami v višini 10 milijonov dolarjev in pristop slovenskih partnerjev.

1 Introduction

The Qualcomm Tricorder XPRIZE competition [1] is a global competition with the aim to stimulate the development of technologically advanced medical devices bringing accurate health diagnostics to the consumers. Currently, there are 30 international teams actively registered in the competition, which represents around one tenth of those that expressed initial interest. Teams will compete in terms of diagnostic accuracy and functionality, as well as user experience. The submission deadline for the Qualifying Round is May 2014, and from those that will submit the desired contributions, ten selected teams will advance in September to the Final Round scheduled to take place in the first half of 2015. Finally, up to three teams will be awarded a prize in total sum of $10 million.

The device envisioned for the competition will integrate innovative sensing hardware with advanced artificial-intelligence techniques. The convenient and portable design of the device will allow for anytime, anywhere, reliable health assessment, independent of medical professionals. The application of such devices could improve the utilization of healthcare resources by reducing unnecessary doctor’s appointments, as well as improve personalized health care. This radical new approach will give individuals far greater choice in their own health-care by delivering health-care tools directly into their hands. A significant emphasis is put on user experience to ensure that the user will be able to use it correctly (no medical background needed) and that they will want to use it.

The target device is also based on a permanent and even continuously increasing demand for simple, easily accessible, and reliable diagnostic methods and systems. Currently, one of the most widely available tools for health assessment are online symptoms analysers, for instance the WebMD [2] or Mayo Clinic symptom checkers [3]. However, these are known to be unreliable, especially in the absence of additional diagnostics or healthcare professional consultation [4]. Recently, new innovative devices for remote diagnosis had been developed. One of such products for physical examination is the Tyto™ care, which combines multiple different technologies in their telemedical device [5]. Additionally, there is the need for continuous monitoring of vital signs, which is not provided by the aforementioned solutions. Various devices for monitoring vital signs are available in the market and range from wearable chest straps to more user friendly wristbands with integrated medical-grade sensors [6]. A novel, innovative solution should encompass both the continuous monitoring and reliable, advanced diagnostic methods in a single device. One of the most successful attempts in this direction represents the Scanadu Scout™ [7], one of the favourites in the Qualcomm Tricorder XPRIZE competition.

Moreover, there has been a persistent desire over several decades to design artificial intelligence (AI) assistants for various human tasks [8]. In recent years, humans use advanced assistants that might be perceived as somehow intelligent, on a regular basis: Google Now [9], question-answering systems like iOS Siri [10] or Android Assistant [11]. However, even though these systems are massively used, none of them exhibits true intelligence. True, they seem intelligent to naïve users and they are able to solve the tasks they are designed for reasonably well; however, they fail in a couple of sentences in particular when examined by those familiar with the Turing test. The idea of intelligent assistants emerges once again in the next generation of diagnostic devices – why not use these assistants and integrate them into a medical system? However, a viable, state-of-the-art solution that would successfully integrate assistants and other successful AI methods as quite diverse technologies in a single system, remains a challenge.

This paper describes our entry in the Tricorder competition, with the goals specified. Our approach currently mainly integrates machine learning techniques and multiple diagnostic hardware modules. It allows for a highly scalable solution with a possibility of subsequent extensions.

2 Our approach

Our approach is based on a combination of the concept of the Sci-Fi Tricorder – a multifunction hand-held device used for sensor scanning, data analysis, and recording data [8] – and the engineering and market approach, resulting in a prototype that will be able to perform well in the experimental tests. These two conditions seem to be orthogonal on each other and finding the compromise is the central issue of the competition.

The AI part of our approach integrates the following:

1. Design an AI-based medical assistant that will gather all information from the sensors on the user and in the environment and other sources of knowledge, with the goal to use this information for the benefit of the user.
2. The agent-based system will be highly modular and interdisciplinary.
3. The system will be able to communicate with the user through the graphical user interface (GUI) and in natural language.
4. The system will diagnose not only the diseases set out in the Tricorder competition, but also additional diseases and conditions, and if possible also predict probability of future complications.
5. The task is to infer short-, mid- and long-term conclusions about the user's medical situation, tuned to characteristics of each user.

(1) Concepts like the *Internet of things* [12][13] and *Sensor fusion* [14] enable the integration of several independent sources of data into meaningful information. Our approach combines the two concepts and also provides an integration at a novel level, as presented in [15] where each of the sensor was treated as a context and all the other sensors as input data from which the machine learning model of the domain was constructed. We call this approach *multiple-context sensor fusion*. For example, if the level of physical activity is taken as the context, and all the other sensor data are used for machine learning, we can reason as follows: in the context of low physical activity, a high heartbeat (and some other characteristics) indicates an alarming situation. Use of context is essential for quality performance in real-life circumstances and we expect the same will happen when it will be implemented in the second stage of the competition. Currently, we have implemented the algorithm for several tasks and expect no problem implementing it also in the Tricorder prototype. However, one should note that successful use of several AI modules mostly designed in previous projects with several tens of thousands lines of code each, far exceeds the capacities of current mobile devices. Therefore, services in the cloud are the only operational option for now.

(2) The ability to gather information from any devices that can be contacted through predefined standard communication protocols is already an indication of modularity and interdisciplinarity. The system needs to gather all available information from all sources and make use of it. Integrating data from devices included in the Tricorder, and learning from observation, may – for example – enable automatic learning that lower temperature in a room can result in a common cold for an individual user. If the temperature in a living room during regular monitoring significantly decreases for a substantial amount of time, a warning is thus issued that a low temperature previously caused a cold and that it is recommended that the temperature is increased. Currently, this functionality is not implemented yet.

In addition, the Tricorder system should be modular in a sense that if another device component is connected, it should be easily incorporated into the whole system. For example, connecting a heart-rate monitor or disconnecting it should not cause any errors in the system. The solution for this has been long known – agent systems enable the most flexible architecture. Our approach is based on JADE [16] designing a research prototype with the desired flexibility. We have previously designed MASDA system for analyses of soccer strategies [17] and designing cognitive and behavioural clones for teaching team commander how to deal with hostile crowds [18]. Currently, we are using agent architecture in smart houses and smart cities [19][20]. These systems already demonstrate large amount of elasticity, autonomy, interdisciplinarity and ability to deal with several heterogeneous sources of knowledge and integrating them into one functional system. To implement it in our Tricorder device, the current architecture will have to be upgraded in the second stage of the competition. However, several versions of the systems were already designed at a level of independent prototypes, including systems for activity recognition based on accelerometers of a mobile phone, in a bracelet, or separately attached sensors to various body parts. Some of these systems already used a sophisticated architecture [22]. Another of our systems [23] combines text and image marked and sent through a smart phone to estimate the probability of the Lyme disease. The current systems is already flexible in a way that adding or deleting one or more sensors is a rather simple task; however, no complex architecture or agent approach is currently implemented in our system.

(3) User communication can be performed through a classical text or highly visual interface. In addition, a natural language assistant is planned to be a part of our Tricorder. Two examples of our already implemented natural language assistants are Robi for the Jozef Stefan homepage [23] or Svizec for the union homepage [25]. These systems are implemented in a cloud and are available for major mobile platforms like Android or iOS. Adding a natural language interface into our Tricorder system would demand a couple of months of work. An example of practical use would be improved communication when the user does not understand the question or a message displayed on the screen. Similarly, the system is supposed to talk to the user not only through predefined speech sequences, but also from the dynamically generated text. We have designed such a platform for several man-machine projects [26].

(4) In previous experiments, diseases that manifest in the patient's motion, like the Parkinson's disease, were already detected with over 90% accuracy on simulated and real patients [27]. The methods used were based on integration of machine learning, dynamic time warping (DTW) and semantic attributes for each disease. It was demonstrated that each of these mechanisms improves the classification accuracy and that the most successful combination integrates them all. While only 5 diseases were tested so far, this research was an indication that several diseases can be successfully detected from movements only. In addition, the increase in these characteristic signs also enables the prediction of how fast the condition will deteriorate, thus enabling preventive actions. Several systems of this kind have been developed at the Jozef Stefan institute. It is feasible to implement these services in the cloud in several months each. The usefulness of these modules is clearly dependent on the success rate in practical tests – one of the future-work tasks. But noticing that a person limps for several hours or that the hands shake more than normally, or that a spasm is present in one arm for some time period are sufficient to call for help. The system will be able to communicate with the user or even bypass the user and call immediately for help in case of emergence or lack of user response – if set up in this way for example for an elderly living home alone.

(5) Several of our systems already deal with short-, mid- and long-term situations. They observe the performance of a particular user and learn his/her habits and performances [28][29]. For example, if a person already limps, then only differences to the common gait are looked for. Similarly, blood pressure of 140/90 can be an alarming news for one person and a good news for another. The team has designed several systems already that proved their performance, e.g. the Jozef Stefan Institute has won the live activity and fall detection EvAAL competition [30] and demonstrated another system at the European AI conference [31].

While the team has been designing all these subsystems for specific tasks, meaning that we have these prototypes developed and several of them in regular use, they are used for different tasks (see the relevant publications). Due to time constraints we were not able to introduce these functionalities into our existing prototype, but we estimate it is feasible to modify and integrate the already designed subsystems in half a year, at least at a prototype level with most of the computing performed in the cloud. Currently, we can only demonstrate each of the subsystem on its own.

However, an early version of a system focused on the practical part of the Tricorder competition has already been implemented and tested, and is presented in the next section. Partially, it already encompasses some of the AI sub-models presented in this section.

3 MESI Tricorder competition entry

In this section we describe our working prototype designed for prototype use at the Qualifying Round.

3.1 The system

Our Tricorder system was already presented at the CeBIT fair [32]. Its schema is presented in Figure 1. It consists of a bracelet for monitoring the vital signs, a mobile device with application for communication with the user and several applications including local computing methods and connection to the services in the cloud, and specialized modules for additional tests to determine specific diseases.

The bracelet measures vital signs designed for everyday use:

1. The ECG is the recording of the electrical activity of the heart. The signal is obtained by touching three electrodes, two of them with a wrist and the third with a finger of the other hand. Although only two electrodes are enough to measure single channel ECG, we added a third one, which is used for the noise reduction – similar to a right-leg drive (RLD) electrode in standard ECG devices. We improved and minimized the technology to fit our tight

housing and occupy only one square centimetre of printed circuit board (PCB).

2. The oxygen saturation (SpO2) is a measure of percentage of haemoglobin that has already bounded with oxygen. It is calculated by measuring the reflected red and infrared light that is sent to the fingertip. After post-processing the acquired data the device calculates respiratory rate.
3. The third sensor inside the bracelet is the temperature sensor. For best performance and quicker measurement, we use an infrared sensor.

For accurate and continuous measurements of vital sign user is required to use the so called *shield.* It consists of wireless cuff for measuring blood pressure using an automatic oscillometric method and a patch located on the ribs for measuring oxygen saturation, temperature, electrocardiogram (ECG), respiratory rate and activity tracking. Data obtained by vital signs and activity recognition help diagnose several diseases such as atrial fibrillation, hypertension and sleep apnea.

The second part of device consists of four in-depth modules for diagnosing 15 diseases:

1. The first module is so-called *To see*. A charge-coupled device (CCD) image camera with controlled standard and polarized white light is used for detecting melanoma and streptococcal pharyngitis. Special mount enables standard distance from the skin for obtaining real size of melanoma. For detecting otitis media wideband technology is used. High frequency sound impulses similar to Dirac impulse are transmitted to ear channel to move the eardrum. The reflected sound is recorded with high sensitivity digital microphone and processed with microcontroller. It calculates the fast Fourier transform (FFT) of the reflected signal and compares it with the pre-collected database.
2. The second module is *To hear*. Smart electronic stethoscope is designed for detecting pulmonary diseases. It has a microphone with high signal-to-noise ratio (SNR) attached to a specially designed bell, similar to the standard stethoscope chest piece, to record the low frequency sounds. We are developing software that is able to identify sounds and noises typical for each of pulmonary conditions. On the other side of the module, there is a second microphone for measuring air lung volume and speed or flow of inhaled and exhaled air. It is used to diagnose chronic obstructive pulmonary disease (COPD).
3. The *Urine module* analyses urine using test strips that are scanned by camera and automatically processed with computer vision. The mount on the camera enables standardization of light and distance for accurate results.
4. The fourth is the *Blood module*. To achieve the best possible user experience we are avoiding invasive methods and taking advantages of the spectroscopy technology. The blood module is able to diagnose anaemia and diabetes.

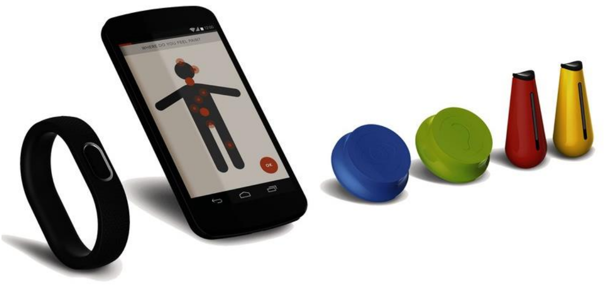


Figure 1. The CeBIT version of our Tricorder system consists of the bracelet, a mobile device with mobile applications and additional hardware devices for particular tests.

3.2 User experience

The proposed device aims to despecialize an aspect of primary health-care by giving end consumers the insight into their health on an instant basis. Through a user-centric design process we intend to understand the basic human needs and appropriate the technology at hand in such a way that it makes sense to a non-specialist. At the same time we want to engage the user more frequently, encouraging involvement and provided a detailed insight.

The system additionally tries to record vital signs in a greater extent that can be used as a detailed insight into the state before visiting a general practitioner (GP). Our aim was to solve as much of the diagnosing process only through a user-friendly questionnaire while the separate modules are intended for diagnosis confirmation. The questionnaire on its own already enables a more informed referral to a GP. Many specialist units outperformed a single multi-functional unit in our user testing for readability and ease-of-use. Modules are essentially multi-sensor units, packaged as “digital senses” for the end consumer. These units act together as an internet-of-things ecosystem and can be expanded as necessary in the future.

3.3 The diagnostic algorithm

We developed a novel algorithm for the initial assessment of the user’s medical condition, which could be either a healthy condition or one of the preselected 14 diseases. The algorithm combines two approaches: user interaction with a questionnaire and machine learning to make a tentative diagnosis.

The algorithm is implemented on a mobile device (see Figure 2) and focuses on the first of the two stages in the diagnostic procedure. Namely, during the first stage the user answers questions from a questionnaire to establish an initial diagnosis that is primarily informative. In the second stage, a special, additional device – a diagnostic module – is used for medical test and final diagnosis. The diagnosis can be additionally confirmed by an expert, if the user chooses to seek professional help in given circumstances. However, the questionnaire algorithm is one of the key components in this diagnostic process since it results in the first diagnosis to be later confirmed or rejected.

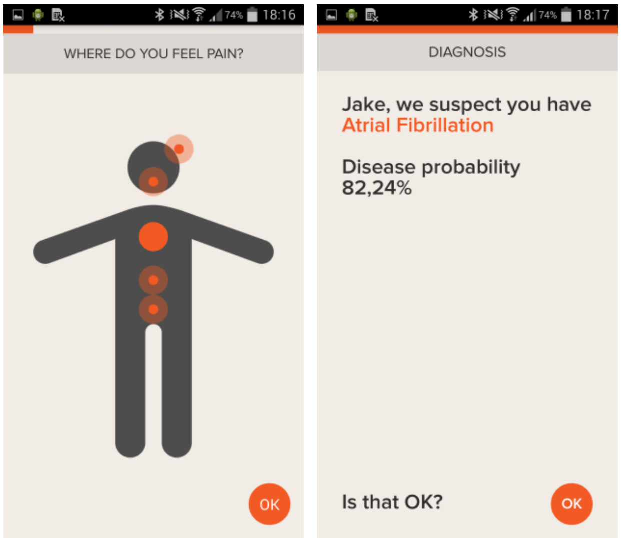
The algorithm consists of five steps:

1. Information such as the user's profile, vital-signs measurements and pain symptoms serve as initial inputs to the algorithm.
2. The initial inputs are used for the system to deduce a list of probable symptoms, from which the user selects main symptoms that he/she is experiencing.
3. The selected main symptoms, together with the initial inputs serve as the input to the algorithm for the medical-condition probability calculation. The medical conditions are shown in Table 1.
4. The user is asked for additional symptoms until the medical conditions emerge out of an uncertainty range that represents an undefined condition.
5. The medical condition with the highest probability is presented to the user as a tentative diagnosis, which should be further confirmed by a diagnostic hardware module.

The output of the first step of the algorithm is a list of probable symptoms, created by utilization of association rules and information-gain (IG) ranking. The input about the symptoms is used by the third and fourth step to provide relevant questions to the user. A new question is chosen according to the expected most informative symptom (symptom with highest IG). Each time the user answers a question, the probability for each medical condition is retrieved from the J48 classifier, dedicated to that condition. The classifiers were trained on data set of 15.000 virtual patients, generated using expert medical knowledge. The virtual patients were generated utilizing the table that relates symptoms to the diseases. For example, a patient with a single disease is generated with one or several symptoms for a randomly chosen disease. The probability of a patient having any of the medical conditions was uniformly distributed. Additionally, the probabilities of frequent combinations of the diseases had been set according to the medical experts. Only the last two percentages of all virtual patients were designed to have symptoms of a randomly selected combination of any two diseases.

The calculated probabilities of medical conditions may fall within the so called certainty or uncertainty range. The uncertainty range is a range of probability values where specific medical condition is neither very probable nor very improbable. The thresholds (low and high) define the uncertainty range.

The final output of the algorithm consists of the predicted medical condition, its probability and a selection of a further diagnostic module for the test. In this test, user is subdued to a single medical test, according to the initial diagnosis. A particular test is capable of recognizing any medical condition that belongs under the same diagnostic module (see Table 2) and is as such not necessarily limited to confirming only the initially predicted condition. This means that even an incorrect initial diagnosis and a correct diagnostic module selection can lead towards correct final diagnosis. Therefore, patients that are initially incorrectly diagnosed in the first stage, will have correct diagnosis in the second stage, if the diagnostic module will be able to induce the correct diagnose and will perform reliably.



Jake, we suspect you have Atrial Fibrillation

Disease probability 82,24%

Is that OK?

Figure 2. Examples of questionaries' menus and diagnostics.

4 Experiments

While learning was performed on 15.000 patients, testing was performed on a separate dataset of 1500 patients, where each medical condition was present in at least 100 patients. The test set was generated in a similar way as the original data set, in a way that the examples were different for each of the used sets.

Table 1. Initial prediction of the algorithm.

Initial medical condition prediction	Sensitivity	Specificity
Healthy	0.61	0.62
Hypertension	0.99	0.88
Atrial fibrillation	0.99	0.95
Acute haemorrhagic stroke	0.96	0.98
Obstructive sleep apnea	0.86	0.97
Hepatitis A	0.91	0.99
Otitis media	0.94	0.84
Streptococcal pharyngitis	0.95	0.97
Tuberculosis	0.96	0.99
COPD	0.99	0.95
Acute viral pneumonia	0.93	0.86
Lower urinary tract bacterial infection	0.99	0.99
Microcytic iron deficiency anaemia	0.83	0.95
Leucocytosis	0.59	0.6
Diabetes type 2	0.76	0.75

The results of initial medical condition prediction (initial diagnosis) are shown in Table 1.

Sensitivity is the probability that a person with a certain disease is correctly identified. Sensitivity is defined as a number of patients with the disease, who are correctly classified (true positives, TP), divided by the sum of both correctly classified and incorrectly classified patients with the disease (false negatives, FN). The formula for sensitivity is as follows:

$$\text{SENSITIVITY} = \text{TP} / (\text{TP} + \text{FN}).$$

Specificity is the probability that a person, identified as having the disease, is correctly identified. It is defined as a number of patients with the disease, who are correctly classified (TP), divided by the sum of both correctly classified patients with the disease and incorrectly classified patients without the disease (false positives, FP). The formula for specificity is as follows:

$$\text{SPECIFICITY} = \text{TP} / (\text{TP} + \text{FP}).$$

From Table 1 one can notice that healthy patients are least successfully identified. The reason for this is that both our diagnostics and the competition are primarily focused on correctly identifying the people in which a disease is present. However, on average, the obtained sensitivity of 0.884, specificity of 0.886 and accuracy of 0.883 seem quite acceptable for real-life trials.

Table 2. Corrected prediction of the algorithm.

Corrected medical condition prediction	**Sensitivity**	**Specificity**
Healthy	0.61	0.62
Hypertension	0.99	0.88
Atrial fibrillation	0.99	0.95
Acute haemorrhagic stroke	0.96	0.98
Obstructive sleep apnea	0.86	0.97
Hepatitis A	0.91	0.99
Camera module (*To see*)		
Otitis media	0.94	0.84
Streptococcal pharyngitis	0.95	0.97
Microphone module (*To hear*)		
Tuberculosis	0.96	0.99
COPD	0.99	0.99
Acute viral pneumonia	0.99	0.86
Urine module		
Lower urinary tract bacterial infection	0.99	0.99
Blood module		
Microcytic iron deficiency anaemia	0.92	0.95
Leucocytosis	0.63	0.77
Diabetes type 2	0.86	0.81

As mentioned in the previous chapter, there are some patient examples, which are incorrectly classified in the first stage, but their classification is corrected during the second stage of the diagnostic process, as a result of the correct module selection. The results that additionally account for these examples are presented in Table 2.

The average sensitivity is then 0.903, the average specificity 0.904 and average accuracy is 0.901.

The length of the list with probable symptoms was predefined to seven symptoms. From this list, on average two symptoms had been selected as present by the patient. The algorithm asked less than four additional questions, on average, to make the initial diagnosis.

The results suggest that such a questionnaire is both user friendly and efficient in terms of diagnostic accuracy.

5 Discussion

The Qualcomm Tricorder XPRIZE competition aims at revolutionizing home medical care through advances in hardware such as electronics and mobile devices, and software such as advanced applications and AI services. The last is related to the growing sense of optimism in the AI community. More and more established AI researchers believe that we are already in the transition period according to the “singularity theory” [33] According to Kurzweil [34] advances in electronics and artificial intelligence will enable the human civilization to jump ahead as use of metal enabled a jump from the stone- to metal age.

Our prototype already enables first analyses of vital signs and diagnoses of the selected 15 medical conditions. First tests were performed on virtual patients, since we were not able to perform the tests on live patients under the given circumstances. As a consequence, one might argue that the obtained results (sensitivity, specificity, etc.) were too optimistic. We agree that only clinical trials can represent efficiency in real life; however, the first experiments provide certain hope that this approach will be fruitful. One should bear in mind that according to many publications, internists achieve substantially lower performance in the first trial without specialized tests at other locations. Additionally, people without the necessary medical knowledge would benefit enormously if the device achieved comparable results than the one in our experiments.

Another cause why we have obtained such good results might be that number of the diagnoses was limited to 15 and each disease was very well characterized with its symptoms. Furthermore, each attribute was treated as correct without errors in judgment about the particular symptom. In real life, proper detecting of symptoms is a difficult task on its own. All these issues are to be addressed in future work.

In addition to correct tests with real-life patients and improvements of the existing device, another major issue is prevalent: to introduce the advanced AI and sustain high functionality in real-life circumstances we plan to modify and integrate the individual functional modules that partners had developed so far. That alone would enable design of a far more capable and intelligent system with agent structure, advanced multiple learning capabilities, sensor fusing, full internet of things, modular and interdisciplinary design, communication through both GUI and in natural language, user adaptability and adaptation. For the Final Round we intend to upgrade the system in that direction. The system is targeted to become a medical assistant making permanent observations and taking care of the user in a manner of intelligent agent assistants with certain degree of autonomy. The improved assistant observations would also play an important role in case the user wants analyses of his/hers medical conditions and assess probabilities of diseases and need of professional medical help.

In summary: as predicted by the organizers of the Qualcomm Tricorder XPRIZE competition, the device to revolutionize home care is on the brink of a break-through. We hope to contribute to these world-wide efforts.

References

[1] Qualcomm Tricorder XPRIZE, http://www.qualcommtricorderxprize.org/

[2] WebMD, http://symptoms.webmd.com/

[3] Mayo Clinic Symptom Checker, http://www.mayoclinic.org/symptom-checker/select-symptom/itt-20009075

[4] GUALTIERI, L.N. The doctor as the second opinion and the internet as the first Information. Technology Interfaces, 2009. ITI '09. Proceedings of the ITI 2009 31st International Conference on: June 22-25, 2009, Dubrovnik, Croatia.

[5] Tyto™ care, http://tytocare.com/

[6] W/Me medical-grade wristband, https://www.kickstarter.com/projects/723246920/finally-a-wearable-device-that-can-improve-your-li

[7] Scanadu Scout™, https://www.scanadu.com/

[8] Tricorder, http://en.wikipedia.org/wiki/Tricorder

[9] Google Now, http://bgr.com/2012/11/15/google-now-wins-popular-science-award/

[10] iOS Siri, http://www.apple.com/ios/siri/

[11] Android Assistant, https://play.google.com/store/apps/details?id=com.speaktoit.assistant

[12] IEEE Intelligent systems: Web of things, Vol. 28, No. 6, November/December 2013

[13] IEEE Computer: Internet of things, Vol. 46, No. 2, February 2013

[14] Information Fusion, An International Journal on Multi-Sensor, Multi-Source Information Fusion, Elsevier.

[15] Gjoreski, Hristijan, Kaluža, Boštjan, Gams, Matjaž, Milić, Radoje, Luštrek, Mitja. Ensembles of multiple sensors for human energy expenditure estimation. V: The 2013 ACM International Joint Conference on Pervasive and Ubiquitous Computing, September 8-12, 2013, Zurich, Switzerland. UBICOMP 2013. Washington: Association for Computing Machinery = ACM, pp. 359-362.

[16] Bellifemine, Fabio Luigi, Caire, Giovanni Greenwood, Dominic. Developing Multi-Agent Systems with JADE, Wiley, 2007.

[17] Chong, Nak-Young, Mastrogiovanni, Fulvio. Handbook of Research on Ambient Intelligence and Smart Environments: Trends and Perspectives, IGI Global, 2011

[18] Tavčar, Gašper, Gams, Matjaž, Kvassay, Marcel, Laclavík, M., Hluchý, L., Schneider, B., Bracker, H. Graph-based analysis of data from human behaviour simulations. V: SAMI 2012: IEEE 10th Jubilee International Symposium on Applied Machine Intelligence and Informatics: proceedings: January 26-28, 2012, Herl'any, Slovakia. Piscataway: IEEE, cop. 2012, pp. 421-426.

[19] IEEE Internet Computing: Smart Cities, Vol. 17, No. 6, November/December 2013

[20] http://www.artemis-ia.eu/project/index/view?project=50

[21] OpUS, http://dis.ijs.si/Opus/

[22] Kozina, Simon, Gjoreski, Hristijan, Gams, Matjaž, Luštrek, Mitja. Efficient activity recognition and fall detection using accelerometers. V: BOTÍA, Juan A. (ur.). Evaluating AAL systems through competitive benchmarking: International Competitions and Final Workshop, EvAAL 2013, July and September 2013, [Madrid, Valencia]: proceedings, (Communications in Computer and Information Science, ISSN 1865-0929, 386). Heidelberg [etc.]: Springer, 2013, pp. 13-23.

[23] Čuk, Erik, Gams, Matjaž, Možek, Matej, Strle, Franc, Maraspin-Čarman, Vera, Tasič, Jurij F. Supervised visual system for recognition of erythema migrans, an early skin manifestation of lyme borreliosis. Strojniški vestnik, ISSN 0039-2480, Feb. 2014, vol. 60, no. 2, pp. 115-123, ilustr., doi: 10.5545/sv-jme.2013.1046.

[24] Robi, http://www.ijs.si/

[25] Svizec, http://www.sviz.si/svizec/

[26] Asistent, http://dis.ijs.si/projekt-asistent/?page_id=100

[27] Pogorelc, Bogdan, Gams, Matjaž. Detecting gait-related health problems of the elderly using multidimensional dynamic time warping approach with semantic attributes. Multimedia tools and applications, ISSN 1380-7501, 2013, vol. 66, no. 1, pp. 95-114, doi: 10.1007/s11042-013-1473-1.

[28] Kaluža, Boštjan, Cvetković, Božidara, Dovgan, Erik, Gjoreski, Hristijan, Mirchevska, Violeta, Gams, Matjaž, Luštrek, Mitja. A Multiagent care system to support independent living. International journal on artificial intelligence tools, ISSN 0218-2130, [in press] 2014, 10 pp., doi: 10.1142/S0218213014400016.

[29] Cvetković, Božidara, Kaluža, Boštjan, Luštrek, Mitja, Gams, Matjaž. Adapting Activity Recognition to a Person with Multi-Classifier Adaptive Training. Journal of Ambient Intelligence and Smart Environments, accepted for publication (2014).

[30] EvAAL competition, http://dis.ijs.si/?p=1320

[31] Luštrek, Mitja, Kaluža, Boštjan, Cvetković, Božidara, Dovgan, Erik, Gjoreski, Hristijan, Mirchevska, Violeta, Gams, Matjaž. Confidence:

ubiquitous care system to support independent living. V: RAEDT, Luc de (ur.). ECAI 2012: 20th European Conference on Artificial Intelligence, 27-31 August 2012, Montpellier, France including Prestigious Applications of Artificial Intelligence (PAIS-2012) systems demonstrations track, (Frontiers in artificial intelligence and applications, ISSN 0922-6389, v. 242). Amsterdam: IOS Press, cop. 2012, pp. 1013-1014.

[32] CeBIT, http://www.cebit.de/product/qualcomm-tricorder-xprize/470055/Q558791?source=pkl

[33] IJCAI 13, Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence, Peking, China, ISBN 978-1-57735-633-2, (2013).

[34] Kurzweil, Ray. The Singularity is Near. New York: Viking Books, (2005).

[35] MESI team, http://simplifyingdiagnostics.com/xp/#team

From Qualitative to Quantitative Evaluation Methods in Multi-criteria Decision Models

Biljana Mileva Boshkoska
Faculty of information studies, Ulica talcev 3, Novo Mesto
E-mail: biljana.mileva@fis.unm.si
Jožef Stefan Institute, Jamova 39, Ljubljana
E-mail: biljana.mileva@ijs.si

Thesis Summary

Keywords: option ranking, DEX methodology, copula functions

Received: March 26, 2014

This article presents a summary of the doctoral dissertation of the author with title "From qualitative to quantitative evaluation methods in multi-criteria decision models".

Povzetek: Članek predstavlja povzetek doktorske disertacije avtorja z naslovom "Od kvalitativnih do kvantitativnih metod vrednotenja v večparameterskih odločitvenih modelih".

1 Introduction

The thesis [1] addresses the decision making problematic of ranking a finite set of qualitative options that are sorted into a set of classes. In the decision problems of interest, the options are represented with qualitative attributes that form a decision table. The decision maker's preferences split the decision table into subsets of equally preferred options, called classes, so that options belonging to the same class are considered indistinguishable. In practice, this is often inadequate and hence one wants to further distinguish between options belonging to the same class, and consequently rank them. Furthermore, the wish is to obtain such rankings with least effort, i.e., using only the information already available in the decision table.

The thesis presents a modeling approach that combines qualitative and quantitative models.

The resulting quantitative model should be in some way consistent with the original, qualitative one and should be preferably constructed in an automatic or semi-automatic way from the information contained in the qualitative model. These are very important questions, both theoretically and practically. Theoretically, it is important for bridging the gap between both types of models and involves a number of theoretically interesting sub-problems, such as finding a suitable representation of a decision problem in different forms for different computational process, within the same decision-making process Practically, bridging this gap is important to overcome some limitations of qualitative models, such as low sensitivity and limited applicability for the ranking of options.

2 Methods used

The problem addressed here is directly motivated by decision expert (DEX) methodology [2, 3], that, in the process of developing a decision model, produces decision tables which can be interpreted either as a set of options or a set of decision rules governing the preference evaluation. The existing qualitative-quantitative method (QQ) [4] developed for solving the ranking problem, is based on the assumption that when qualitative data are suitably mapped into discrete quantitative ones, they form monotone or nearly linear functions. The main limitation of QQ is that in many cases it fails to model non-linear functions. There are other qualitative MCDM methods that also deal with this issue, however, none of them solves the problem stated above. To solve this issue, we propose and evaluate four different QQ-based methods for estimating a regression function.

The first method includes investigation of different impurity functions for estimation of coefficients in the linear regression equation used by QQ. The main contribution arising from this method is the usage of different non-linear functions instead of the standard least squares algorithm, that lead to full rankings of many non-monotone decision tables, for which QQ provides equal rankings (ties) of options or fails to fulfill the monotonicity of the rankings.

The second method introduces polynomial functions instead of the linear one in QQ. For that purpose the methods Constrained Induction of Polynomial Equations for Regression (CIPER) and New CIPER are employed for heuristic search of the best polynomial for a given decision table. Although polynomial functions outperform QQ, they usually fail to provide full ranking of options.

The third method redefines the option ranking problem as constraint optimization problem, and as such, investigates the usage of linear programming for defining its solution. This intuitive approach mainly leads to overly stringent constraints that rarely form a feasible region for solutions.

The fourth research approach, which is the main focus of the thesis, changes the view of the decision tables from deterministic to stochastic. In this approach [5], the attributes are considered as random variables. Copulas are functions which connect marginal distributions of random variables and their joint distribution. The copula function is highly sensitive to small variations of input variables, thus providing distinct results for cases where linear regression used in QQ fails. One-parametric multivariate copulas are used for evaluation of symmetric decision tables, and fully nested Archimedean construction (FNAC) and partially nested Archimedean construction (PNAC), for nonsymmetric decision tables. For the use copulas, the thesis presents new quantile regression equations for different position of the dependent variable in the FNAC and PNAC. The results from the real

Extensive numerical experiments for evaluation of the performance and applicability of the proposed methods were conducted which confirmed the usefulness of the methods, in particular the usage of copula-based method for ranking non-linear decision tables. Finally, the copula-based methods were successfully applied to two real-world cases: ranking of EC motors [6] and ranking of workflows.

3 Conclusion

The newly developed decision support methods for qualitative option evaluation and ranking within classes can be associated with three most relevant results. Firstly, the used approaches extend the space of solvable monotone and linear decision tables to the space of general discrete decision tables. Secondly, methods bridge the gap between qualitative and quantitative models in terms of improving qualitative methods' low sensitivity and limited applicability for the ranking of options within classes. Finally, the methods are applicable for ranking of qualitative options specified with non-linear and or non-monotone decision tables.

References

[1] B. Mileva Boshkoska (2013) *From qualitative to quantitative evaluation methods in multi-criteria decision models*, PhD Thesis, IPS Jožef Stefan, Ljubljana, Slovenia.

[2] M. Bohanec and V. Rajkovič (1990) DEX: An expert system shell for decision support.*Sistemica*, pp.145-157.

[3] M. Bohanec and V. Rajkovič and I. Bratko and B. Zupan and M. Žnidaršič (2012) DEX methodology: Thirty three years of qualitative multi-attribute modelling, *Proceedings of the 15th International Conference Information Society IS 2012*, pp. 31-34.

[4] M. Bohanec (2006) *Odločanje in modeli*, DMFA.

[5] B. Mileva Boshkoska and M. Bohanec (2012) A method for ranking non-linear qualitative decision preferences using copulas *International Journal of Decision Support System Technology*, pp. 1-17.

[6] B. Mileva Boshkoska, M. Bohanec, P. Boškoski and D. Juričić (2013) Copula-based decision support system for quality ranking in the manufacturing of electronically commutated motors, *Journal of Intelligent Manufacturing*, doi: 10.1007/s10845-013-0781-7

Classifier Generation by Combining Domain Knowledge and Machine Learning

Violeta Mirchevska
Department of Intelligent Systems, Jožef Stefan Institute, Jamova 39, Ljubljana, Slovenia
E-mail: violeta.mircevska@ijs.si

Thesis Summary

Keywords: prior domain knowledge, inductive machine learning, small-dataset learning, ambient intelligence

Received: March 25, 2014

This article presents a summary of the doctoral dissertation of the author, which addresses the task of classifier generation by combining domain knowledge and machine learning.

Povzetek: Prispevek predstavlja povzetek doktorske disertaicje avtorice, ki obravnava naloge kreiranja klasifikatorjev s kombiniranjem domenskega znanja in strojnega učenja.

1 Introduction

The field of machine learning (ML) is concerned with the development of algorithms that enable computer programs to learn and automatically improve with experience [1]. ML algorithms have been successfully applied to a wide variety of domains, such as credit-card fraud detection, book recommendations and creating helicopter control logic. They may automatically extract comprehensive concept models solely from concept examples, finding even patterns that are too subtle to be detected by humans. However, their success greatly depends on the quality and the completeness of the available concept examples.

Despite the exponential growth of digital data, there are still domains for which data is scarce. We assume there are at least two reasons for scarce data: (1) sufficient general-purpose data may be costly or otherwise difficult to obtain, possibly due to great domain variation, and (2) general-purpose data may be inappropriate for some deployments, for example, because they are user-specific. One such domain is fall detection. The available training data in the fall-detection domain partially captures the domain's properties because it is difficult to record fall examples due to ethical issues and injury danger. In addition, for reliable performance, fall-detection classifiers need to be tuned to user-specific data. A classifier which suits a user with diminished motor skills may not be appropriate for a user which regularly exercises in the living room.

When learning from training examples which partially capture the domain properties, the learner may create a classifier from patterns, which although representative of the available examples, are not characteristic for the learned concept [2]. Such classifier would perform poorly in real life because it does not capture the essence of the learned concept. This issue may be partially tackled by introducing domain knowledge (DK) as an additional information source in the learning process. Expert DK complements ML as it may contain patters which are not captured in the available concept examples. An expert may verify a classifier's patterns and/or supplement them with patterns from DK. Therefore, classifier generation by a combination of DK and ML is beneficial in domains with insufficient general-purpose data.

Classifier adaptation to user needs may be performed online after system deployment when real-life user-specific data becomes available. In order to pose minimal burden on the user, we consider obtaining user-specific data through occasionally given user feedback which contains information about false negatives (i.e., the system did not detect the class of interest when there was one) or false positives (i.e., the system detected the class of interest when there was none). Such user feedback may be considered as a reward signal given to the system. Learning from rewards is often applied in sequential decision making domains, where the reward function is considered as the most parsimonious description of a task. Online classifier adaptation may be represented as a sequential decision-making task. Therefore, rewards extracted from user feedback may be used for online classifier adaptation in the cases when user-specific data may be obtained after deployment.

The dissertation [2] proposes a novel method, named CDKML (Combining Domain Knowledge and Machine Learning), for classifier generation in the case of scarce data. It combines DK and ML when learning from insufficient general-purpose data, and leverages user feedback for online classifier adaptation to user needs.

The article is organized as follows. Section 2 gives an overview of the CDKML method. Section 3 summarizes the evaluation results. The dissertation's scientific contributions are outlined in Section 4 together with plans for future work.

2 The CDKML method

The CDKML method [3] is a three-phase approach to learning consisting of initialization, refinement and online adaptation. In the initialization phase, an expert specifies a set of patterns important for distinguishing the concept of interest. The patterns may be extracted from DK or be obtained using interactive data mining. In the refinement phase, an optimization algorithm is used for finding the most suitable general-purpose pattern-parameter values by maximizing the classifier's accuracy on available general-purpose data. In the online adaptation phase, user feedback is used to fine-tune the pattern-parameter values to user needs. The online adaptation problem is formulated as a Markov decision process.

3 Evaluation

CDKML was evaluated in three behavior modeling domains: behavioral cloning, posture recognition and fall detection. It's performance was compared to the performance of five classifiers built using ML algorithms in Weka [4]: SMO, RandomForest, NaiveBayes, JRip and J48.

CDKML's refined classifier showed the best performance in the fall-detection domain where it considerably outperformed all five ML algorithms, the posture-recognition domain followed, while it did not show improvement in comparison to standard ML in the behavioral-cloning domain. We attribute the improvement in performance primarily to the contribution of the expert in CDKML's initialization phase where the expert extracted the classifier's patterns using DK and interactive data mining. The improvement was the most evident in the fall-detection domain where DK provided clear instructions: "If a person is lying or sitting on the ground for a longer period of time then a fall happened". Formulating the patterns for the posture-recognition classifier was, however, not simple. In this case, interactive data mining played an important role, helping the expert to incorporate DK into the classifier. In the behavioral-cloning domain, we did not have available DK.

The evaluation of CDKML's online adaptation phase showed that the proposed online adaptation approach is capable of adjusting the refined classifier to correctly recognize events not present in the available general-purpose examples, making tradeoffs between contradictory user feedback based on the cost of each misclassification.

4 Conclusions

The dissertation addresses the problem of classifier generation from scarce data. It proposes a new, three-phase method, named CDKML, for extraction of reliable classifiers in domains where the training examples partially represent the domain properties, but human experts can contribute with their DK. The main contributions of the dissertation are:

- A novel method, named CDKML, for classifier generation and online adaptation which leverages both ML and DK. The novelty is in the way of integration of three phases: initialization, refinement and online adaptation;
- A novel classifier adaptation based on user feedback using Markov decision processes. This, third phase of the CDKML method, is novel on its own.

CDKML achieved higher accuracy than classical ML algorithms when learning from scarce data by leveraging the available DK and user feedback.

As future work, we plan to examine two CDKML improvements. First, exploitation of DK captured in ontologies needs to be considered. The Web offers huge amounts of unstructured, textual data, and approaches to extracting domain patterns and ontology development from that kind of data are emerging [5]. It would be interesting to research possibilities for automating CDKML's initialization by utilizing DK available on the Web. Second, CDKML's online classifier adaptation relies only on user feedback. However, the more real-life examples of the learned concept become available, the better the capability of ML to induce a reliable concept classifier. CDKML's online adaptation may be accompanied with ML classifier re-induction. A combination of the two classifiers in which the ML classifier's influence on the final classification increases as more data becomes available seems reasonable.

Acknowledgement

The research leading to the dissertation was partially financed by the European Union, European Social Fund.

References

[1] T. M. Mitchell (1997). *Machine Learning*. McGraw-Hill, Inc., New York, NY, USA.

[2] V. Mirchevska (2013) *Behavior Modeling by Combining Machine Learning and Domain Knowledge*, PhD Thesis, IPS Jožef Stefan, Ljubljana, Slovenia.

[3] V. Mirchevska, M. Luštrek, M. Gams (2013) Combining domain knowledge and machine learning for robust fall detection. *Expert Systems*, preprint published online.

[4] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, I. H. Witten (2009) The Weka data mining software: An update. *SIGKDD Explorations Newsletter* 11, pp. 10-18.

[5] B. Dalvi, W. W. Cohen, J. Callan (2012) Collectivly representing semi-structured data from the Web. *Proceedings of the Joint Workshop on Automatic Knowledge Base Construction and Web-Scale Knowledge Extraction*, Association of Computational Linguistics, Stroudsburg, PA, USA, pp. 7-12.

JOŽEF STEFAN INSTITUTE

Jožef Stefan (1835-1893) was one of the most prominent physicists of the 19th century. Born to Slovene parents, he obtained his Ph.D. at Vienna University, where he was later Director of the Physics Institute, Vice-President of the Vienna Academy of Sciences and a member of several scientific institutions in Europe. Stefan explored many areas in hydrodynamics, optics, acoustics, electricity, magnetism and the kinetic theory of gases. Among other things, he originated the law that the total radiation from a black body is proportional to the 4th power of its absolute temperature, known as the Stefan–Boltzmann law.

The Jožef Stefan Institute (JSI) is the leading independent scientific research institution in Slovenia, covering a broad spectrum of fundamental and applied research in the fields of physics, chemistry and biochemistry, electronics and information science, nuclear science technology, energy research and environmental science.

The Jožef Stefan Institute (JSI) is a research organisation for pure and applied research in the natural sciences and technology. Both are closely interconnected in research departments composed of different task teams. Emphasis in basic research is given to the development and education of young scientists, while applied research and development serve for the transfer of advanced knowledge, contributing to the development of the national economy and society in general.

At present the Institute, with a total of about 900 staff, has 700 researchers, about 250 of whom are postgraduates, around 500 of whom have doctorates (Ph.D.), and around 200 of whom have permanent professorships or temporary teaching assignments at the Universities.

In view of its activities and status, the JSI plays the role of a national institute, complementing the role of the universities and bridging the gap between basic science and applications.

Research at the JSI includes the following major fields: physics; chemistry; electronics, informatics and computer sciences; biochemistry; ecology; reactor technology; applied mathematics. Most of the activities are more or less closely connected to information sciences, in particular computer sciences, artificial intelligence, language and speech technologies, computer-aided design, computer architectures, biocybernetics and robotics, computer automation and control, professional electronics, digital communications and networks, and applied mathematics.

The Institute is located in Ljubljana, the capital of the independent state of **Slove**nia (or S♡nia). The capital today is considered a crossroad between East, West and Mediterranean Europe, offering excellent productive capabilities and solid business opportunities, with strong international connections. Ljubljana is connected to important centers such as Prague, Budapest, Vienna, Zagreb, Milan, Rome, Monaco, Nice, Bern and Munich, all within a radius of 600 km.

From the Jožef Stefan Institute, the Technology park “Ljubljana” has been proposed as part of the national strategy for technological development to foster synergies between research and industry, to promote joint ventures between university bodies, research institutes and innovative industry, to act as an incubator for high-tech initiatives and to accelerate the development cycle of innovative products.

Part of the Institute was reorganized into several high-tech units supported by and connected within the Technology park at the Jožef Stefan Institute, established as the beginning of a regional Technology park "Ljubljana". The project was developed at a particularly historical moment, characterized by the process of state reorganisation, privatisation and private initiative. The national Technology Park is a shareholding company hosting an independent venture-capital institution.

The promoters and operational entities of the project are the Republic of Slovenia, Ministry of Higher Education, Science and Technology and the Jožef Stefan Institute. The framework of the operation also includes the University of Ljubljana, the National Institute of Chemistry, the Institute for Electronics and Vacuum Technology and the Institute for Materials and Construction Research among others. In addition, the project is supported by the Ministry of the Economy, the National Chamber of Economy and the City of Ljubljana.

Jožef Stefan Institute
Jamova 39, 1000 Ljubljana, Slovenia
Tel.:+386 1 4773 900, Fax.:+386 1 251 93 85
WWW: http://www.ijs.si
E-mail: matjaz.gams@ijs.si
Public relations: Polona Strnad

INFORMATICA

AN INTERNATIONAL JOURNAL OF COMPUTING AND INFORMATICS

INVITATION, COOPERATION

Submissions and Refereeing

Please submit a manuscript at: http://www.informatica.si/Editors/PaperUpload.asp. At least two referees outside the author's country will examine it, and they are invited to make as many remarks as possible from typing errors to global philosophical disagreements. The chosen editor will send the author the obtained reviews. If the paper is accepted, the editor will also send an email to the managing editor. The executive board will inform the author that the paper has been accepted, and the author will send the paper to the managing editor. The paper will be published within one year of receipt of email with the text in Informatica MS Word format or Informatica LaTeX format and figures in .eps format. Style and examples of papers can be obtained from http://www.informatica.si. Opinions, news, calls for conferences, calls for papers, etc. should be sent directly to the managing editor.

QUESTIONNAIRE

□ Send Informatica free of charge

□ Yes, we subscribe

Please, complete the order form and send it to Dr. Drago Torkar, Informatica, Institut Jožef Stefan, Jamova 39, 1000 Ljubljana, Slovenia. E-mail: drago.torkar@ijs.si

Since 1977, Informatica has been a major Slovenian scientific journal of computing and informatics, including telecommunications, automation and other related areas. In its 16th year (more than twenty years ago) it became truly international, although it still remains connected to Central Europe. The basic aim of Informatica is to impose intellectual values (science, engineering) in a distributed organisation.

Informatica is a journal primarily covering intelligent systems in the European computer science, informatics and cognitive community; scientific and educational as well as technical, commercial and industrial. Its basic aim is to enhance communications between different European structures on the basis of equal rights and international refereeing. It publishes scientific papers accepted by at least two referees outside the author's country. In addition, it contains information about conferences, opinions, critical examinations of existing publications and news. Finally, major practical achievements and innovations in the computer and information industry are presented through commercial publications as well as through independent evaluations.

Editing and refereeing are distributed. Each editor can conduct the refereeing process by appointing two new referees or referees from the Board of Referees or Editorial Board. Referees should not be from the author's country. If new referees are appointed, their names will appear in the Refereeing Board.

Informatica is free of charge for major scientific, educational and governmental institutions. Others should subscribe (see the last page of Informatica).

ORDER FORM – INFORMATICA

Name: ....

Title and Profession (optional): ....

....

Home Address and Telephone (optional): ....

....

Office Address and Telephone (optional): ....

....

E-mail Address (optional): ....

Signature and Date: ....

Informatica WWW:

http://www.informatica.si/

Referees from 2008 on:

A. Abraham, S. Abraham, R. Accornero, A. Adhikari, R. Ahmad, G. Alvarez, N. Anciaux, R. Arora, I. Awan, J. Azimi, C. Badica, Z. Balogh, S. Banerjee, G. Barbier, A. Baruzzo, B. Batagelj, T. Beaubouef, N. Beaulieu, M. ter Beek, P. Bellavista, K. Bilal, S. Bishop, J. Bodlaj, M. Bohanec, D. Bolme, Z. Bonikowski, B. Bošković, M. Botta, P. Brazdil, J. Brest, J. Brichau, A. Brodnik, D. Brown, I. Bruha, M. Bruynooghe, W. Buntine, D.D. Burdescu, J. Buys, X. Cai, Y. Cai, J.C. Cano, T. Cao, J.-V. Capella-Hernández, N. Carver, M. Cavazza, R. Ceylan, A. Chebotko, I. Chekalov, J. Chen, L.-M. Cheng, G. Chiola, Y.-C. Chiou, I. Chorbev, S.R. Choudhary, S.S.M. Chow, K.R. Chowdhury, V. Christlein, W. Chu, L. Chung, M. Ciglarič, J.-N. Colin, V. Cortellessa, J. Cui, P. Cui, Z. Cui, D. Cutting, A. Cuzzocrea, V. Cvjetkovic, J. Cypryjanski, L. Čehovin, D. Čerepnalkoski, I. Ćosić, G. Daniele, G. Danoy, M. Dash, S. Datt, A. Datta, M.-Y. Day, F. Debili, C.J. Debono, J. Dedič, P. Degano, A. Dekdouk, H. Demirel, B. Demoen, S. Dendamrongvit, T. Deng, A. Derezinska, J. Dezert, G. Dias, I. Dimitrovski, S. Dobrišek, Q. Dou, J. Doumen, E. Dovgan, B. Dragovich, D. Drajic, O. Drbohlav, M. Drole, J. Dujmović, O. Ebers, J. Eder, S. Elaluf-Calderwood, E. Engström, U. riza Erturk, A. Farago, C. Fei, L. Feng, Y.X. Feng, B. Filipič, I. Fister, I. Fister Jr., D. Fišer, A. Flores, V.A. Fomichov, S. Forli, A. Freitas, J. Fridrich, S. Friedman, C. Fu, X. Fu, T. Fujimoto, G. Fung, S. Gabrielli, D. Galindo, A. Gambarara, M. Gams, M. Ganzha, J. Garbajosa, R. Gennari, G. Georgeson, N. Gligorić, S. Goel, G.H. Gonnet, D.S. Goodsell, S. Gordillo, J. Gore, M. Grčar, M. Grgurović, D. Grosse, Z.-H. Guan, D. Gubiani, M. Guid, C. Guo, B. Gupta, M. Gusev, M. Hahsler, Z. Haiping, A. Hameed, C. Hamzaçebi, Q.-L. Han, H. Hanping, T. Härder, J.N. Hatzopoulos, S. Hazelhurst, K. Hempstalk, J.M.G. Hidalgo, J. Hodgson, M. Holbl, M.P. Hong, G. Howells, M. Hu, J. Hyvärinen, D. Ienco, B. Ionescu, R. Irfan, N. Jaisankar, D. Jakobović, K. Jassem, I. Jawhar, Y. Jia, T. Jin, I. Jureta, Đ. Juričić, S. K, S. Kalajdziski, Y. Kalantidis, B. Kaluža, D. Kanellopoulos, R. Kapoor, D. Karapetyan, A. Kassler, D.S. Katz, A. Kaveh, S.U. Khan, M. Khattak, V. Khomenko, E.S. Khorasani, I. Kitanovski, D. Kocev, J. Kocijan, J. Kollár, A. Kontostathis, P. Korošec, A. Koschmider, D. Košir, J. Kovač, A. Krajnc, M. Krevs, J. Krogstie, P. Krsek, M. Kubat, M. Kukar, A. Kulis, A.P.S. Kumar, H. Kwaśnicka, W.K. Lai, C.-S. Laih, K.-Y. Lam, N. Landwehr, J. Lanir, A. Lavrov, M. Layouni, G. Leban, A. Lee, Y.-C. Lee, U. Legat, A. Leonardis, G. Li, G.-Z. Li, J. Li, X. Li, X. Li, Y. Li, Y. Li, S. Lian, L. Liao, C. Lim, J.-C. Lin, H. Liu, J. Liu, P. Liu, X. Liu, X. Liu, F. Logist, S. Loskovska, H. Lu, Z. Lu, X. Luo, M. Luštrek, I.V. Lyustig, S.A. Madani, M. Mahoney, S.U.R. Malik, Y. Marinakis, D. Marinčič, J. Marques-Silva, A. Martin, D. Marwede, M. Matijašević, T. Matsui, L. McMillan, A. McPherson, A. McPherson, Z. Meng, M.C. Mihaescu, V. Milea, N. Min-Allah, E. Minisci, V. Mišić, A.-H. Mogos, P. Mohapatra, D.D. Monica, A. Montanari, A. Moroni, J. Mosegaard, M. Moškon, L. de M. Mourelle, H. Moustafa, M. Možina, M. Mrak, Y. Mu, J. Mula, D. Nagamalai, M. Di Natale, A. Navarra, P. Navrat, N. Nedjah, R. Nejabati, W. Ng, Z. Ni, E.S. Nielsen, O. Nouali, F. Novak, B. Novikov, P. Nurmi, D. Obrul, B. Oliboni, X. Pan, M. Pančur, W. Pang, G. Papa, M. Paprzycki, M. Paralič, B.-K. Park, P. Patel, T.B. Pedersen, Z. Peng, R.G. Pensa, J. Perš, D. Petcu, B. Petelin, M. Petkovšek, D. Pevec, M. Pičulin, R. Piltaver, E. Pirogova, V. Podpečan, M. Polo, V. Pomponiu, E. Popescu, D. Poshyvanyk, B. Potočnik, R.J. Povinelli, S.R.M. Prasanna, K. Pripužić, G. Puppis, H. Qian, Y. Qian, L. Qiao, C. Qin, J. Que, J.-J. Quisquater, C. Rafe, S. Rahimi, V. Rajkovič, D. Raković, J. Ramaekers, J. Ramon, R. Ravnik, Y. Reddy, W. Reimche, H. Rezankova, D. Rispoli, B. Ristevski, B. Robič, J.A. Rodriguez-Aguilar, P. Rohatgi, W. Rossak, I. Rožanc, J. Rupnik, S.B. Sadkhan, K. Saeed, M. Saeki, K.S.M. Sahari, C. Sakharwade, E. Sakkopoulos, P. Sala, M.H. Samadzadeh, J.S. Sandhu, P. Scaglioso, V. Schau, W. Schempp, J. Seberry, A. Senanayake, M. Senobari, T.C. Seong, S. Shamala, c. shi, Z. Shi, L. Shiguo, N. Shilov, Z.-E.H. Slimane, F. Smith, H. Sneed, P. Sokolowski, T. Song, A. Soppera, A. Sorniotti, M. Stajdohar, L. Stanescu, D. Strnad, X. Sun, L. Šajn, R. Šenkeřík, M.R. Šikonja, J. Šilc, I. Škrjanc, T. Štajner, B. Šter, V. Štruc, H. Takizawa, C. Talcott, N. Tomasev, D. Torkar, S. Torrente, M. Trampuš, C. Tranoris, K. Trojacanec, M. Tschierschke, F. De Turck, J. Twycross, N. Tziritas, W. Vanhoof, P. Vateekul, L.A. Vese, A. Visconti, B. Vlaovič, V. Vojisavljević, M. Vozalis, P. Vračar, V. Vranić, C.-H. Wang, H. Wang, H. Wang, H. Wang, S. Wang, X.-F. Wang, X. Wang, Y. Wang, A. Wasilewska, S. Wenzel, V. Wickramasinghe, J. Wong, S. Wrobel, K. Wrona, B. Wu, L. Xiang, Y. Xiang, D. Xiao, F. Xie, L. Xie, Z. Xing, H. Yang, X. Yang, N.Y. Yen, C. Yong-Sheng, J.J. You, G. Yu, X. Zabulis, A. Zainal, A. Zamuda, M. Zand, Z. Zhang, Z. Zhao, D. Zheng, J. Zheng, X. Zheng, Z.-H. Zhou, F. Zhuang, A. Zimmermann, M.J. Zuo, B. Zupan, M. Zuqiang, B. Žalik, J. Žižka,

Informatica

An International Journal of Computing and Informatics

Web edition of Informatica may be accessed at: http://www.informatica.si.

Subscription Information Informatica (ISSN 0350-5596) is published four times a year in Spring, Summer, Autumn, and Winter (4 issues per year) by the Slovene Society Informatika, Litostrojska cesta 54, 1000 Ljubljana, Slovenia.
The subscription rate for 2014 (Volume 38) is
– 60 EUR for institutions,
– 30 EUR for individuals, and
– 15 EUR for students
Claims for missing issues will be honored free of charge within six months after the publication date of the issue.

Typesetting: Borut Žnidar.
Printing: ABO grafika d.o.o., Ob železnici 16, 1000 Ljubljana.

Orders may be placed by email (drago.torkar@ijs.si), telephone (+386 1 477 3900) or fax (+386 1 251 93 85). The payment should be made to our bank account no.: 02083-0013014662 at NLB d.d., 1520 Ljubljana, Trg republike 2, Slovenija, IBAN no.: SI56020830013014662, SWIFT Code: LJBASI2X.

Informatica is published by Slovene Society Informatika (president Niko Schlamberger) in cooperation with the following societies (and contact persons):
Robotics Society of Slovenia (Jadran Lenarčič)
Slovene Society for Pattern Recognition (Janez Perš)
Slovenian Artificial Intelligence Society (Dunja Mladenić)
Cognitive Science Society (Urban Kordeš)
Slovenian Society of Mathematicians, Physicists and Astronomers (Andrej Likar)
Automatic Control Society of Slovenia (Sašo Blažič)
Slovenian Association of Technical and Natural Sciences / Engineering Academy of Slovenia (Vojteh Leskovšek)
ACM Slovenia (Andrej Brodnik)

Informatica is financially supported by the Slovenian research agency from the Call for co-financing of scientific periodical publications.

Informatica is surveyed by: ACM Digital Library, Citeseer, COBISS, Compendex, Computer & Information Systems Abstracts, Computer Database, Computer Science Index, Current Mathematical Publications, DBLP Computer Science Bibliography, Directory of Open Access Journals, InfoTrac OneFile, Inspec, Linguistic and Language Behaviour Abstracts, Mathematical Reviews, MatSciNet, MatSci on SilverPlatter, Scopus, Zentralblatt Math

Volume 38 Number 1 March 2014 ISSN 0350-5596

Informatica

An International Journal of Computing and Informatics

Editors's Introduction to the Special Issue on "Advances in Semantic Information Retrieval"	V. Klyuev, M. Mozgovoy	**1**
Automatic Detection of Antisocial Behaviour in Texts	M. Munezero, C.S. Montero, T. Kakkonen, E. Sutinen, M. Mozgovoy, V. Klyuev	**3**
Leveraging User Experience through Input Style Transformation to Improve Access to Music Search Services	M. Purgina, A. Kuznetsov, E. Pyshkin	**11**
User Annotations as a Context for Related Document Search on the Web and Digital Libraries	J. Ševcech, R. Mŕo, M. Holub, M. Bieliková	**21**
SOAROAD: An Ontology of Architectural Decisions Supporting Assessment of Service Oriented Architectures	P. Szwed, P. Skrzyński, G. Rogus, J. Werewka	**31**
End of Special Issue / Start of normal papers		
Artificial Immune Based Cryptography Optimization Algorithm	X. Zhou, K. Liu, Z. Jin, S. Tian, Y. Fu, L. Qin	**43**
Bilinear Grid Search Strategy Based Support Vector Machines Learning Method	L. Lin, Z. Xiaolong, Z. Kai, L. Jun	**51**
An Imperative of a Poorly Recognized Existential Risk: Early Socialization of Smart Young Generation in Information Society	V.A. Fomichov, O.S. Fomichova	**59**
Semantic Searching of Biological Documents Using Gene Ontology	M. Mostafa, E.M.F. El Houby, A. Salah	**71**
Tricorder: Consumer Medical Device for Discovering Common Medical Conditions	M. Somrak, M. Luštrek, J. Šušteršič, T. Krivc, A. Mlinar, T. Travnik, L. Stepan, M. Mavsar, M. Gams	**81**
From Qualitative to Quantitative Evaluation Methods in Multi-criteria Decision Models	B.M. Boshkoska	**89**
Classifier Generation by Combining Domain Knowledge and Machine Learning	V. Mirchevska	**91**

Informatica **38** (2014) Number 1, pp. 1–93