Integrated Medical Systems for Improvement of Consultations Between Physicians

Conference Paper · September 2010

CITATIONS
0

READS
475

3 authors, including:

Drashko Nakikj
Columbia University
12 PUBLICATIONS **125** CITATIONS
SEE PROFILE

Vladimir Trajkovik
Ss. Cyril and Methodius University in Skopje
282 PUBLICATIONS **1,647** CITATIONS
SEE PROFILE

M. Gusev (Editor)

ICT Innovations 2010, Web Proceedings

ISSN 1857-7288

M. Gusev (Editor)

ICT Innovations 2010, Web Proceedings

ISSN 1857-7288

On line edition published on http://ictinnovation.org, 2010

Editor: Marjan Gusev
Technical support: Pance Ribarski, Sasko Ristov, Kiril Kiroski
Graphic Design: Innovation, LTD
Illustrator: Alen Aleksovski

Preface

Macedonian Society on Information and Communication Technologies (ICT-ACT) organizes the "ICT Innovations" conference from 2009. It is assumed to be a platform for academics, professionals and practitioners to interact and share ideas, knowledge and expertise in their innovative fundamental and applied research in ICT.

ICT Innovations 2010 Conference was held in Ohrid, Macedonia, on September 12-15, 2010. 70 presentations were given at the conference focused on issues concerning a variety of ICT fields organized in the following 6 sessions:

- Internet applications and Services
- Artificial intelligence and Bioinformatics
- Internet, Mobile and Wireless Technologies
- Multimedia Information Systems
- Computer Networks and Systems
- Computer Security Systems.

Mr. Ivo Ivanovski, Minister of Information Society, and Prof. Velimir Stojkovski, Rector of the Ss. Cyril and Methodius University in Skopje, addressed the participants at the opening session.

155 authors from 13 countries submitted 104 papers for the conference. 33 papers were selected to be published in the Springer Verlag printed edition in Communication in Computer and Information Science via extensive reviewing process. This edition of web proceedings contains 26 regular papers and 20 posters presented on the conference. The reviewing process was realized by 99 reviewers. The average was 4,5 reviews for each submitted paper. The selection of papers was realized by a program committee of 73 members, 28 of which from the Republic of Macedonia.

Ohrid,
September 2010

Conference chairmen
Marjan Gusev

Program committee

Ackovska Nevena, UKIM, Macedonia
Amata Garito Maria, Uninettuno - International Telematic University, Italy
Andova Suzana, TUE, Netherlands
Andonovic Ivan, University of Strathclyde, UK
Antovski Ljupco, UKIM, Macedonia
Atanassov Emanouil, IPP BAS, Bulgaria
Bakeva Verica, UKIM, Macedonia
Bosnacki Dragan, TUE, Netherlands
Cakmakov Dusan, UKIM, Macedonia
Celakovski Sasko, ITgma, Macedonia
Chitkuchev T. Lubomir, Boston University, USA
Davcev Danco, UKIM, Macedonia
Dika Zamir, SEEU, Macedonia
Dimitrova Nevenka, Philips Research, USA
Dimov Zoran, Microsoft - Vancouver, Canada
Eleftherakis George, CITY, Greece
Fullana Pere, ESCI - Barcelona, Spain
Furht Borko, Florida Atlantic University, USA
Gavrilovska Liljana, UKIM, Macedonia
Gievska-Krliu Sonja, GWU, USA
Gjorgjevik Dejan, UKIM, Macedonia
Gligoroski Danilo, Univ. Trondheim, Norway
Grünwald Norbert, Hochschule Wismar, Germany
Haak Liane, University Oldenburg, Germany
Hadzi-Velkov Zoran, UKIM, Macedonia
Ivanovic Mirjana, UNS, Serbia
Jonoska Natasha, Univ. of South Florida, USA
Josimovski Saso, UKIM, Macedonia
Junker Horst, IMBC, Germany
Jurca Ioan, UTT, Romania
Kalajdziski Slobodan, UKIM, Macedonia
Kalpic Damir, FER, Croatia
Madevska Ana, UKIM, Macedonia
Markovski Smile, UKIM, Macedonia
Marovic Branko, University of Belgrade, Serbia
Marx Gomez Jorge, Oldenburg University, Germany
Milentijevic Ivan, Univ. Nis, Serbia
Milosavljevic Milan, Singidunum University Belgrade, Serbia
Misev Anastas, UKIM, Macedonia
Mitreski Kosta, UKIM, Macedonia
Mitrevski Pece, UKLO, Macedonia
Mustafa Blerim, MT, Macedonia
Manolopoulos Yannis, Aristotle University, Greece
Nanevski Aleksandar, IMDEA, Spain
Olariu Stephan, ODU, USA
Paris Francois, University of Houston, USA
Parychek Peter, Univ. Donau, Austria
Patel Dilip, LSBU, UK
Patel Shushma, LSBU, UK
Profumo Francesco, Politecnico di Torino, Italy
Pudlowski Zenon J., WIETE - Melburn", Australia
Radevski Vladimir, NYUS, Macedonia
Sjursen Harold, New York University, USA
Stojanov Georgi, American University of Paris, France
Stojcev Mile, Univ. Nis, Serbia
Tasic Jurij, Univ. of Ljubljana, Slovenia
Tochtermann Klaus, Univ. Graz, Austria
Trajanov Dimitar, UKIM, Macedonia
Trajkovic Ljiljana, SFU, Canada
Trajkovik Vladimir, UKIM, Macedonia
Trajkovski Igor, UKIM, Macedonia
Trichet Francky, Nantes University, France
Vasileska Dragica, ASU, USA
Velinov Goran, UKIM, Macedonia

Contents

Regular papers

Posters

Data Driven Intelligent Systems

Ana Madevska Bogdanova, Nevena Ackovska

Institute of Informatics, Faculty of Natural Sciences and Mathematics,
Arhimedova 5, 1000 Skopje, Macedonia
{ana,nevena}@ii.edu.mk

Abstract. We provide an overview of the current paradigm in Machine Learning (ML) domain. The real data are full of uncertainty and noise and the amount of available data is growing at exponential rate. When one builds applications that deal with the real world problems, the systems that can learn as they obtain new data are needed. These problems use data from different connected sources and the solution needs to be done in real time. We give a brief history of the Machine Learning field and application of the Bayesian framework in the robotics intelligent systems.

Keywords: Machine Learning, Bayesian model, Intelligent Systems, Robotics.

1 Introduction

The current ML methods have to consider the fact that the real data are full of uncertainty, noise, and high dimensionality. Also, today we are dealing with exponential growth in stored data (genomic data, geographical information, social networks, e-health records). In building applications that solve real world problems, the ability to learn as they obtain new data becomes essential. The data sets can be integrated from a cloud of different data sources and the system has to extract the information from each individual source. Systems that use methods that can learn as they navigate the data can be called Data Driven Intelligent Systems. The evolution of the Intelligent Systems (IS) has always followed the technology leading edge of knowledge. Nowadays, computers learn abstractions, patterns, conditional probability distributions, etc.
This paper contains three main parts – the overview of ML evolution, the Bayes framework as an essential necessity for building Intelligent Systems with huge amount of noise data and the use of this framework in several very important robotics problems.

2 The overview of the paradigms in building Intelligent Systems

The Machine Learning paradigm began its development in the 1960' that led to building different knowledge systems. It is the time when pattern recognition emerged

as a new field [1], [2] and the symbolic concept induction was introduced. Latter, in the 1980', the decision trees, rule learning, planning, diagnostics, design and control –a plethora of many research domains emerged, including also learning theory, Neural Network (NN) Learning algorithms, Expert Systems (ES), Genetic Algorithms [3], [4], [5]. The 1990' has brought a maturity of the ML domain – Data mining [6], Support Vector Machine (SVM) and other kernel methods [7], Bayesian networks, Voting, Bagging, Boosting [8], etc.

From this entire asset of methods we pull the following line:

- classic Artificial Intelligence (AI) (1960 – 1980),
- black-box techniques as NN and SVM (1980 - current) – statistical techniques
- Bayesian formulation that changes posterior probabilities as more data are taken into consideration.

The classic AI has had its time with the use of Expert Systems [9] that gave solutions in certain domains, but they weren't general solutions for developing machine intelligence. Usually there is a combinatorial explosion of choices with the ES and it is characterized with man-made rules, according to the experts of the field. Today, some ES are used as tutorial systems [10].

It was a big step to move from the written rules to a knowledge learned from a given data set. The limitations of the ES approach have led to the black box techniques that learn from data (NN, SVM) [4], [7]. These techniques still give excellent results in various problems in Pattern Recognition in different domains, regression, and ranking. Their main limitation is the inability to incorporate the background knowledge in order to solve more accurately the problems that include huge amount of data. Also, the disadvantage of the black-box techniques is the inability to explain the obtained results.

Bayesian theory is used with the Artificial NNs [12], SVMs. The posterior probability, i.e. the values obtained from the output layer neurons in the Neural Networks or the SVM outputs from the probability filter [11], [13], [14], [15] gives the relevance of the recognized pattern presented in probability terms. The filter is actually the sigmoid activation function (Fig. 1) in the ANNs. This way one can use the benefits of the probability theory to describe the uncertainty of the outputs [4].

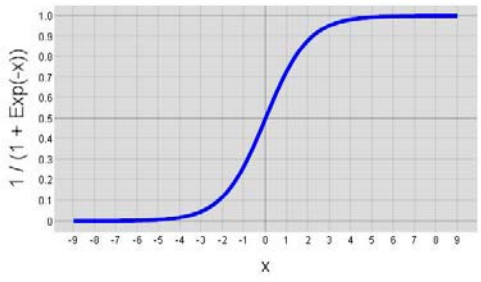


Fig. 1. Sigmoid squashing function 1/(1+Exp(-x)).

The use of the Bayesian framework in the Machine Learning problems is not a new idea. It also gives the opportunity to incorporate the background knowledge into the

system that learns (the prior probability). The latter, of course is not an easy task, but if it is possible, one can give a very accurate estimation of the system outcome (with the posterior probability).

3 Open Machine Learning framework

Over the time, a new paradigm for Machine Intelligence has emerged that brings together the best features of the two previous paradigms - allows prior knowledge from domain experts to be incorporated in the statistical machine learning techniques. With the use of appropriate algorithms, this enables a development of a new generation of large-scale applications.

In this sense, the structured knowledge has to be used – data with recognizable background, i.e. domain knowledge. Unstructured data can be used with the black-box techniques, but it can not work with the cloud computing problems.

To learn from data, one creates the models. In the real world application there isn't analytical expression for the model of the domain.

The Bayesian formulation of the probabilities provides a powerful tool for dealing with the uncertainty. In Machine Learning applications the uncertainty lies in the noisy data (ambiguous, contradictory, missing data) and the unknown model parameters. This uncertainty must be described mathematically in order to measure its quantity. Probability theory is used in order to meet this requirement. It provides consistent framework for quantification and manipulation of uncertainty.

3.1 Bayesian approach in on-line learning

Bayesian approach provides the framework to measure and narrow the uncertainty. For example, in an n-dimensional space, one can represent n quantities, and their values may be unknown. Those values can be described by the probability distributions. So, at the beginning, before any information from the data at hand is presented to the system, the positions of the most probable values of the parameters are known - the prior probability. In the same space of the parameters lives a likelihood function that provides the information where are the values of the parameters that has most likely given a rise to the data and where are the values of the parameters that less likely have generated that data. The Bayesian formulation multiplies the values of the prior probability and the likelihood function and the result is the **posterior probability**, again in the same space. This way a 'squashed' prior distribution is obtained, which means that the uncertainty has been restricted - something is learned. In the paradigm of probability, learning means the reduction of the uncertainty as new data are collected.

As a second step, the obtained posterior probability is the new prior probability when the new data are observed. Again the prior is multiplied with the new likelihood function and the result is the new posterior probability, which has reduced the uncertainty even further – something more is learned.

The learning process includes adjusting the model every time a new data is obtained. The Bayesian approach gives the opportunity to use the learned parameter

values of the system as starting values of the system that compute the information from the new data.

If one wants to make predictions, it shouldn't be taken only the most probable parameter values, but by using the probability theory every parameter value from that space is used. Only, the parameters from the area with higher probabilities have bigger weight, while the parameters with the lower probability posses lesser weight. The computational task is to integrate those parameters. However, the real life applications carry very high dimensionality and they require heavy integration and a lot of computation.

The key computational step is to go through all the parameters in the space. This is how we can reduce the number of computations (example in two-dimensional space):

$$\sum_x \sum_y xy = x_1y_1 + x_2y_1 + x_1y_2 + x_2y_2 = (x_1 + x_2)(y_1 + y_2) \ . \quad \textbf{(1)}$$

In order to describe the model of the process, we need the probability distribution. With the real world problems, the dimensionality of its parameters can be thousands or millions and there is no analytical expression to model the intractable distribution, so some approximations should be done. Until recently there was only one, general purpose technique for calculating the distributions, called Monte Carlo (MC) chain methods [16], [17]. The main idea is to draw a lot of samples at random from this distribution, so where the distribution is high, there would be a lot of samples, and where the distribution is low – less samples. In practice, MC methods are computationally very intensive. So they have been used on problems that scale from few hundred to few thousand points. With some new deterministic methods, instead of drawing random samples, we can work with approximations from family of distributions, i.e. the real distribution curve might be approximated by simpler family of distributions. Depending on the used family of distributions, and the meaning of the expression 'best fit', there are several algorithms (Variational Bayes [18], [20], Loopy Belief Propagation [19], Expectation Propagation [21]). This approach has allowed developing ML that can deal with huge amount of data with high dimensionality, in real time [22].

3.2 Bayesian learning - definition

The idea behind the Bayesian learning [23], [22] is the following. If one sets a model with data represented with the vector **w**, which has certain number of parameters (the dimensionality of the feature space), and denote the training set by *D*, then the conditional probability distribution is defined as a $p(D|\mathbf{w})$. As a function of **w**, it is known as a likelihood function. In a Bayesian framework the uncertainty in the vector **w** is expressed through a probability distribution $p(\mathbf{w})$, known as the prior distribution. This is problem related and it captures the background knowledge about the vector **w**.

The information from the training data is expressed through the likelihood function. Combined with the prior distribution, we obtain the posterior distribution, using Bayes' theorem

$$p(\mathbf{w} / D) = \frac{p(D / w)p(w)}{p(D)} . \quad (2)$$

As we mentioned, in the Bayesian framework we don't get the estimations of the point values (the parameters), but we consider their distributions, i.e. we consider the distributions over the variables. The Gaussian distribution serves well, in the sense that it allows possibility of a certain feature to have values around the main value (Fig.2). The Gaussian distribution is narrowing as more information (data) concerning the values of that feature is obtained.

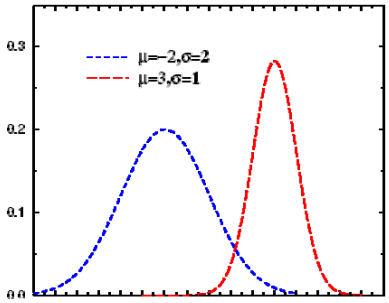


Fig. 2. Two Gaussians. The left one ($\mu = -2, \sigma = 2$), represents a distribution of a parameter that hasn't been inspected yet, whiles the other Gaussian with $\mu = 3, \sigma = 1$ is a distribution of a parameter that has 'learned' and the choice of its possible values is narrowed.

4 Bayesian framework in modeling robotics systems

Robotics is one very important area of Intelligent Systems research. It embraces a vast area of today's technological and modeling state of the art paradigms. This area shows great evolution, from mechanical beings controlled by humans [35], to autonomous robots, and robots constructed by following natures buildings mechanisms [34], [36], [37]. Since the robotics has grown far away from merely creating mechanical services for humans and has entered the realm of sensory perception and vast data computation, new algorithms for handling these data were needed. However, the Bayesian methods have proven to work well for most of the major robotics problems. All of them build upon the framework described above – learning is obtained by reintroducing the posterior probability from a specific iteration, as prior to the next.

These problems are far from trivial. They include robot perception, localization, mapping, exploring, movement tracking [39] and many more. Shortly we will provide an overview of two of these important and difficult for handling problems – localization and mapping.

Localization makes the robots autonomous. If a robot does not know its location, it can be difficult to determine what to do next. In order to localize itself, a robot has access to some measurements, which give the robot feedback about its moving actions and the situation of the environment around the robot. Given this information, the robot has to determine its location as accurately as possible. What makes this difficult is the existence of uncertainty in both the movement and the sensing of the robot. This uncertain information needs to be combined in an optimal way, so the errors in localization become minimal.

However, it has been shown that the current probabilistic methods for position estimation still face considerable problems. These problems are closely related to the type of representation used to represent probability densities over the robot's state space. Many local approaches aim at tracking the position of a robot once its starting location is known. These approaches usually use Kalman filters [33], [32] to integrate sensor information over time. Existing approaches of this class have shown to be efficient and accurate, but due to the assumptions underlying these methods, they typically are not able to localize the robot globally. For global localization, several researchers proposed a new localization paradigm, called Markov Localization [32]. This technique uses a richer representation for the state space of the robot and therefore is able to localize a robot from scratch, i.e. without knowledge of its starting location.

Building maps for mobile robots is one of the fundamental robotics problems. A decade ago it has been proven that building maps when the robot’s location is known is relatively easy task [27]. Also, if a map already exists, the localization of a robot is proven to be successful [28].The problem of mapping while moving is often referred to as the simultaneous mapping and localization problem. Most of the approaches to solve this important problem build maps incrementally, by iterating localization and incremental mapping for each new sensor scan the robot receives [29], [30]. These methods work well in real time, but they usually fail while mapping large cyclic environments. The reason for this is the robot’s cumulative error. In such environments it can grow without bounds, and when closing the cycle error has to be corrected backwards in time (which most existing methods are incapable of doing). However, recent studies have shown that this problem gives good solutions when using likelihood estimators [31]. When closing a cycle in cyclic environments, the global uncertainty is getting smaller if the existence of the cycle is recognized by the robot (this is shown in Fig. 3).

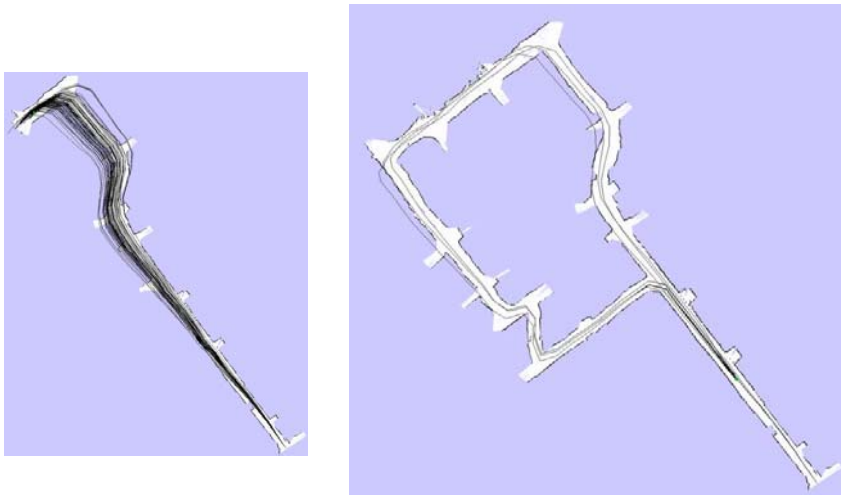


Fig. 3. The uncertainty grows as the robot advances in building the map. However, once a cycle has been recognized, the global uncertainty is minimal [38] .

The above explained Bayesian line of reasoning applies also in the multi-robot environments [31], robot exploration [32] and many other robotics related problems. However it has been modified to serve better with the large amount of real time data that comes from the ongoing sensing in the modern robotic systems.

5 Conclusion

The main feature of solving the real life problems is allowing the manipulation of data as new information arrives. The outcomes vary depending of the new obtained knowledge from the new information and lessen the uncertainty of the problem. The Bayesian approach enables the computational machine to solve the problems in real time, which is not possible with traditional Monte Carlo methods because of the complex computation.

The integration of the real data from different on-line sources can give more and more accurate classification systems, which accuracy grows with the amount of the new arrived data. By using the right algorithms in the Bayesian framework, one can take advantage of the huge amount of data in real time to test and refine the system in construction.

There are many examples of applications that are build on this paradigm: ranking system in on-line games, which can very quickly estimate the game player skills in order to match her/him with other player with similar skills [24], recognition of network attacks, e-medical charts, and many more [25], [26].

The robotics deal with uncertainty at all times. There are robots with many sensors, environments with dynamic obstacles (like moving people in crowded museums),

exploration of unknown spaces etc. All of these problems deal with enormous quantity of data that should be processed in real time. The Bayesian framework has proven to work well, although on all the levels it has been upgraded with novel methods [32]. However, the basic of the learning framework we were discussing in this paper has remained. It is obtained by reintroducing the posterior probability from a specific iteration, as prior to the next.

References

1. Koutroumbas, K., Theodoridis, S.: Pattern Recognition (4th ed.). Boston: Academic Press. ISBN 1-59749-272-8 (2008)
2. Fukunaga, K.: Introduction to Statistical Pattern Recognition (Second ed.). Academic Press(1990)
3. Kulikowski, C. A., Weiss, M.:. Computer Systems That Learn: Classification and Prediction Methods from Statistics, Neural Nets, Machine Learning, and Expert Systems. Machine Learning. San Francisco. Morgan Kaufmann Publishers. ISBN 1-55860-065-5. (1991)
4. Bishop, C.M.: Neural Networks for Pattern Recognition. Oxford University Press, (1996)
5. Darlington, K.:. The Essence of Expert Systems. Pearson Education. ISBN 0-13-022774-9, (2000)
6. Han, J.,Kamber, M.: Data Mining: Concepts and Techniques. Morgan Kaufmann, (2001)
7. Burges, C. J. C.: A tutorial on support vector machines for pattern recognition. In: KnowledgeDiscovery and Data Mining **2**(2), 121–167. (1998).
8. Bühlmann, P. Hothorn, T.: Boosting algorithms: regularization, prediction and model fitting. In: Statistical Science, 22(4):477-505, (2007)
9. Giarratano, J.C., Riley, G.: Expert Systems: Principles and Programming, (Fourth Edition), PWS Publishing Company, (1998)
10. Madevska Bogdanova, A., Celikik M., Erakovik, Z.: Expert System for Taxonomy in Entomology. In: 10th World Multiconference on Systemic, Cybernetics and Informatics, Orlando, USA (2006)
11. Kwok, J.: Moderating the outputs of support vector machine classifiers. In: IEEE Transactions on Neural Networks 10(5): 1018-1031 (1999)
12. MacKay, D. J. C.: Bayesian methods for backprop networks. In: E. Domany, J. L. van Hemmen, and K. Schulten (eds.), Models of Neural Networks, III, Chapter 6, pp. 211–254. Springer (1994)
13. Platt, J. C.: Probabilities for SV machines. In: A. J. Smola, P. L. Bartlett, B. Scholkopf, and
D. Shuurmans (eds.), Advances in Large Margin Classifiers, pp. 61–73. MIT Press (2000)
14. Madevska-Bogdanova A., Nikolic D., Curfs L.: Probabilistic SVM outputs for pattern recognition using Analytical Geometry. In: NEUROCOMPUTING, An International Journal, Elsevier 2004; Vol. 62C, 293-303, (2004)
15. Madevska, A., Nikolic, D.: A New Approach of Modifying the SVM output. In: IEEE-INNS-ENNS International Joint Conference on Neural Networks, IJCNN'2000, Italy (2000)
16. Robert, C. P.,Casella, G.: Monte Carlo Statistical Methods (2nd ed.). Springer, New York, ISBN 0387212396 (2004)
17. Rubinstein, R. Y.,Kroese, D. P.: Simulation and the Monte Carlo Method (2nd ed.). John Wiley & Sons, New York. ISBN 9780470177938 (2007)
18. Ishikawa, Y., Takeuchi, I., Nakano, R.: Variational Bayes from the primitive initial point for Gaussian mixture estimation. Computer Science, 5863 LNCS (PART 1), pp. 159-166 (2009)

19. Shachter, R. D. Peot , M.:. Simulation approaches to general probabilistic inference on belief networks. In P. P. Bonissone, M. Henrion, L. N. Kanal, and J. F. Lemmer (eds.), Uncertainty in Artificial Intelligence, Volume 5. Elsevier (1990)
20. May, P., Ehrlich, H.C., Steinke, T.: ZIB Structure Prediction Pipeline: Composing a Complex.
21. Biological Workflow through Web Services. In: Nagel, W.E., Walter, W.V., Lehner, W. (eds.) Euro-Par 2006. LNCS, vol. 4128, pp. 1148--1158. Springer, Heidelberg (2006)
22. Vaclav, S., Quinn, A.: The Variational Bayes Method in Signal Processing. Springer, (2005)
23. Minka, T.: Expectation propagation for approximate Bayesian inference. In J. Breese and D. Koller (Eds.), Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence, pp. 362–369. Morgan Kaufmann. (2001).
24. Bishop, C.: Pattern Recognition and Machine Learning, Springer, (2006)
25. Neal, R. M.: Bayesian Learning for Neural Networks. Springer. Lecture Notes in Statistics
118. (1996)
26. Dangauthier, P., Herbrich, R., Minka T., Graepel T.: TrueSkill Through Time: Revisiting the History of Chess. In Advances in Neural Information Processing Systems 20, MIT Press, (2008)
27. Zhang, X., Graepel, T., Herbrich, R.: Bayesian Online Learning for Multi-Label and Multi-Variate Performance Measures. In: Thirteenth Conference on Artificial Intelligence and Statistics AISTATS (2010)
28. D. Fouskakis, I. Ntzoufras, and D. Draper D.: Bayesian variable selection using cost-adjusted BIC, with application to cost-effective measurement of quality of health care. In: Annals of applied Statistic, Volume 3, Number 2, 663-690, (2009).
29. H. P. Moravec and A. Elfes. High resolution maps from wide angle sonar. In ICRA-85, (1985)
30. Borenstein J., Everett B., Feng L.: Navigating Mobile Robots: Systems and Techniques. A. K. Peters, (1996)
31. J.J. Leonard, H.F. Durrant-Whyte, and I.J. Cox. Dynamic map building for an autonomous mobile robot. International Journal of Robotics Research, 11(4), (1992).
32. W.D. Rencken. Concurrent localisation andmap building for mobile robots using ultrasonic sensors. In *IROS-93* (1993)
33. Thrun S., Burgard W., Fox. D.: A Real-Time Algorithm forMobile Robot MappingWith Applications to Multi-Robot and 3D Mapping. In IEEE International Conference on Robotics and Automation, San Francisco, April (2000)
34. Thrun S., Burgard W., Fox. D.: Probabilistic Robotics, MIT Press, 2005
35. Kalman R. E.: A New Approach to Linear Filtering and Prediction Problems, Research Institute for Advanced Study, Baltimore, (1960)
36. Ackovska N. Robotics: From Classical to Bio-nano Robots, RoboMak Magazine 2010 (in print)
37. Bozinovski S., Jovancevski, G., Ackovska, N. Distributed Interactive Robotics Classroom. Optoelectronics Information-Power Technologies 2(4): 5-9 (2002)
38. Morin J.-F., Shirai Y., Tour J. M., “En Route to a Motorized Nanocar”, Organic Letters, Vol. 8, No. 8, pp. 1713-1716 (2006)
39. Ackovska N., Bozinovski S., Jovancevski G.: Real-Time Systems – Biologically Inspired Future, Journal of Computers, Vol 3. No. 3, pp.56-63 (2008)
40. Homepage Sebastian Thrun, http://robots.stanford.edu/videos.html
41. Panangadan A., Mataric M., Sukhatme G. S.: Tracking and Modeling of Human Activity Using Laser Rangefinders, In. International Journal of Social Robotics, 2: 95–107, Springer (2010)

Computational Methods for Gene Finding in Prokaryotes

Mihaela Angelova, Slobodan Kalajdziski, Ljupco Kocarev

Faculty of Electrical Engineering and Information Technologies, Skopje, Macedonia
mihaelaangelova@yahoo.com

Abstract. Gene finding is crucial in understanding the genome of a species. The long genomic sequence is not very useful, unless its biologically functional subsequences (genes) are identified. Along with the ongoing revolution in sequencing technology, the number of sequenced genomes has increased drastically. Therefore, the development of reliable automated techniques for predicting genes has become critical.

Automatic gene prediction is one of the essential issues in bioinformatics. Many approaches have been proposed and a lot of tools have been developed. This paper compiles information about some of the currently most widely used gene finders for prokaryotic genomes, explaining the underlying computational methods and highlighting their advantages and limitations. Finally, the gene finders are tested on a strain with high GC-content.

Keywords: *ab initio* prediction, gene prediction, homology-based search, prokaryotes

1 Introduction

The development of automated sequencing technologies with dramatically lower cost and higher throughput has revolutionized biological research, allowing scientists to decode genomes of many organisms [1]. Prokaryotic genomes are sequenced at an increasing rate. After a genomic sequence is reconstructed from the sequencing data, the next and most important step is to understand the content of the genome i.e. identify the gene loci and their functions. These genes then become the basis for much further biological research.

In the earliest days, genes were identified with experimental validation on living cells and organisms, which is the most reliable method, but costly and labor intensive. Since there can be thousands of genes in one bacterial genome, computational methods are essential for automatic analysis of uncharacterized genomic sequences.

At present, there are many prokaryotic gene finders, based on different approaches. Generally, the gene prediction approaches can be divided into two classes: intrinsic (*ab initio*) and extrinsic (homology-based). It is common for gene finders of both types to be used in a gene finding project, owing to their complementary nature. They all have their own advantages and weaknesses. Therefore, high-quality gene annotation of microbial genomes remains an ongoing challenge.

2 Computational Methods for Gene Prediction

DNA sequences that encode proteins are not a random combination of codons (triplets of adjacent nucleotides designating a specific amino acid). On the contrary, the order of codons obeys certain biological rules and is maintained during evolution. Certain patterns in the codon arrangements have been recognized. For example, conserved preference for certain codon pairs within the coding region is confirmed in the three domains of life [2]. Moreover, the genetic code is used differently in different bacterial species. In most bacteria, the synonymous codon usage (usage of codons that represent the same amino acid) varies not only between organisms [3], but also within an organism, since the horizontally transferred genes tend to have different codon usage from the host. These patterns in the genetic code and their conservancy with evolution can be very helpful for the computational gene finding. They enable algorithms for gene prediction to rely on statistics that describe gene patterns or on sequences' resemblance to conserved annotated proteins.

Two general classes of computational methods are adopted: *ab initio* prediction (intrinsic) and homology-based search (extrinsic) [4]. The first method uses gene structure as a guide to gene detection. The latter one, which is based on the observation that coding sequences are more conserved than the non-coding genes and intergenic regions, compares the genome to the available gene sequences and searches for significant homology.

2.1 *Ab initio* Gene Finders

Ab initio approaches do not use extrinsic information for gene prediction. Instead, they inspect the input sequence and search for traces of gene presence. Intrinsic methods extract information on gene locations using statistical patterns inside and outside gene regions as well as patterns typical of the gene boundaries. There are specific DNA motifs called signals that indicate a neighboring gene, *e.g.* promoters, start and stop codons. Apart from signals, *ab initio* methods discover gene signposts based on content search, looking for patterns of codon usage specific for the organism. Mainly, *ab initio* algorithms implement intelligent methods to represent these patterns as a model of the gene structure in the organism. The most widespread algorithms for gene finding in prokaryotes are based on Markov models and dynamic programming.

Prokaryotic gene features useful for *ab initio* prediction. The simplest *ab initio* method is to inspect Open Reading Frames (ORFs). An ORF is a sequence of bases encompassed by the translation initiation and termination site *i.e.* the coding sequence of prokaryotic organisms [5]. Every coding Deoxyribonucleic Acid (DNA) has six possible reading frames: three on the direct (positive) strand and three in the complementary strand read in the opposite direction of the double-stranded DNA. The nucleotides on the positive strand are grouped into triplets starting from the first (+1), second (+2) or third (+3) nucleotide in the start codon. Therefore, there are three possible reading frames in the positive strand. The same follows for the negative strand, by looking at the sequence from the opposite side. An ORF consists of

consecutive triplets and terminates with the first stop codon it encounters. Typically, only one reading frame, the ORF, is used to translate the gene. Therefore, the prokaryotic gene finder should primarily be able to identify which of the six possible reading frames contains the gene i.e. is an ORF. In general, bacterial genes have long ORFs. This is a hint for gene finding. For example, if the frame +1 has the longest sequence without a stop codon, then its amino acid sequence most probably leads to the gene product.

Nevertheless, this is a good, but not assuring indication for selecting the correct ORF. Not every ORF is a coding region. Telling the difference between genes and random ORFs is the most important goal of the gene finding process [5]. Even if we tune a certain length threshold and define that ORFs longer than that threshold are genes, differentiating between short genes and occasional ORFs remains a problem.

There are some other characteristics of the prokaryotic gene that pose difficulties for the gene finding process. Identifying the right ORFs is deteriorated when two ORFs overlap. Although this is considered to happen rarely in prokaryotes, it is difficult to automatically resolve the problem.

Moreover, there are multiple start codons. In most cases, ATG is the start codon that suggests initiation of translation. Occasionally, GTG and TTG act as initiation sites [6]. Multiple start codons can cause ambiguities, because their presence does not ensure translation initiation.

In conclusion, there is no straightforward way to find genes based on their features. Therefore, *ab initio* gene finders rely not only on signal sensors (start and stop codons, promoters, etc.), but they also use content sensors, such as patterns of codon usage or other statistically inferred features.

Markov Model Based Algorithms. Several highly accurate prokaryotic gene-finding methods are based on Markov model algorithms.

The *GeneMark* family [7] includes two major programs, called GeneMark [8] and GeneMark.hmm [9]. Analysis of DNA from any prokaryotic species without a pre-computed species-specific statistical model is enabled by a self-training program, GeneMarkS [10].

GeneMark uses a Bayesian formalism to assess the *a posteriori* probability that a given short fragment is part of a coding or non-coding region. These calculations are performed using Markov chain models. The idea behind this is that there are specific correlations between adjacent nucleotides in chromosomal DNA sequences. Markov chains have shown to be appropriate in inferring the statistical description of the gene structure.

In mathematical terms, a Markov chain is a discrete random process that evolves through the states from the set $S = \{s_1, s_2,\ldots,s_r\}$. The conditional distribution of any future state depends only on the k preceding variables, for some constant k. In the context of gene prediction, the sequence of random variables $X_1, X_2,\ldots, X_k$ take on values from the set of bases (A, C, G, T) and a Markov chain models the probability that a given base b follows the k bases immediately prior to b in the sequence. Using a training set, a Markov chain captures statistical information about a sequence by computing the probability that a certain nucleotide x_i appears after a sequence s_i e.g. $p(x_i = A \mid s_i = TTGCA)$, k = 5. The three codon positions have different nucleotide frequency statistics. Therefore, in order to model the codon usage, normally the

variables of the Markov chain are sets of three nucleotides (codons) or multiples of codons. For this reason, the orders of the Markov chains, k, used for prediction are 2, 5, 8, and so on. For the purpose of modeling protein-coding regions, GeneMark utilizes a three-periodic inhomogeneous Markov model (transition probabilities change with time), because the DNA composition and features vary among different species [11]. Ordinary (homogenous) Markov models are found to be appropriate for non-coding DNA.

GeneMark is the oldest method based on Markov models. It does not offer high accuracy, because it lacks precision in determining the translation initiation codon [9]. Markov chain model of the DNA sequences is firstly introduced in GeneMark. The initial success of GeneMark has paved the way for further research in this direction.

GeneMark.hmm is designed to improve GeneMark in finding exact gene starts. Therefore, the properties of GeneMark.hmm are complementary to GeneMark. GeneMark.hmm uses GeneMark models of coding and non-coding regions and incorporates them into hidden Markov model framework. In short terms, Hidden Markov Models (HMM) are used to describe the transitions from non-coding to coding regions and *vice versa*. GeneMark.hmm predicts the most likely structure of the genome using the Viterbi algorithm, a dynamic programming algorithm for finding the most likely sequence of hidden states. To further improve the prediction of translation start position, GeneMark.hmm derives a model of the ribosome binding site (6-7 nucleotides preceding the start codon, which are bound by the ribosome when initiating protein translation). This model is used for refinement of the results.

Both GeneMark and GeneMark.hmm detect prokaryotic genes in terms of identifying open reading frames that contain real genes. Moreover, they both use pre-computed species-specific gene models as training data, in order to determine the parameters of the protein-coding and non-coding regions.

Acceleration of microbial genome sequencing has led to the need for non-supervised gene finding methods. *GeneMarkS* combines GeneMark.hmm and GeneMark with a self-training procedure. The main focus of GeneMarkS is detecting the correct translation initiation sites. It creates a statistical model, runs the GeneMark.hmm program, and corrects the model based on the results. The steps are repeated iteratively until convergence.

Glimmer3.0 [12] The core of Glimmer is Interpolated Markov Model (IMM), which can be described as a generalized Markov chain with variable order. After GeneMark introduces the fixed-order Markov chains, Glimmer attempts to find a better approach for modeling the genome content. The motivational fact is that the bigger the order of the Markov chain, the more non-randomness can be described. However, as we move to higher order models, the number of probabilities that we must estimate from the data increases exponentially. The major limitation of the fixed-order Markov chain is that models from higher order require exponentially more training data, which are limited and usually not available for new sequences. However, there are some oligomers from higher order that occur often enough to be extremely useful predictors. For the purpose of using these higher-order statistics, whenever sufficient data is available, Glimmer IMMs.

Glimmer calculates the probabilities for all Markov chains from 0-th order to 8-th.

If there are longer sequences (*e.g.* 8-mers) occurring frequently, IMM makes use of them even when there is insufficient data to train an 8-th order model. Similarly, when the statistics from the 8-th order model do not provide significant information, Glimmer refers to the lower-order models to predict genes.

Opposed to the supervised GeneMark, Glimmer uses the input sequence for training. The ORFs longer than a certain threshold are detected and used for training, because there is high probability that they are genes in prokaryotes. Another training option is to use the sequences with homology to known genes from other organisms, available in public databases. Moreover, the user can decide whether to use long ORFs for training purposes or choose any set of genes to train and build the IMM.

There are many annotation services that incorporate Glimmer or GeneMark in their pipelines such as RAST [13], Maker [14] and JCVI Annotation Service [15].

AMIGene [16]. The reason for including AMIGene in the list of gene finders revised in this paper is that AMIGene can be very helpful in some cases. The interesting thing about AMIGene is that it serves as substitution for manual curation, because it searches the most likely CoDing Sequences (CDSs) in the output of a GeneMark-like program.

AMIGene predicts the genome structure in the same way as GeneMark. In addition to that, AMIGene investigates codon usage patterns and relative synonymous codon usage in the predicted CDSs, using multivariate statistical technique of factorial correspondence analysis (FCA) and k-means clustering. AMIGene uses these results to evaluate and filter predicted genes. The construction of gene classes based on codon usage can uncover small genes, which are difficult to spot using the typical model.

AMIGene is not yet suitable for identifying true translation initiation sites and does not take into account overlaps between adjacent CDSs. Considering these drawbacks and considering that AMIGene predicts only the most likely CDSs, it follows that it is a good idea to use AMIGene in combination with other gene finders.

FGenesB [17] is another Markov chain-based algorithm, claimed to be more accurate than GeneMarkS and Glimmer. Unlike them, it finds tRNA and rRNA genes, in addition to coding sequences. Initial predictions of ORFs are used as training set for 5^{th} order in-frame Markov chains for coding regions, 2^{nd} order Markov models translation and termination sites. FGenesB uses genome-specific parameters, automatically trained using only genomic DNA as an input.

FGenesB annotates the genes i.e. identifies their functions by homology with protein databases. As the rRNA genes are highly conserved with evolution, FGenesB identifies them easily in the genome, by comparing them against bacterial and archaeal rRNA databases, using the Basic Local Alignment Search Tool (BLAST) [18], which is described in the section for homology based search.

In prokaryotic cells, functionally related genes are usually found grouped together in clusters called operons and transcribed as one unit. FGenesB is able to predict operons based on distances between ORFs and frequencies of different genes neighboring each other in known bacterial genomes.

In conclusion, FGenesB integrates model-based gene prediction with homology-based annotation, accompanied by operon, promoter and terminator prediction in

bacterial sequences.

Dynamic programming. Opposing to the other gene finders described so far, *Prodigal* [19] does not rely on the assumption that long ORFs are potential genes with high probability, because it can be misleading for gene prediction in GC-rich organisms. Because the stop triplets (TAA, TGA, TAG) are AT rich, their frequency is lower in organisms with high GC (guanine-cytosine) content. Hence, the probability that long ORFs occur by chance increases proportionally to the GC content [20].

Prodigal self-trains by detailed analysis of the GC frame plot. It calculates the statistical significance of the bases G and C in different frame positions. The GC frame plot consists of three graphs, depicting the GC content of the 1st, 2nd, and 3rd nucleotide from each codon in each open reading frame. In coding DNA, the GC content of the third base (GC3) is often higher in genes, relative to non-coding regions [21]. Based on this, Prodigal builds its gene model, looking for a bias for G or C in the 1st, 2nd and 3rd position of each codon. After determining the potential genes, Prodigal filters them, by examining the translation initiation site, ribosomal binding site (RBS), and the lengths of ORFs. The refined set of genes is used as training data.

Prodigal utilizes the same dynamic programming algorithm both for its preliminary training phase and for its final gene calling phase. It scores each ORF, start-stop pair, some motifs, etc. and uses a dynamic programming procedure to find the optimal pathway among a series of weighted steps.

The disadvantage of Prodigal is that there are some genes such as laterally transferred genes, genes in phage regions, proteins with signal peptides and other that do not match the typical GC frame bias for the organism in question.

In summary, *ab initio* gene finders find most of the genes, but have a significantly bigger number of false positives. At the present, no *ab initio* gene finder is able to clearly distinguish short non-coding ORFs from real genes. Moreover, most gene finders rely on the assumption that long ORFs in prokaryotes are genes, which usually leads to incorrect results in microbes with high GC. This problem is addressed by Prodigal. On the other hand, Glimmer for example, uses long ORFs as its training set. As a consequence of the low frequency of stop codons in GC-rich organisms i.e. increased likelihood of random long ORFs, Glimmer lengthens the genes. Mainly, the predicted genes are longer than the actual ones. Therefore, it is important for gene finders to be GC content indifferent.

Not only the GC content can influence on the accuracy of gene finders, but horizontally transferred DNA sequences can also affect the statistical model. These sequences are not evolutionary connected to the rest of the genome, which is why they differ significantly in the context of codon usage, GC frame bias, etc. The Markov models of GeneMark and Glimmer differentiate between these regions, whereas Prodigal fails to recognize them.

2.2 Homology-Based Search

The lower evolution rate of the coding regions enables that genes are identified by comparison with existing protein sequences. Given a library of sequences of other

organisms, we search the target sequence in this library and identify library sequences (known genes) that resemble the target sequence [22].

Local alignment and global alignment are two methods based on similarity searches. The most common local alignment tool is the BLAST family of programs. BLAST is a widely-used tool for searching similarity by homology-based gene finders. It identifies regions of similarity by first breaking down the query sequence into a series of DNA or protein sequences, and then it searches a local or NCBI database. Once a match is found, it tries to align the two sequences i.e. identify every matching letter, insertion, deletion and substitution.

This evidence-based approach is the most reliable method for gene prediction [20]. It is able to find biologically relevant genes. Moreover, it is based on a very simple concept. This approach helps not only find the gene loci, but also annotate (infer the function of) that region, because homologous sequences are supposed to have the same or similar functions.

The biggest limitation of this approach is that only an insufficient number of genes have significant homology to genes in external databases. This number is even smaller for organisms whose closest relatives are not sequenced, because there are many species-specific genes that are not present in databases.

In conclusion, the most reliable way to identify a gene in a newly sequenced genome is to find a close homolog from another organism. Homology-based search is the simplest and is characterized with high accuracy. However, it requires huge amounts of extrinsic data and finds only half of the genes. Many of the genes still have no significant homology to known genes.

3 Comparison of the Gene Finding Tools

Gene finders may differ on the type of genes they are able to recognize (non-coding RNA or proteins); some of them accept only one genomic sequence as input, whereas others can process multiple sequences; different gene finders may produce output files in different formats. The following table summarizes the features of all the gene finders that are described or mentioned in this paper.

Table 1. Comparison of Some Features for Gene Finders

Gene finders	CDS	stable RNA[a]	FA[b]	developed for :	Output files format
Prodigal	y	n	n	bacterial & archaeal	GBK, GFF or SCO
GeneMark.hmm	y	n	n	prokaryotes	algorithm-specific
GeneMark	y	n	n	prokaryotes	algorithm-specific
GenMarkS	y	n	y	prokaryotes	algorithm-specific
RAST	y	y	y	bacteria and archarea	GTF,GFF3,GenBank,EMBL

JCVI Annotation Service	y	y	y	prokaryotes	algorithm-specific
AMIGene	y	n	n	prokaryotes	EMBL, GenBank, GFF
Glimmer3	y	n	n	prokaryotes	algorithm-specific
EasyGene	y	n	n	prokaryotes	GFF2
Maker	y	n	y	small eukaryotes and prokaryotes	GFF3
Augustus	y	n	y	eukaryotes	GTF (similar to gff). GFF

[a] stable RNA refers to rRNA, tRNA, tmRNA, RNA Component of RNaseP

[b] FA stands for functional annotations i.e. mRNA, operons, promoters, terminators, protein-binding sties, DNA bends

The gene finders listed in Table 1 were tested on the bacterial strain *Pseudomonas aeruginosa* LESB58 (*P.a.* LESB58), which has a high GC-content (~66.3%). Most of these gene finders are specialized for prokaryotes. Although Augustus is developed only for eukaryotes, it offers the option to be trained on a given set of genomes. Therefore, Augustus was trained on the 10 closest genomes of *Pseudomonas aeruginosa* LESB58.

Table 2. Results from Testing the Gene Finders on *P.a.* LESB58

Gene Finder	# Genes	# Genes on the + Strand	# Genes on the - Strand	#Correct Genes	% Correct Genes (compared to the Original)	% Correct Genes from (from all found genes)
Original	6061	2993	3067	6061	100,00%	100,00%
Prodigal	6055	3014	3041	5286	89,14%	87,30%
FGenesB	6197	3094	3103	5070	85,50%	81,81%
Glimmer3.0	6276	3100	3176	5043	85,04%	80,35%
GeneMarkS	6100	3043	3057	5006	84,42%	82,07%
JCVI	6270	3098	3172	5036	83,10%	80,32%
GeneMarkHMM	6129	3055	3074	4920	82,97%	80,27%
Rast	6297	3116	3181	4940	81,52%	78,45%
MED	7475	3708	3767	4747	80,05%	63,51%
Maker with model	6149	3065	3084	4588	75,71%	74,61%
Maker	5884	2904	2980	4370	72,11%	74,27%
Augustus	5268	2587	2681	3529	59,51%	66,99%
AMIGene	6154	3077	3077	2967	50,03%	48,21%
EasyGene	3150	0	3150	2570	43,34%	81,59%

Expectedly, Prodigal coped robustly with the high GC content of the strain *P.a.* LESB58. From the Markov model-based algorithms for gene prediction, FGenesB generated the biggest number of correct genes, performing slightly better than Glimmer and GeneMarkS. Glimmer lengthened genes, resulting in drastically higher average gene length. AMIGene found many genes that were not recognized by other gene finders; however more than half of the genes it predicted were not correct.

Approximately 11.6% from all the genes in P.a.LESB58 were detected by every gene finder.

The Results suggest that Prodigal is preferable for gene prediction in high GC genomes. However, for the purpose of testing the gene finders, the hypothetical proteins were not excluded from the published annotation.

4 Future Directions in Microbial Gene Prediction

Microbial gene identification is a well-studied problem. Since the early eighties of the twentieth century, there has been great progress in the development of computational gene prediction. There is still much room for improvement, especially in understanding the translation initiation mechanisms.

The accuracy of most of the gene finding methods drops considerably, when high GC content genomes are observed. Moreover, most methods tend to predict too many genes, mainly because of the problem of predicting short genes. Although many short genes without a BLAST hit might be real, the likelihood is that the most are false positives.

The evaluation system of gene prediction programs is still in need of improvement. The authors of all the gene finders mentioned in this review estimated the accuracy of the tools by predicting genes in complete genomes and then comparing the output to the "known" genes. However, it is estimated that 10-30% of the annotated genes are not protein-coding genes, but rather ORFs that occur by chance [20]. The gene finders exclude hypothetical proteins for testing purposes, because published annotations are not 100% accurate; therefore, the question remains open as to how accurate these predictions really are. The need for more reasonable criterions for evaluation of gene prediction programs is apparent.

It is important to improve current methodologies to obtain higher quality gene predictions, translation initiation site prediction and reduction in the number of false positives, in order to minimize the need for manual curation. Future gene finders should enable automatic gene prediction without human intervention.

5 Conclusion

At the present, there is no tool for gene prediction that automatically finds all the genes in a given genomic DNA sequence with 100% accuracy. The most reliable method for identifying genes is by similarity to a protein in other organism. Genes with no match to known proteins can be predicted using statistical measures.

Every algorithm for gene prediction has its advantages and limitations. Currently, the best approach seems to be a combination of gene finders, followed by evidence-based manual curation.

Acknowledgments. Ljupco Kocarev thanks ONR Global (Grant number N62909-10-1-7074) and Macedonian Ministry of Education and Science (grant 'Annotated graphs in system biology') for partial support.

References

1. Kahvejian,A., Quackenbush,J., and Thompson,J.F.: What would you do if you could sequence everything? Nat Biotechnol, 26,1125-1133 (2008).
2. Tats,A., Tenson,T., and Remm,M.: Preferred and avoided codon pairs in three domains of life. Bmc Genomics, 9 (2008).
3. Ermolaeva,M.D.: Synonymous codon usage in bacteria. Curr.Issues Mol Biol, 3,91-97 (2001).
4. Borodovksy, M., Hayes, W.S., and A.V.Lukashin: In R.L.Charlebois (ed), Organization of Prokaryotic Genomes. ASM Press,pp 11-33 (1999).
   Charlebois,R.L.: Statistical Predictions of Coding Regions in Prokaryotic Genomes by Using Inhomogeneous Markov Models. American Society Microbiology, pp 11-33 (1999).
5. Jin Xiong: Essential Bioinformatics. Cambridge University Press,pp 97-112 (2006).
6. Besemer,J., Borodovsky,M.: GeneMark: web software for gene finding in prokaryotes, eukaryotes and viruses. Nucleic Acids Res, 33,W451-W454 (2005).
7. Borodovksy, M., McIninch J.D. GeneMark: Parallel Gene Recognition for Both Strands. Comupt Chem, 17,123-133 (1993).
8. Lukashin,A.V., Borodovsky,M.: GeneMark.hmm: new solutions for gene finding. Nucleic Acids Res, 26,1107-1115 (1998).
9. Besemer,J., Lomsadze,A., and Borodovsky,M.: GeneMarkS: a self-training method for prediction of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions. Nucleic Acids Res, 29,2607-2618 (2001).
10. Borodovksy, M., McIninch J.D.: Recognition of genes in DNA sequence with ambiguities. BioSystems, 30,161-171 (1993).
11. Delcher,A.L., Bratke,K.A., Powers,E.C., and Salzberg,S.L.: Identifying bacterial genes and endosymbiont DNA with Glimmer. Bioinformatics, 23,673-679 (2007).
12. NMPDR, http://rast.nmpdr.org/
13. Yandell Lab, , http://www.yandell-lab.org/software/maker.html
14. J. Craig Venter Institute, http://www.jcvi.org/cms/research/projects/prokaryotic-annotation-pipeline/overview/
15. Bocs,S., Cruveiller,S., Vallenet,D., Nuel,G., and Medigue,C.: AMIGene: Annotation of MIcrobial genes. Nucleic Acids Research, 31,3723-3726 (2003).
16. Softberry Inc. FGENESB Suite of Bacterial Operon and Gene Finding Programs, http://linux1.softberry.com/berry.phtml?topic=fgenesb&group=help&subgroup=gfindb 2010. 6-1-2010.
17. Altschul,S.F., Gish,W., Miller,W., Myers,E.W., and Lipman,D.J. (1990) Basic Local Alignment Search Tool. Journal of Molecular Biology, 215,403-410.
18. Doug Hyatt, Gwo-Liang Chen, Philip F.LoCascio, Miriam L.Land, Frank W.Larimer, and Loren Hauser (2010) Prodigal: prokaryotic gene recognition and translation initiation site identification. Bmc Bioinformatics, 11-119.
19. Skovgaard,M., Jensen,L.J., Brunak,S., Ussery,D., and Krogh,A. (2001) On the total number of genes and their length distribution in complete microbial genomes. Trends in Genetics, 17,425-428.
20. Bibb,M.J., Findlay,P.R., and Johnson,M.W.: The relationship between base composition and codon usage in bacterial genes and its use for the simple and reliable identification of protein-coding sequences. Gene, 30,157-166 (1984).
21. Baxevanis, A.D., and Ouellette, B.F.F.: ”Sequence alignment and Database Searching” in *Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins*. John Wiley & Sons, New York, pp. 187-212 (2001)

Naïve Bayes technique for diatoms classification with discretised input

Andreja Naumoski, Kosta Mitreski

University Ss. „Cyril and Methodius“Faculty of Electrical Engineering and Information Technologies, Skopje, Karpos 2 bb, Skopje, Macedonia
{andrejna,komit}@feit.ukim.edu.mk

Abstract. The challenge to discover knowledge from environmental data that has led to usage of methods and techniques such as data mining tools, can bridge the knowledge gap between the biological experts and organisms. This research aimed to assess relationships between the diatoms and the indicators of the environment with Naïve Bayes method. Diatoms are ideal indicators of certain physical-chemical parameters and they can be classified into one of the water quality classes (WQCs). The classification models are induced by using Naïve Bayes technique. The input dataset that is supplied for the naïve Bayes method is discretised. Based on the evaluation results, several models are presented and discussed. The obtain results from the models are verified with existing diatom ecological preference and for some diatoms new knowledge is added. To best of our knowledge, this is the first time the prosed method to be applied for diatom classification of any ecosystem.

Keywords: Naïve Bayes, diatom classification, indicators, water ecosystem

1 Introduction

The water quality classes define in the traditional way can be interpreted as a classification problem in the terms of data mining point of view. In this paper research, this property is used to discover the appropriate environment conditions for newly found diatom, which are an ideal bio-indicator of a certain physico-chemical parameter. Considering these facts, we deal with the typical classification problem, when we try to build a model that classifies the correct diatoms into one of the WQ classes.

In this domain, classical statistical approach, such as canonical correspondence analysis (CCA), detrended correspondence analysis (DCA) and principal component analysis (PCA), are most widely used as modelling techniques [18]. Although these techniques provide useful insights in the data, they are limited in terms of interpretability. Obvious progress in this research area in a direction of interpretability, have been made using data mining techniques, mainly decision trees. These methods, improves the interpretability and increases the prediction power of the models. First attempt to model diatom-indicator relationship for Lake Prespa, have been made by [4]. Several of the model produced, knowledge about the newly

discovered diatom's relationships with the environment for the first time [4]. New class of multi-target decision trees later were used, in order to reveal the dynamic nature of the entire set of physical-chemical parameters of this lake ecosystem [15]. These methods were more precise and also have increased the interpretability. Nevertheless, these methods were not robust on data change. This is an important property, because the environmental condition inside of the lake changes over small periods of time.

Many empirical comparisons between naive Bayes and modern decision tree algorithms such as C4.5 (Quinlan 1993) showed that naive Bayes predicts equally well as C4.5 [3, 5, 7] for many real data domains. To best of our knowledge, this is the first usage of Naïve Bayes classifier for diatom classification. The good performance of naive Bayes is surprising because it makes an assumption that is almost always violated in realworld applications: given the class value, all attributes are independent. This is also true for the diatoms that are independent; one diatom can be indicator of one water quality class and one for another. Nevertheless, from ecological point of view it is very important to estimate the degree that diatom depends from the certain environmental conditions.

Domingos and Pazzani [8] present an explanation that naive Bayes owes its good performance to the zero-one loss function. In [9] the authors have shown that the performance of naive Bayes is much worse when it is used for regression (predicting a continuous value). Moreover, evidence has been found that naive Bayes produces poor probability estimates [11, 12]. That's way the input dataset that we will use in the experiments shown in the paper are discrete and we donate a certain class for each diatom.

The rest of the paper is organized as follows: Section II provides the definitions for the Naïve Bayes classifier. In Section III we present the diatoms abundance water quality datasets as well as the experimental setup. Section IV gives the experimental results and the verification of the model results and finally, Section V concludes the paper and the research directions are outlined.

2. Naïve Bayes method

Classification is a fundamental issue in machine learning and data mining. In classification, the goal of a learning algorithm is to construct a classifier given a set of training examples with class labels. Typically, an example E is represented by a tuple of attribute values (x_1, x_2,.., x_n), where x_i is the value of attribute X_i. Let C represent the classification variable, and let c be the value of C.

A classifier is a function that assigns a class label to an example. From the probability perspective, according to Bayes Rule, the probability of an example E = (x_1, x_2,…, x_n) being class C is:

$$p(c \mid E) = \frac{p(E \mid c)p(c)}{p(E)} . \quad (1)$$

Assume that all attributes are independent given the value of the class variable; that is,

$$p(E \mid c) = p(x_1, x_2, ..., x_n \mid c) = \prod_{i=1}^{n} p(x_i \mid c) . \quad (2)$$

The resulting classifier is then:

$$f_{NB}(E) = p(C) \prod_{i=1}^{n} p(x_i \mid C) . \quad (3)$$

The function f_{NB} (E) is called a naive Bayesian classifier, or simply Naive Bayes (NB) (see eq.3). This is called conditional independence. In our paper it is obvious that the conditional independence assumption is true, meaning that each diatom is independent from one water quality class.

In order to estimate the probability that one diatom belongs into one water quality class we will use standardize normal distribution, express as:

$$F_X(x) = \Phi\left(\frac{(x-\mu) \pm (precision/2)}{\sigma}\right), where\ \Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-t^2/2} dt, . \quad (4)$$

The precision number is estimated by the Bayes classifier, together with the μ and σ for diatom and each WQC. The *x* value is inputted as discrete class terms, because of the ecological (uncertainty) nature of the diatom dataset, and the better performance reported by [11, 12]. The Naïve Bayes classifier algorithm was used from the WEKA machine learning toolkit [19]. The discrete class values are given below.

3. Data description and experimental setup

The datasets used in the experiments consist from 13 input parameters representing the TOP10 diatom species (diatom species that exist in Lake Prespa [2]) with their abundance per sample, plus the three WQC for conductivity, pH and Saturated Oxygen. Then one dataset is crated for each WQ class with the TOP10 diatoms.

Table 1. Water quality classes for the physical-chemical parameters [16, 17]

Physical-chemical parameters	Name of the WQC	Parameter range	Name of the WQC	Parameter range
Saturated Oxygen	oligosaprobous	SatO > 85	α-mesosaprobous	25-70
	β-mesosaprobous	70-85	α-meso / polysaprobous	10-25
pH	acidobiontic	pH < 5.5	alkaliphilous	pH > 7.5
	acidophilous	pH > 5.5	alkalibiontic	pH > 8
	circumneutral	pH > 6.5	Indifferent	pH > 9
Conductivity	fresh	Conduc < 20	brackish fresh	90 – 180
	fresh brackish	Conduc < 90	brackish	180 - 900

These measurements were made as a part of the TRABOREMA project [6]. The WQCs are defined according to the three physical-chemical parameters: Saturated Oxygen [16], Conductivity [17] and pH [16, 17] which are given in Table 1. Among the input parameters 10 are numerical parameters and the rest 3 are nominal with a number of possible classes from 3 to 6.

The experimental setup estimates the highest probability of diatom with water quality class. After the data is process by the algorithm, full classification model for each water quality class, then probability measured using normal distribution is estimated. The normal distribution takes as input value one discretised class term from the Table 2.

Table 2. Discretised input dataset into probability estimator

Diatoms	**DTerm 1 – DT1**	**DTerm 2 – DT2**	**DTerm 3 – DT3**	**DTerm 4 – DT4**	**DTerm 5 – DT5**
	Bad	Weak	Good	Very Good	Excellent
APED	0	3.25	6.5	9.75	13
CJUR	0	21.5	43	64.5	86
COCE	0	20.25	40.5	60.75	81
CPLA	0	10	20	30	40
CSCU	0	10.25	20.5	30.75	41
DMAU	0	3	6	9	12
NPRE	0	4.75	9.5	14.25	19
NROT	0	6	12	18	24
NSROT	0	7.75	15.5	23.25	31
STPNN	0	5.25	10.5	15.75	21

5. Experimental results

In this section, three models are given for each water quality class, to show the probability estimates from the Naïve Bayes classification. Later the results from the classification models with the known ecological reference of the diatoms are verified.

5.1 Interpretation of the classification models

All the induced classification models have a define range of discretised class terms, which later will be commented. Each diatom for certain class has a probability estimate, which is important measure of indicator properties of the diatom.

The results from the classification model for Conductivity water quality class are presented in Table 3. According to the classification model, the APED diatom is a bad indicator of brackish water with 99.73% of probability, while he is weak indicator of brackish fresh waters with probability of 14.46%. The model also identifies the APED diatom as good indicator of fresh brackish waters with probability of 5.35%, while the other estimates are very low. Similar conclusion can be made for all the TOP10 diatoms. We will summarize just a few of them. For example, COCE diatom is weak indicator of brackish waters (2.18%), on other hand he is good indicator of brackish

fresh waters. According to the classification model the DMAU diatom is a weak indicator of brackish waters, while good indicator of brackish waters.

Table 3. Evaluation results from the classification model for Conductivity water quality class

Diatoms	Bad	Weak	Good	Very Good	Excellent
Class APED	brackish 99.73%	Brackish fresh 14.46%	fresh brackish 5.35%	fresh brackish 0.60%	fresh brackish 0.02%
Class CJUR	fresh brackish 99.73%	brackish 2.18%	brackish 0.00%	brackish 0.00%	Fresh brackish 0.00%
Class COCE	fresh brackish 99.73%	brackish 2.64%	Brackish fresh 2.13%	Brackish fresh 1.15%	Brackish fresh 0.29%
Class CPLA	fresh brackish 99.73%	Brackish fresh 0.37%	Brackish fresh 0.00%	Fresh brackish 0.00%	Fresh brackish 0.00%
Class CSCU	Fresh brackish 95.28%	brackish 6.09%	Brackish fresh 2.90%	Brackish fresh 0.44%	Brackish fresh 0.02%
Class DMAU	fresh brackish 99.73%	Brackish fresh 14.19%	brackish 5.40%	brackish 0.67%	brackish 0.03%
Class NPRE	fresh brackish 99.73%	Brackish fresh 14.50%	Brackish fresh 1.12%	Brackish fresh 0.01%	Brackish fresh 0.00%
Class NROT	fresh brackish 99.73%	brackish 12.07%	brackish 0.89%	brackish 0.01%	brackish 0.00%
Class NSROT	fresh brackish 99.73%	Brackish fresh 9.69%	brackish 0.60%	brackish 0.00%	brackish 0.00%
Class STPNN	fresh brackish 99.73%	Brackish fresh 10.81%	Brackish fresh 1.92%	Brackish fresh 0.06%	Brackish fresh 0.00%

The STPNN diatom has weak indicator properties for brackish fresh waters, while the NROT diatom for brackish waters. It is interesting to note that the low indicator properties is not a of impropriate method for classification, but more to the quality and quantity of the data. This was concluded for this diatom dataset in experiments with previous methods [15]. The classification model, classified the diatoms as bad indicators, because most of the data contained values of diatoms abundance near 0. We have assumed that low abundance of certain diatoms is bad indicator of given water quality class, but it was unknown for which class.

The evaluation results for the pH water quality class are presented in Table 4. From the model, it is easy to note that APED diatom is a good indicator of *alkaliphilous* waters, and weak indicator of *alkalibiontic* and bad indicator of *circumneutral* waters. NPRE diatom is bad indicator of *acidophilous* waters, but good to excellent indicator of *acidophilous* waters. NROT, NSROT and STPNN diatoms are good to excellent indicators of indifferent waters, but with low probability according the model. All the

diatoms more or less have around 15% probability to be good indicators of certain water quality class.

Table 4. Evaluation results from the classification model for pH water quality class

Diatoms	Bad	Weak	Good	Very Good	Excellent
Class	circumneutral	alkalibiontic	alkaliphilous	alkaliphilous	alkaliphilous
APED	19.19%	14.16%	8.00%	2.51%	0.34%
Class	acidophilous	alkalibiontic	alkalibiontic	alkalibiontic	acidophilous
CJUR	30.31%	3.24%	0.00%	0.00%	0.00%
Class	acidophilous	Indifferent	Indifferent	alkaliphilous	alkaliphilous
COCE	1.55%	2.51%	2.38%	0.80%	0.17%
Class	Indifferent	alkalibiontic	alkalibiontic	alkalibiontic	alkalibiontic
CPLA	95.46%	6.11%	0.14%	0.00%	0.00%
Class	Indifferent	circumneutral	alkalibiontic	alkalibiontic	alkalibiontic
CSCU	99.73%	7.06%	3.82%	1.01%	0.09%
Class	circumneutral	Indifferent	acidophilous	acidophilous	acidophilous
DMAU	14.08%	19.27%	7.75%	3.96%	1.24%
Class	acidophilous	alkaliphilous	alkaliphilous	alkaliphilous	alkaliphilous
NPRE	99.73%	12.20%	3.60%	0.33%	0.01%
Class	acidobiontic	Indifferent	Indifferent	Indifferent	Indifferent
NROT	99.73%	13.96%	2.70%	0.08%	0.00%
Class	acidobiontic	Indifferent	Indifferent	Indifferent	Indifferent
NSROT	99.73%	10.00%	2.69%	0.13%	0.00%
Class	acidobiontic	Indifferent	Indifferent	Indifferent	Indifferent
STPNN	99.73%	8.87%	0.79%	0.01%	0.00%

Concerning the last water quality class – Saturated Oxygen, the results from the classification model (see Table 5) shows that the TOP10 diatoms have all bad indicator properties for the *polysaprobous* WQC class.

Table 5. Evaluation results from the classification model for Saturated Oxygen water quality class

Diatoms	Bad	Weak	Good	Very Good	Excellent
Class	Poly saprobous	Oligo saprobous	β-meso saprobous	β-meso saprobous	β-meso saprobous
APED	99.73%	14.89%	5.07%	0.42%	0.01%
Class	Poly saprobous	α-meso saprobous	α-meso saprobous	α-meso saprobous	α-meso saprobous
CJUR	99.73%	6.54%	0.16%	0.00%	0.00%
Class	Poly saprobous	β-meso saprobous	α-meso saprobous	α-meso saprobous	α-meso saprobous
COCE	99.73%	2.79%	2.15%	0.80%	0.11%
Class	Poly saprobous	Poly saprobous	β-meso saprobous	β-meso saprobous	β-meso saprobous
CPLA	99.73%	10.26%	1.54%	0.07%	0.00%
Class	Poly saprobous	α-meso saprobous	Oligo saprobous	Oligo saprobous	Oligo saprobous

CSCU	99.73%	6.82%	2.80%	0.39%	0.02%
Class	Poly saprobous	β-meso saprobous	α-meso saprobous	α-meso saprobous	α-meso saprobous
DMAU	99.73%	15.07%	6.33%	1.33%	0.11%
Class	Poly saprobous	Oligo saprobous	Oligo saprobous	Oligo saprobous	Oligo saprobous
NPRE	99.73%	13.25%	1.68%	0.03%	0.00%
Class	Poly saprobous	Oligo saprobous	α-meso saprobous	α-meso saprobous	α-meso saprobous
NROT	99.73%	12.29%	0.98%	0.01%	0.00%
Class	Poly saprobous	Oligo saprobous	Oligo saprobous	Oligo saprobous	Oligo saprobous
NSROT	99.73%	9.70%	0.74%	0.00%	0.00%
Class	Poly saprobous	α-meso saprobous	Oligo saprobous	Oligo saprobous	Oligo saprobous
STPNN	99.73%	9.78%	0.49%	0.00%	0.00%

The APED diatom and the CPLA diatom, according the classification model are weak indicator of *oligosaprobous* water class, but good to excellent indicators for *β-mesosaprobous* with very low probability. The CSCU, NPRE, NROT, NSROT and STPNN diatoms according to the models are weak to very good indicators of *oligosaprobous* waters. The rest of the diatoms are less or more weak to good indicators of *α-mesosaprobous* waters. Once more, the classification model has low values for the probability estimates, except for the first class.

5.2 Verification of the results from the models

Ecological references for the TOP10 diatom are taken from the latest diatom ecology publications [14], used in several recently published papers [1, 2, 4, 15], and database (European Diatom Database - http://craticula.ncl.ac.uk/Eddi/jsp/index.jsp). Concerning ecological reference of the TOP10 dominant diatoms in Lake Prespa, CJUR and NPRE are newly described taxa (diatoms) with no record for their ecological preferences in the literature. Therefore, some of the results from the classification models are the first known ecological reference for certain WQC classes.

In the relevant literature APED diatom is known to be *alkaliphilous*, *fresh-brackish*, nitrogen-autotrophic (tolerates elevated concentrations of organically bound nitrogen), high oxygen saturation (>75%), *β-mesosaprobic* and *eutrophic* (because of Organic N tolerance) diatom indicator [14]. According to the classification models the APED diatoms is found to be an *alkaliphilous* and *fresh-brackish* indicator. Regarding the Saturated Oxygen WQ classes, APED is a weak indicator of *oligosaprobous,* but good indicator of *β-mesosaprobic* environment.

Concerning CSCU diatom indicator affinity, the model for pH WQC has revealed that this diatom is *alkalibiontic*. According to the models for conductivity WQ class the CSCU diatom is *brackish* to *brackish fresh* diatom, while the Saturated Oxygen WQC model shows the weak affinity of this diatom to *α-mesosaprobous*, but with

good indicator properties for *oligosaprobous* waters. In the relevant literature the CSCU is known as *alkalibiontic*, *freshwater* to *brakish* water taxon, being *oligosaprobic* indicators with eutrophic references [14].

The COCE diatom is known as *meso-eutro* taxon [14], while concerning the pH properties of this diatom, there is no known ecological reference. According to the classification models, the COCE diatom is relative good indicator for *brackish fresh* waters, *indifferent* to *alkaliphilous,* and for the saturated oxygen demand he is a weak indicator of *β-mesosaprobous* but relatively good for *α-mesosaprobous* environments. Further experiments to investigate the trophic indicator affinity of this diatom should be made by using trophic state index classes.

The STPNN diatom, in the literature is known as *hyper-eutrophic* (*oligo-eutrophic*; indifferent) taxon frequently found on moist habitats, while the classification models have been found to be *alkalibiontic* taxon. According to the classification models, this diatom is a weak to good indicator of *brackish* waters, while *indifferent* taxon for pH WQC. The model for the saturated oxygen showed that this diatom is weak indicator of *α-mesosaprobous* waters, but relatively good for *oligosaprobous*.

The other ecological references for the rest of the diatoms are new and they have to be further investigated, before any solid conclusion is made. Nevertheless, many of the known ecological references are verified with the classification method, thus proving the reliability of the proposed method for diatom classification.

6. Conclusion

The proposed method has verified the known diatom ecological knowledge and for some of them, added a new knowledge. Classifying the diatoms from measure data can be greatly improved with the proposed method, not just from Lake Prespa, but from any lake ecosystem, since the geographical location plays no role in the bio-indicator properties of certain diatom [13].

The experiments on diatom WQC datasets show that the Naïve Bayes method can be a good tool for diatom classification. For each of the defined WQ classes, the method has found a relationship between the diatoms and the indicator with certain probability. The input data in the proposed method is divided into classes, with labeled a term, associate with a define range. With this process, the classification accuracy of the proposed method is higher, based on the research work done previously on other datasets. Also, another fact is the changing ecosystem conditions, which adds a degree of uncertainty in the process of diatom classification. That's way, we use the Naïve Bayes classifier, because estimates the probability of a diatom in a certain WQC class and reduces the uncertainty which is accompanied with the environmental data.

More important is the interpretation of the classification models, compared with the classical statistical methods such as: PCA, CCA, DCA and other methods, used previously, the proposed method is more directly interpretable. The obtained models have openly stated prediction and probability in terms of finding correct diatom-indicator relationship. The experiments showed that machine learning tools can

extract valuable knowledge in a relatively comprehensible form, even when the application area is so extremely complex for humans and the data are far from being perfect.

We believe that studies like ours that combines the ecological together with information technologies, especially in the area of eco informatics, are necessary to provide understanding of the physical, chemical and biological processes and their relationship to aquatic biota for predicting a certain effect. Verification of the obtained models showed that the proposed method, have successfully classified certain known diatoms, and added new ecological knowledge for the unknown diatoms for certain WQCs.

Further research needs to be focused on developing classification models base on the Naïve Bayes method for trophic state index classes. Other methods for classification could be suitable for diatoms classification that needs to be explored.

References

1. Krstič, S.: Description of sampling sites. FP6-project TRABOREMA: Deliverable 2.2. (2005).
2. Levkov. Z., Krstič. S., Metzeltin. D., Nakov. T.: Diatoms of Lakes Prespa and Ohrid (Macedonia). Iconographia Diatomologica, vol. 16, pp. 603 (2006)
3. Langley, P., Iba, W., and Thomas, K.: An analysis of Bayesian classifiers. In Proceedings of the Tenth National Conference of Artificial Intelligence. AAAI Press, pp. 223—228 (1992)
4. Naumoski, A., Kocev, D., Atanasova, N., Mitreski, K., Krtić, S., Džeroski, S.: Predicting chemical parameters of water quality form diatoms abundance in lake Prespa and its tributaries. The 4th International ICSC Symposium on Information Technologies in Environmental Engineering - ITEE 2009. Springer Berlin Heidelberg press, Thessaloniki, Greece pp. 264--277. (2009)
5. Kononenko, I.: Comparison of inductive and naïve Bayesian learning approaches to automatic knowledge acquisition. In Wielinga, B., ed., Current Trends in Knowledge Acquisition. IOS Press, (1990)
6. TRABOREMA Project WP3.: EC FP6-INCO project no. INCO-CT-2004-509177 (2005-2007)
7. Pazzani, M. J.: Search for dependencies in Bayesian classifiers. In Fisher, D., and Lenz, H. J., eds., Learning from Data: Artificial Intelligence and Statistics V. Springer Verlag, (1996)
8. Domingos, P., Pazzani, M.: Beyond independence: Conditions for the optimality of the simple Bayesian classifier. Machine Learning 29, pp. 103--130 (1997)
9. Friedman, J.: On bias, variance, 0/1-loss, and the curse of dimensionality. Data Mining and Knowledge Discovery 1, (1996)
10. Quinlan, J. C4.5: Programs for Machine Learning. Morgan Kaufmann: San Mateo, CA, (1993)
11. Bennett, P. N.: Assessing the calibration of Naïve Bayes' posterior estimates. In Technical Report No. CMUCS00-155 (2000)
12. Monti, S., Cooper, G. F.: A Bayesian network classifier that combines a finite mixture model and a Naïve Bayes model. In Proceedings of the 15th Conference on Uncertainty in Artificial Intelligence. Morgan Kaufmann. Pp. 447--456 (1996)
13. Gold, C., Feurtet-Mazel, A., Coste, M., Boudou, A.: Field transfer of periphytic diatom communities to assess shortterm structural effects of metals (Cd Zn) in rivers. Water Research Vol. 36, pp. 3654--3664 (2002)

14. Van Dam, H., Martens, A., Sinkeldam, J.: A coded checklist and ecological indicator values of freshwater diatoms from the Netherlands. Netherlands Journal of Aquatic Ecology Vol. 28, Issue. 1, pp. 117—133 (1994)
15. Kocev. D., Naumoski. A., Mitreski. K., Krstić. S., Džeroski. S.: Learning habitat models for the diatom community in Lake Prespa. Journal of Ecological Modelling, vol. 221, No. 2, pp. 330--337 (2009)
16. Krammer. K., Lange-Bertalot. H.: Die Ssswasserflora von Mitteleuropa 2: Bacillariophyceae. 1 Teil. pp. 876, Stuttgart: Gustav Fischer-Verlag (1986)
17. Van Der Werff. A., Huls. H.: Diatomeanflora van Nederland. Abcoude - De Hoef (1957, 1974)
18. Stroemer. E. F., Smol. J. P.: The diatoms: Applications for the Environmental and Earth Sciences, Cambridge University Press, Cambridge (2004)
19. WEKA 3.6.2 - Machine Learning Toolkit - http://www.cs.waikato.ac.nz/ml/weka/

Hierarchical Protein Classification based on Gene Ontology and Decision Trees

Ilinka Ivanoska, Georgina Mirceva, Kire Trivodaliev and Slobodan Kalajdziski

Ss. Cyril and Methodious University, Faculty of Electrical Engineering and Information Technologies, Karpos 2 bb, 1000 Skopje, Macedonia
{ilinkai,georgina,kiret,skalaj}@feit.ukim.edu.mk

Abstract. Proteins are the most important cell parts, therefore, knowing their exact function is of a great significance. However, the function of large amount of proteins is still unknown. In addition, today, biologists persist on hierarchical organization the living world, and thus in protein databases also. There are many protein classification algorithms proposed determining the protein function, but, only a few of them take into consideration these hierarchical structures. The Gene Ontology (GO) is a protein and gene database structured as a controlled hierarchical vocabulary of terms to describe protein functions. This paper introduces a new hierarchical multi-label protein classifier that uses the relationships among the GO terms. First, protein descriptors are extracted from the structural coordinates stored in the Protein Data Bank (PDB) files. Then, a modified C4.5 algorithm is applied to select the most appropriate descriptor features for protein classification based on the GO hierarchy. An evaluation of this approach is presented, and the results show that the hierarchical structure of GO is important for improving the accuracy of the classification problem at higher levels.

Keywords: C4.5 Classification, Gene Ontology, Protein function prediction.

1 Introduction

Proteins are the most important cell molecules and they are included in every living organisms function, starting from immune system regulation, to muscles structure, hormones, metabolism, breathing, genes control etc. Thus, knowing their structure plays a vital role in determining their function, and this is important in finding methods for prevention and treatment of many diseases. In addition, the knowledge of protein function is a crucial link in the development of new drugs, better crops, and even development of synthetic bio-chemicals.

Proteins are constructed by long chains of amino acid residues folding into complex three-dimensional polypeptide chain structures. This three-dimensional representation of a residue sequence and the way this sequence folds in the 3D space are very important to understand the logic in which a function of a protein is based on. In fact, the concept of function typically acts as an specific term for all types of activities that a protein is involved in, be it cellular, molecular or physiological. Also,

evolutionary evidence could potentially be derived from conserved protein structures existed in multiple species.

A need for a classification of proteins is obvious, which may result in a better understanding of these complicated three-dimensional structures, their functions, and the deeper evolutionary procedures that led to their creation. In molecular biology, many classification schemes and databases (GO [1], CATH [16], FSSP [15] and SCOP [2]) have been developed in order to describe the different kinds of similarity between proteins.

Controlled vocabularies such as the Gene Ontology (GO) are becoming increasingly important for automated protein classification methods. Several systems have been developed that employ pattern recognition, using a variety of features, including sequence similarity[8][9][10], presence of protein functional domains and gene expression patterns, and others, but most of these approaches have not considered the hierarchical structure of the GO. The Direct Acyclic Graph (DAG) represents the functional relationships between the GO terms, thus it should be an important component of an automated classification system. The classification methods usually ignore the hierarchical nature of the controlled vocabulary in order to simplify the classification problem.

Only, several authors have recently used hierarchical information in [11] report that the training set including hierarchical information outperformed similar data in which the hierarchical nature of the data was ignored. In [12] predicted gene function is based on the relationship between GO annotations using decision trees and a Bayesian network.

The Structural Classification of Proteins - SCOP database [2] is another protein database that describes the evolutionary relationships between proteins of known structure and similarly has a hierarchical structure. The main levels of the SCOP hierarchy are Family, Superfamily, Fold, and Class. Using the terminology of the SCOP database, two proteins that belong to the same fold share a common three-dimensional pattern with the same major secondary structure elements (SSEs) in the same arrangement with the same topological connections. In the SCOP hierarchy, folds are grouped into different classes, where a class is defined by the topographical arrangement of the secondary structures of its member proteins.

CATH (Class, Architecture, Topology, Homologous SuperFamily) [16] is other protein classification system that has a hierarchical structure (8 level of hierarchy: Class, Architecture, Topology, Homologous Superfamily, and another four lower levels of hierarchy) which strives for an automated classification, without using the human classification factor.

There are various approaches for protein classification which are trying to offer efficient and completely automated protein classification for these hierarchical classification schemes. These approaches have different characteristics in terms of algorithm for determining protein similarity. Basically, the protein similarity metric used defines the complexity and the efficiency of the classification approach, and usually does not consider the hierarchical structure.

One way to determine protein similarity is to use sequence alignment algorithms like Needleman–Wunch [19], BLAST [18], PSI-BLAST [17] etc. They offer fast and efficient recognition of overlapping subsequences in two protein structures which

leads to detection of closely related protein structures, but these methods cannot recognize proteins with remote homology.

Instead of sequence alignment methods, structure alignment methods like CE [5], MAMMOTH [6], DALI [7], etc. are used to detect and highlight distant homology relations between protein structures. In general, these methods are very precise and efficient and they have high degree of successful mapping of existing structures in new proteins. Structure alignment methods perform one-against-all proteins comparison in order to find the most similar existing protein to a novel protein structure. Having in mind that the number of classified proteins and annotated, for example in Gene Ontology, is ever increasing and that structure alignment methods are quite cost expensive, the speed of classification with these methods is always questioned.

In [21] a hierarchical classification algorithm is used based on decision trees for the Munich Information Center for Protein Sequences (MIPS) scheme [20], but the accuracy of the classification algorithm more than level 4 of the hierarchy is less than 30% which is not satisfactory.

In comparison to Gene Ontology, in [14] the protein classification approach is based on a bottom-up classification flow, and a multi-level classification approach with a combination of protein descriptors and a C4.5 decision tree classification algorithm. There as a classification scheme, the SCOP classification hierarchy is used. In this paper we extend their approach on the Gene Ontology (GO) classification scheme and modify the C4.5 classification algorithm for the hierarchical structure.

In section 2 we present the classification process architecture and the used classification methods. Section 3 presents the experimental results, while the section 4 concludes the paper.

2 Gene Ontology Protein classification architecture based on structural features and decision trees

The classification process architecture has three main features. First, it uses 3D protein descriptor [13] which transforms the protein tertiary structure into N-dimensional feature vector, and additionally gives some other protein structural features. Second, this architecture is based, and it uses the Gene Ontology classification scheme. And finally, as a classification algorithm decision trees trained with C4.5 are used.

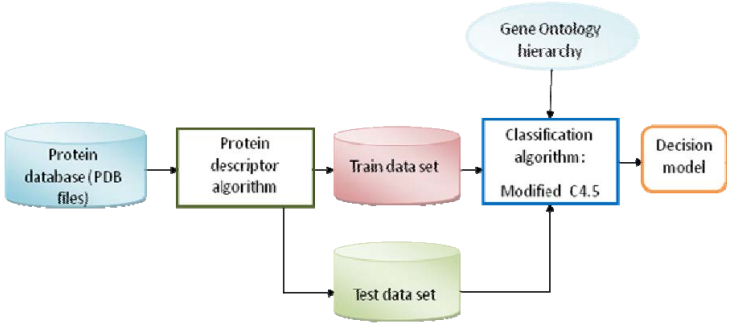


Fig. 1. Classification process architecture

The system (as can be seen from the Fig. 1) is consisted of two phases: training phase and testing or classification phase. The training of offline data flow takes into consideration the knowledge given in the protein database to build predicative classification flow for the GO hierarchy.

2.1 Protein descriptor

The training procedure is the descriptor extraction (shown on Fig. 2). Descriptors consisting of 450 features (416 of them describe the protein's geometry, while 34 of them give information for the primary and secondary protein structure) are generated for each protein forming a training set for the C4.5 decision tree algorithm. This descriptor [13] relies on the geometric 3D structure of the proteins. It consists of four phases: triangulation, normalization, voxelization of the 3D protein structures, and the Spherical Trace Transform applied. As a result, geometry - based descriptors are produced, which are completely rotation invariant.

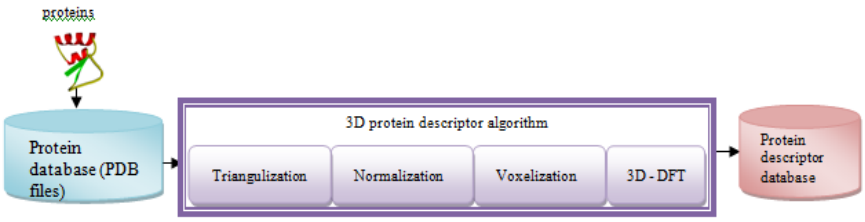


Fig. 2. Protein descriptor generation process

After the protein descriptor process is the process of forming and training of the decision tree for the protein classifier.

2.2 Hierarchical protein classification scheme - Gene Ontology

The process of protein training classifier uses the Gene Ontology (GO) hierarchical protein classification scheme [1]. Therefore, we give a following explanation of its structure. The Gene Ontology defines a set of terms to which any given protein may be annotated and is structured as a directed acyclic graph. Each term has defined relationships to one or more other terms in the same domain, and sometimes to other domains. In this graph the parent–child relationship among terms implies that the child term is either a special case of the parent term (the IS–A relationship) or describes a process or component that is part of the parent process/component (the PART-OF relationship).

The Gene Ontology project covers three domains: cellular component, molecular function, and biological process. A cellular component is just that, a component of a cell, but it is part of some larger object; this may be an anatomical structure (e.g. rough endoplasmic reticulum or nucleus) or a gene product group (e.g. ribosome, proteasome).

A biological process is series of events accomplished by one or more ordered assemblies of molecular functions. It can be difficult to distinguish between a biological process and a molecular function, but the general rule is that a process must have more than one distinct steps.

Molecular function describes activities, such as catalytic or binding activities, that occur at the molecular level. GO molecular function terms represent activities rather than the entities (molecules or complexes) that perform the actions, and do not specify where or when, or in what context, the action takes place. Molecular functions generally correspond to activities that can be performed by individual proteins, but some activities are performed by assembled complexes of proteins.

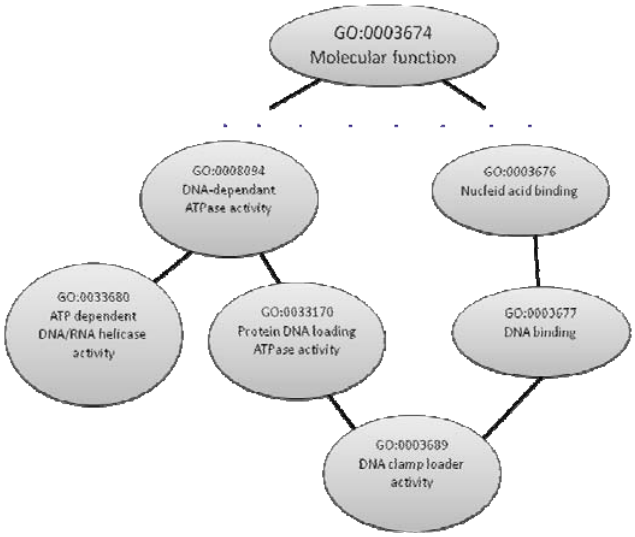


Fig. 3. Representation of the Gene ontology hierarchy

In Fig. 3 a part of the Gene Ontology molecular function domain is presented. Each node in the GO structure represents a protein function which is more specific, but related to the function of the parent nodes to which is connected. Nodes with higher positions in the GO structure represent more general functions, while nodes lower in the structure represent more specific functions, thus, parent GO terms may also be used to describe a protein's function. For example, GO:0003689 in Fig. 3 has two parent terms, GO:0033170 and GO:0003677, which have more general functions than GO:0003689. If a protein is annotated with GO:0003689 then by inference, the protein's function can also be described by the parent nodes (GO:0033170 and GO:0003677 as well as all the linked parent terms). Hence, this protein's function can be described by the set of all GO terms Fig. 3 except for GO:0033680.

In our classifier, we only use the molecular function aspect of Gene Ontology and we consider only protein structures from the Protein Data Bank (PDB) [4] and their annotations.

2.3 Hierarchical multi-level C4.5 algorithm modification

The machine learning algorithm we chose for the classification problem is the decision tree algorithm C4.5. In C4.5 the tree is constructed top down. For each node the attribute is chosen which best classifies the remaining training examples. This is decided by considering the information gain, which is the difference between the entropy of the whole set of remaining training examples and the weighted sum of the entropy of the subsets caused by partitioning on the values of that attribute.Traditional *single-label* classification is concerned with learning from a set of examples that are associated with a single label l from a set of disjoint labels L, $|L|>1$. If $|L|>=2$, then it is called a *multi-class* classification problem.

The problem of classifying proteins in GO hierarchy is a standard hierarchical *multi-class* or *multi-level* classification problem, because a protein (from the Protein Data Bank) can have more than one GO terms annotated (more than one function) associated with itself. By applying the hierarchical multi-label C4.5 classification the end result is one decision tree that can classify new protein structures in branch of the GO hierarchy at once.

For the purposes of the protein classification, we have used and readapted the modification of the C4.5 algorithm [21] for hierarchical multi-label data. In order to automate the GO classification we need to: (1) have information about the Gene Ontology hierarchy (test and store the hierarchy), (2) calculate the entropy, and (3) test the membership of the new protein in the existing hierarchy. We have provided flat text file with the GO hierarchy labels organized in a tree-like manner. The modification of the C4.5 algorithm is made by changing the entropy calculation:

$$entropy(S) = -\sum_{i=1}^{N} p(c_i)\log p(c_i) - \alpha(c_i)\log depth\_of\_DAG(c_i) \quad \textbf{(1)}$$

where $p(c_i)$ = relative frequency of class c_i and $q(c_i) = 1-p(c_i)$.

$\alpha(c_i) = 0$, if $p(c_i) = 0$, or a user-defined constant.

$depth\_of\_DAG(c_i) = 1 + number\ of\ descendants\ of\ class\ c_i$, or the depth of the subtree in the hierarchy below the class c_i.

$\log depth\_of\_DAG(c_i)$ represents the hierarchy depth stream below the class involved.

α is primarily a constant, defined by the user, which allows a weighting to be given to a specific part of the formula. The default value is 1 which means equal weight on the uncertainty in the choice and in the specificity (hierarchy and multi label equal weight). Increasing the value of this constant means having specific classes reported at the expense of homogeneity (hierarchy more important), while decreasing its value would favor more general classes if they made the nodes homogeneous (hierarchy less important than multi label). If the class probability is set to zero also the constant is zero, because if this class is mot used there is no need in transmitting depth information.

Our interest is to present the test and training protein set with the most specific functions in different levels of the GO hierarchy and to get the classified proteins deepest in the hierarchy, which automatically means specifying the most specific functions (classes) and knowing the protein hierarchy automatically above.

3 Experimental Results

All of the experiments were conducted on a PC with a 2.2 GHz Intel Core 2 CPU and 2GB RAM. The GO hierarchy [3] was downloaded and parsed for the molecular function purposes. The protein descriptor generation was made in C++. The hierarchical multi-level modification of the C4.5 algorithm was implemented in C#.NET.

In the training phase, the same data set with 3315 proteins from [14] was used in a way that uniformly represents the database, but using the Gene Ontology hierarchy. Those same proteins were used in order to make a comparison on the same dataset [14] for two different classification schemes: GO and SCOP. We consider that if a protein has a significant place in the SCOP hierarchy, then it would have a similar place in the Gene Ontology hierarchy. However, Gene Ontology hierarchy has 14 levels in depth (in molecular function ontology) and each term can have more than one parent, in comparison to the classification made in [14] where only to the level of Domain (level 5) proteins are classified and each term has only one parent.

In the test phase dataset with 360 proteins was used in a similar manner as in the training phase.

Some experiments regarding the parameter α (the constant that shows how much the hierarchy is important) were made in the hierarchical multi-label classifier. The results are shown on Fig. 4. In the case when the parameter α is equal to 0.9 we get an optimal classification accuracy of 73,6% that decreases if α is increased or decreased, which means that the hierarchical multi-label classifier gives almost equal weight to the hierarchy needed and the multi-label (multi function of one protein) weight. Therefore, the hierarchical structure is equally important as giving more GO

annotation to a protein. This accuracy percent is satisfactory, knowing the fact that 14 levels of hierarchy are used. When α approaches zero hierarchy is not considered, and the classifier gives its worst accuracy of only 53.4%. In addition, using the hierarchical structure lowers the computational cost in the protein classification process.

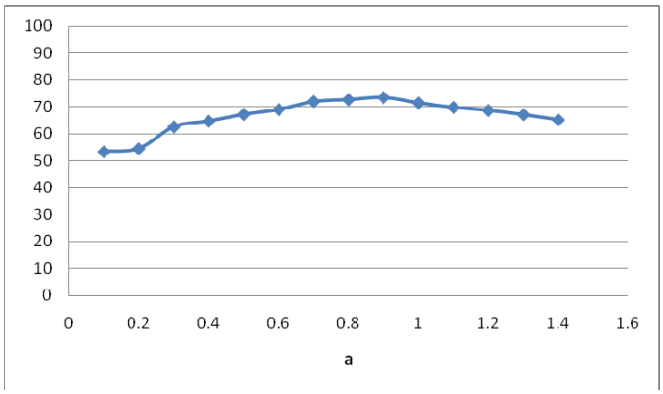


Fig. 4. Results of the hierarchical multi-label C4.5 classifier using GO hierarchical structure.

In Table 1. are given the results from the classification with the multi-label modification in [14]. Although the results in Table 1. go up to 85.29% for the first level of hierachy, they are not taking the hierarchical structure of GO, rather than SCOP with only 5 levels of hierarchy, but only on separate levels. For more levels of hierarchy as in Gene Ontology, where a hierarchical multi-label classification problem is considered, results are not showed in [14].

Table 1. Results of the classification with the multi-level modification of C4.5 in [14].

Level	Accuracy
CLASS	85,29%
FOLD	82,69%
SUPERFAMILY	80,09%
FAMILY	79,38%
DOMAIN	78,88%

4 Conclusion

The aim of this paper was to use the hierarchical structure of a protein database such as the Gene Ontology to help in the protein classification process. We proposed a hierarchical multi-label classifier based on a modification of the standard C4.5 decision trees classification algorithm and a 3D protein descriptor. This classification method can be used even with other hierarchical protein database classification

schemes. The results show that the hierarchical organization structure is important and betters the classifier.

As future work, we can consider methods for changing the protein dataset in order to match a representation of the Gene Ontology structure for improving the classifier's accuracy.

References

1. Smith,, B., Williams, J. and Schulze-Kremer, S.: The ontology of the Gene Ontology. In: Proceedings of AIMA Simposuim (2003)
2. Murzin, A.G., Brenner, S.E., Hubbard, T., Chothia, C.: Scop: a structural classification of proteins database for the investigation of sequences and structures. Journal of Molecular Biology, vol. 247, pp. 536--540 (1995)
3. The Gene Ontology Consortium: http://www.geneontology.org/ [Accessed June 2010]
4. Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., Bourne, P.E.: The Protein Data Bank. Nucleic Acids Res., vol.28, pp. 235--242 (2000)
5. Shindyalov, H.N., Bourne, P.E.: Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. Protein Engineering, vol. 9 , pp. 739--747 (1998)
6. Ortiz, A.R., Strauss, C.E., Olmea, O.: Mammoth: An automated method for model comparison. Protein Science, vol. 11, pp. 2606--2621 (2002)
7. Holm, L., Sander, C.: Protein structure comparison by alignment of distance matrices. Journal of Molecular Biology, vol. 233, pp. 123--138 (1993)
8. S. Khan, G. Situ, K. Decker, and C. Schmidt. GoFigure: Automated Gene Ontology Annotation. *Bioinformatics*, 19(18) pp. 2484–2485 (2003).
9. D. Martin, M. Berriman, and G. Barton. GOtcha: a New Method for Prediction of Protein Function Assessed by the Annotation of Seven Genomes. *BMC Bioinformatics*, 5(1):178 (2004)
10. A. Vinayagam, R. Konig, J. Moormann, F. Schubert, R. Eils, K. Glatting, and S. Suhai. Applying support vector machine for gene ontology based gene function prediction. *BMC Bioinformatics* (2004)
11. R. Eisner, B. Poulin, D. Szafron, P. Lu, and R. Greiner. Improving protein function prediction using the hierarchical structure of the gene ontology. In *CIBCB*, pp. 354–363 (2005)
12. O. D. King, R. E. Foulger, S. S. Dwight, J. V. White, and F. P. Roth. Predicting Gene Function From Patterns of Annotation. *Genome Research*, pp. 896–904 (2003)
13. Kalajdziski, S., Mirceva, G., Trivodaliev, K., Davcev D.: Protein Classification by Matching 3D Structures. In: Frontiers in the Convergence of Bioscience and Information Technologies 2007, pp. 147--152. Jeju Island, Korea (2007)
14. Kalajdziski, S., Pepik, B. Ivanoska, I., Mirceva, G., Trivodaliev, K., Davcev D.: Automated Structural Classification of proteins by using Decision Trees and Structural Protein Features. In: ICT Inovations, Springer, Macedonia (2009)
15. Holm, L., Sander, C.: The FSSP Database: Fold Classification Based on Structure-Structure Alignment of Proteins. Nucleic Acids Research, vol. 24, pp.206--210 (1996)
16. Orengo, C.A., Michie, A.D., Jones, D.T., Swindells, M.B., Thornton, J.M.: CATH - A hierarchic classif. of protein domain structures. Structure, vol. 5(8), pp. 1093--1108 (1997)
17. Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., Lipman. D.J.: Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Research, vol. 25(17), pp: 3389--3402 (1997)
18. Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J.: Basic local alignment search tool. In: Journal of Molecular Biology, vol. 215(3), pp: 403--410 (1990)

19. Needleman, S.B., Wunsch, C.D.: A general method applicable to the search for similarities in the amino acid sequence of two proteins. In: J. Mol. Bio., vol. 48(3), pp: 443--453 (1970)
20. Mewes, H.W., Neumann K.: MIPS: a database for genomes and protein sequences. In: Nucleid Acid Research, vol. 27(1), pp:44-48 (1999)
21. Clare, A.: Machine learning and data mining for yeast functional genomics. PhD thesis, University of Wales Aberystwyth (2003)

Machine Learning Approach for Early Detection of Cardiovascular Deceases (CVD)

Saško Ristov and Aleksandar Pečkov

Institute of Informatics, Faculty of Natural Science, Ss. Cyril and Methodius University, 1000 Skopje, Republic of Macedonia
Sasko@ii.edu.mk,apechkov@gmail.com

Abstract. Cardiovascular deceases are leading cause for mortality in the world. More people are dying from these deceases rather than all other deceases. Health insurance fund in Macedonia defined a method for early detection and prevention of CVD according to risk assessment for patients CVD ailment. We analyze this method, and using machine learning algorithms, as well as comparison of international recommendations, we propose a new improved method for CVD Risk management, such as: smaller target population for better patient management, better risk factor classification and correlation, definition of high (and low) risk factors, treatment recommendations and goals.

Keywords: Machine Learning, Cardiovascular Deceases Risc Factors, Data Mining

1 Introduction

Cardiovascular deceases are chronicle deceases which are one of a leading cause for people mortality in Macedonia, as well as the whole world. Most cases are often undiagnosed for a long time, and in the moment of appearing the first manifestations, they appear as urgent acute coronary conditions. These manifestations appear unexpected quickly and in many cases with fatal end even before can any medical treatment implement.

With early detection of potential risk factors, as well as their in time mutation, the urgent cardiovascular occurrences can be substantially prevented. CVD risk factors [2] that can be prevented or treated include: high blood pressure, high cholesterol, excess weight, physical inactivity, smoking, diabetes, illegal drug use and stress. Unpreventable risks include: previous heart attack, family member with heart disease, increasing age, gender and race.

Because of these reasons, health insurance fund in Macedonia introduced systematic approach to early patient detection with high risk of CVD and in time risk factor treatment.

1.1 Health Insurance Fund Data model

Health insurance fund introduced a model [1] for risk assessment of CVD for early detection and prevention starting 2008. The scope now is to coverage 15%

from whole population between 25[1] and 65 years old yearly with patient personal doctor to make all the necessary analysis and inform patient of his CVD risks.

Eleven risk factors are taken in risk assessment: gender, age, BMI[2], smoking history, systole blood pressure (SBP), cholesterol in blood, serum creatinine, previously heart attack, previously cerebral insultus, and left ventricular hypertrophy determined with ECG[3], and sugar decease. The patients primary doctor determines these factors precisely, and according patient gender, uses table from fig. 1 for men, and table from fig. 2 for women to calculate the patient points to determine risk of CVD, as one of Low level risk, Medium Risk, High Level Risk or Very High Level Risk.

Risk Factor	Points										
Age (years)	25-39	40-44	45-49	50-54	55-59	60-64	65-69	70-74			
points	**0**	**+4**	**+7**	**+11**	**+14**	**+18**	**+22**	**+25**			
Smoking History (according age)	**+9**	**+7**	**+7**	**+6**	**+6**	**+5**	**+4**	**+4**			
SBP (mmHg)	10-19	20-29	30-39	40-49	50-59	60-69	70-79	80-89	90-99	200-209	>210
points	**0**	**+1**	**+2**	**+3**	**+4**	**+5**	**+6**	**+7**	**+8**	**+9**	**+11**
Cholesterol in blood (mmol/L)	<5		5.0-5.9		6.0-6.9		7.0-7.9		8.0-8.9		>9
points	**0**		**+2**		**+4**		**+5**		**+7**		**+9**
BMI (in TTkg/TBm2)	<15		15-19		20-24		25-29		30-34	>35	
points	**0**		**0**		**1**		**+2**		**+3**	**+4**	
Serum Creatin.	<70	70-79		80-89	90-99	100-109		110-119	120-129		>130
points	**0**	**+1**		**+1**	**+2**	**+2**		**+3**	**+3**		**+4**
Male Gender	**+12 points**										
Previously Heart Attack	No – 0 points, Yes - +8 points										
Prev. Cerebral Insultus	No – 0 points, Yes - +8 points										
Left Ventricular Hypertrophy	No – 0 points, Yes - +3 points										
Sugar Decease	No – 0 points, Yes - +2 points										

Fig. 1. Points for CVD risk factors for men.

After calculating sum of points for the patient, the doctor determines the risk factor using table from fig. 3 for men and table from fig. 4 for women.

[1] Initially, the model was to assessment all the patients older than 18 years.
[2] Body Mass Index
[3] Electrocardiogram

Risk Factor	Points										
Age (years)	25-39	40-44	45-49	50-54	55-59	60-64	65-69	70-74			
points	**0**	**+5**	**+9**	**+14**	**+18**	**+23**	**+27**	**+32**			
Smoking History (according age)	**+13**	**+12**	**+11**	**+10**	**+10**	**+9**	**+9**	**+8**			
SBP (mmHg)	10-19	20-29	30-39	40-49	50-59	60-69	70-79	80-89	90-99	200-209	>210
points	**0**	**+1**	**+2**	**+3**	**+4**	**+5**	**+6**	**+7**	**+8**	**+9**	**+11**
Cholesterol in blood (mmol/L)	<5	5.0-5.9		6.0-6.9		7.0-7.9		8.0-8.9		>9	
points	**0**	**+0**		**+1**		**+1**		**+2**		**+2**	
BMI (in TTkg/TBm2)	<15	15-19		20-24		25-29		30-34		>35	
points	**0**	**0**		**1**		**+2**		**+3**		**+4**	
Serum Creatin.	<50	50-59		60-69		70-79		80-89	90-99	100-109	>110
points	**0**	**+1**		**+1**		**+2**		**+2**	**+3**	**+3**	**+4**
Previously Heart Attack	No – 0 points, Yes - +8 points										
Prev. Cerebral Insultus	No – 0 points, Yes - +8 points										
Left Ventricular Hypertrophy	No – 0 points, Yes - +3 points										
Sugar Decease	No – 0 points, Yes - +9 points										

Fig. 2. Points for CVD risk factors for women.

2 Machine Learning Approach

A major focus of machine learning research is to automatically produce (induce) models, such as rules and patterns, from massive data. Machine learning is much related to fields such as "data mining, statistics, inductive reasoning, pattern recognition, and theoretical computer science" and data warehouses.

Data mining techniques are able to identify the high risk patients, to define the most important variables in cardiovascular patients, but, in the same time, they have the capacity to build a model in order to distinguish, in a simple and understandable way, the relationships between any two variables.

The purpose of present study is to compare the capacity of different data mining methods to evaluate and quantify the relationships among cardiovascular risk factors and cardiovascular disease, differently by many factors of patients.

2.1 Software tool - WEKA

Weka [5] [11] is a collection of machine learning algorithms for data mining tasks. The algorithms can either be applied directly to a dataset or called from your own Java code. Weka contains tools for data pre-processing, classification,

Age	Low Level	Medium Level	High Level	Very High Level
25-39	<23	23-30	30-34	>34
40-44	<27	27-33	33-37	>37
45-49	<31	31-36	36-40	>40
50-54	<35	35-40	40-43	>43
55-59	<38	38-43	43-47	>47
60-64	<42	42-46	46-50	>50
65-69	<46	46-50	50-53	>53
70-74	<49	49-52	52-55	>55

Fig. 3. Determination patient risk of CVD according points for men.

Age	Low Level	Medium Level	High Level	Very High Level
25-39	<9	9-14	14-23	>23
40-44	<14	14-19	19-27	>27
45-49	<18	18-24	24-32	>32
50-54	<23	23-29	29-36	>36
55-59	<29	29-34	34-40	>40
60-64	<33	33-38	38-43	>43
65-69	<38	38-42	42-46	>46
70-74	<42	42-45	45-48	>48

Fig. 4. Determination patient risk of CVD according points for women.

regression, clustering, association rules, and visualization. It is also well-suited for developing new machine learning schemes. It was written in Java and developed by the University of Waikato.

2.2 Sample Data and Data Model

We analyze data of around 6200 patients older than 18 years in a period of 2008-2010 year registered in one Private Health Medical Primary Institution. In 2.5 years period, 1682 CVD tests are made on different patients, including results of a risk according to Macedonian Health Insurance Fund Data Model. According to gender, 988 are women, and 694 men. 194 patients are smokers, 40 patients had previously heart attack, 54 had previously cerebral insultus, 62 had left ventricular hypertrophy and 224 had Sugar Decease. We use all risk factors previously mentioned, as an input parameters, including factor of already diagnosed cardiovascular deceases according to [3][4]. 861 Patients (53.75%) were diagnosed with diagnose of circulatory system diseases within the same period as the test shown on fig. 5 as second column with red color.

[4] Current version of International Classification of Diseases

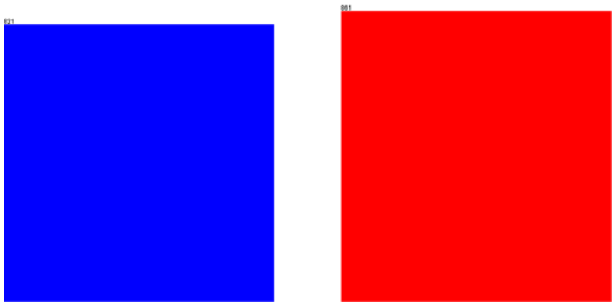


Fig. 5. Health and Diagnosed subject distribution.

3 Machine Learning Algorithms

Machine learning algorithms are used for different applications. The purpose of machine learning algorithms is to use observations (experiences, data, patterns) to improve a performance element, which determines how the agent reacts when it is given particular inputs. The performance element may be a simple classifier trying to classify an input instance into a set of categories or it may be a complete agent acting in an unknown environment. By receiving feedback on the performance, the learning algorithm adapts the performance element to enhance its capabilities.

- **Linear Regression** [9] refers to any approach of modeling the relationship among variables such that the model depends linearly on the unknown parameters to be estimated from the data. Numerous procedures have been developed for parameter estimation in linear regression.
- **M5P** [16] is a decision tree algorithm that models the data as a tree that has linear models in the nodes.
- **Naive Bayes** A Bayes classifier is a simple probabilistic classifier based on applying Bayes' theorem (from Bayesian statistics) with strong (naive) independence assumptions.
- **J48** [14] implements Quinlans C4.5 algorithm for generating a pruned or unpruned C4.5 decision tree. The decision trees generated by J48 are used for classification.
- Ciper [10] [6] is a machine learning algorithm for finding polynomial equations.
- **Classification Via Regression** [15]. The class whose regression model gives the highest value is chosen as the predicted class.

4 Machine Learned Models

First we considered a simpler classifiers and continue to the more complex ones.

– One of the simplest models, a Naive Bayes Classifier has 1215 (72.2354 %) correctly classified instances. If we consider the fact that the attributes are not much correlated we infer why this simple classifier is performing so good.
– Next on the list is classification via linear regression. The resulted model is:

$$-0.013*A-0.090*S1-0.007*P-0.011*BMI0.349*PHA00.063*G11.998$$

$$0.013*A-0.090*S00.007*P0.011*BMI0.349*PHA10.063*G0-1.320$$

for bigger risk, where A stands for age, $S1$ is one if the subject is a smoker, $S0$ if not, P is blood pressure, PHA is the attribute that determines if the patient had previous heart attack, and G determines the gender of the patient.
Results: The older subject has more probability for CVD. We see that it is positively correlated if he is a smoker or not. We also find the correlation about subject gender. If the subject had previous hearth attach, then the risk for CVD is even greater. This model correctly classified 1228 (73.0083 %) instances.
– The next model we are going to try is a decision tree build with J48. The resulting model is given below.

```
AGE <= 49
|   PRESSURE <= 126
|   |   AGE <= 39: 0 (297.0/31.0)
|   |   AGE > 39
|   |   |   SMOKER = 0: 0 (167.0/40.0)
|   |   |   SMOKER = 1
|   |   |   |   BMI <= 23.5: 1 (11.0/2.0)
|   |   |   |   BMI > 23.5: 0 (34.0/11.0)
|   PRESSURE > 126
|   |   PRESSURE <= 135
|   |   |   GENDER = 0
|   |   |   |   BMI <= 24.5: 1 (23.0/8.0)
|   |   |   |   BMI > 24.5: 0 (56.0/23.0)
|   |   |   GENDER = 1: 0 (83.0/24.0)
|   |   PRESSURE > 135
|   |   |   SERUM CREATIN <= 82: 1 (47.0/11.0)
|   |   |   SERUM CREATIN > 82: 0 (17.0/6.0)
AGE > 49
|   PREVIOUS HEARTH ATTACK = 0
|   |   PRESSURE <= 130
|   |   |   AGE <= 56
|   |   |   |   BMI <= 22.5: 0 (24.0/4.0)
|   |   |   |   BMI > 22.5
|   |   |   |   |   SMOKER = 0
|   |   |   |   |   |   PRESSURE <= 110: 0 (16.0/4.0)
|   |   |   |   |   |   PRESSURE > 110
```

```
|   |   |   |   |   |   |   GENDER = 0
|   |   |   |   |   |   |   |   AGE <= 54: 1 (47.0/19.0)
|   |   |   |   |   |   |   |   AGE > 54: 0 (19.0/6.0)
|   |   |   |   |   |   |   GENDER = 1
|   |   |   |   |   |   |   |   SERUM CREATIN <= 69: 0 (12.0/2.0)
|   |   |   |   |   |   |   |   SERUM CREATIN > 69
|   |   |   |   |   |   |   |   |   PRESSURE <= 120: 1 (10.0/2.0)
|   |   |   |   |   |   |   |   |   PRESSURE > 120
|   |   |   |   |   |   |   |   |   |   BMI <= 27.5: 0 (20.0/8.0)
|   |   |   |   |   |   |   |   |   |   BMI > 27.5: 1 (11.0/3.0)
|   |   |   |   |   SMOKER = 1: 1 (26.0/9.0)
|   |   |   AGE > 56: 1 (359.0/109.0)
|   |   PRESSURE > 130: 1 (368.0/70.0)
|   PREVIOUS HEARTH ATTACK = 1: 1 (35.0/2.0)
```

This models correctly classifies 1207 (71.7598 %) instances.
Results: Clearly we can extract some rules from it. For example, a simple rule is if the subject is over 49 years old and have high blood pressure (over 130), he has 5 in 6 chance of getting a CVD.
From a medical point of view, we examined the patients with in the most similar group. For example, Patients (total 297) younger than 39 years and systole blood pressure below 126 weren't diagnosed CVD. They look pretty young and healthy. Only 31 of them are categorized incorrectly.
Another example is group of patients, between 39 and 49 years (total 167), with systole blood pressure below 126, but no smokers, weren't diagnosed CVD. Here we also see some correlation among the parameters. Some parameter values for age and systole blood pressure rely on smoking history and then BMI.
Next huge group creates the patients older than 49 years, (368 patients), with systole blood pressure above 130, are very risk patients of CVD. Maybe in the same group are patients older than 56 years, no matter how much their blood pressure is.
We can conclude that the pressure is correlated with age between 49-56 years.

- Next, we tried classification via model trees. This models correctly classifies 1223 (72.7111 %) instances. Because of the complexity of the model we chose not to write it here. The model has two trees one for each class (0 and 1), that have linear models in them leaves. We can say that the models are similar with the model obtained with a classification via linear regression. Notice that that model is much simpler and a little bit more accurate (1228 correctly classified instances).
- We tried also support vector machines [12] but, they are impossible to interpret and also, it turned out that they are not more accurate as we expected. We tried more kernels and other different parameters and the best result we could get is 1229 (73.0678 %) correctly classified instances. So still the basics linear regression classifier has good performance. We tried also bagging and ensembles but there was no improvement.

– Finally we tried Ciper. The algorithm produces polynomial or rational functions for models. Its best variant produces a set of rational models that classify 1230 instances correctly.
  Results: A simpler variant of the algorithm produced the following model:

  $$-1.1051+0.0075 \cdot Pressure+0.0024 \cdot Pressure \cdot PreviousHearthAttack+0.012 \cdot Age$$

  From the model we can conclude that the blood pressure and the age of the subject are positively correlated. Also in the cases when the subject had previous heart attack, then the blood pressure is even more relevant.

For the purpose of testing the algorithms for better predictive capabilities we used a paired t-test [9]. The number of correctly classified instances on the folds was used in the test for statistical difference and determination of algorithms that have better predictive capabilities. We found that

– Naive Bayes has statistically worse performance then linear regression (p-value 0.051) and support vector machines (p-value 0.0477)
– J48 has statistically worse performance then linear regression (p-value 0.044) and support vector machines (p-value 0.0477)
– Ciper, SMO, LR, and M5P have statistically equal performance. This means that we cannot choose one which is the best from all of them.

If we just count the number of correctly classified instances then the ordering is Ciper, SMO, LR, M5P and in the end NB. But statistically the predictive capabilities for this domain, of Ciper, SMO, LR and M5P is the indifferent. If we would like to choose a simple model then classification via linear regression is the best alternative.

5 Conclusion and further works

We analyzed the existing Macedonia Health insurance fund model on the sample data, and experiment machine learning algorithms on the same data.

Our study purpose has three goals:

– Existing Health Insurance Fund Model Deficiencies
– New improved method for Early detection CVD

5.1 Existing Health Insurance Fund Model Deficiencies

We found several deficiencies:

Huge Disproportion of low and medium level with subjets already diagnozed CVD. After data analysis on existing Macedonia Health insurance fund model, and subject's diagnoses of circulatory system diseases according to ICD-10, we found a huge disproportion. In fact on 610 of 1302 low level risk patients has been diagnosed CVD minimum once, which is almost 47%. medium risk patients has diagnosed CVD. 66 of 112 (59%) high level risk patients has

been diagnosed CVD as shown on fig. 6 where columns from left to right are Low level risk, Medium Risk, High Level Risk and Very High Level Risk, and Blue color are health subjects, and red are diagnosed subjects. This means that exactly half of patients with low or normal level of risk were diagnosed CVD within the same year as CVD tests, which should classify them in the higher level risk factor, as those patients are already diseased of CVD [2].

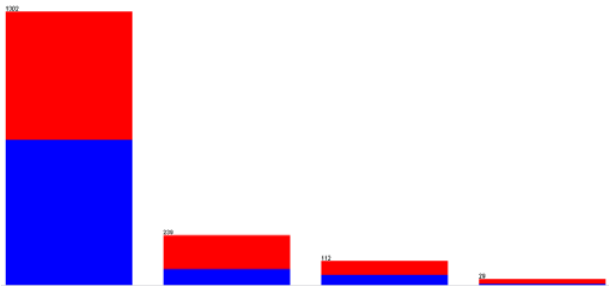


Fig. 6. Risk Level Distribution accordind Health insurance fund model.

Not clearly defined Target population for risk assessment. In the first 2 years, target population was above 18, now it is changed as 25-65 years old.

Complex method. The method is very complex, and is time consuming. **Other Deficiencies**. There are other identified international CVD risk factors [2] which are not included, such as low blood pressure, patient lifestyle, menopausal at women, etc.

5.2 New improved method for Early detection of CVD

In this study we propose new method for early detection of CVD analyzing machine learning methods in the sample data with more than 70% correctness, and using international CVD risk factors.

Our new method propose:

Target population. We propose to test all population above 40 years, as well as all smokers, and women who are postmenopausal (mostly over 50). This will reduce costs for test, and will provide better patients management.

Definition of high risk factors. All machine learning algorithms identified: Age over 50, High blood pressure above 135, Smoking, and previously CVD as high level risk factors.

Definition of low risk factors. All machine learning algorithms identified that high cholesterol and diabetes are low level CVD risk factors.

Treatment recommendations. We propose to treatment all patients over 50 years with High blood pressure goal in maximum 130. We also propose smoking cessation.

5.3 Further work

We used several Machine Learning models, which offered better results than the existing Health Insurance Fund Model. But, although they have a huge correctness, they are not 100% correct.

We should use other parameters which affects as CVD risk factors defined in [2].

Define New Risk Calculator. The sample data we used is from one town and small number which implies to use a huge various sample data to define an improved risk calculator.

References

1. Macedonian Health insurance fund Model. http://www.fzo.org.mk/ (*CVD Prevention)*
2. *Cardiovascular Risk Management*. Edited by F.D. Richard Hobbs and Bruce Arroll 2009 Hobbs FDR & Arroll B. ISBN: 978-1-405-15575-5
3. World Health Organization, ICD-10. http://www.who.int/classifications/icd/en/
4. R. Bayardo. Constraints in data mining. *SIGKDD Explorations*, 4(1), 2002.
5. I. H. Witten and E. Frank. *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*. Morgan Kaufmann, San Mateo, CA, 1999.
6. Pečkov Aleksandar, Džeroski Sašo, Todorovski Ljupčo *A Minimal Description Length Scheme for Polynomial Regression*. PAKDD 2008: 284-295
7. Breiman, L., Friedman, J. H., Olshen, R. A., Stone, C. J. (1984). *Classification and Regression Trees*. Wadsworth International, Belmont, Ca.
8. Grünwald, P., Myung, I., & Pitt, M. (Eds.). (2005). *Advances in minimum description length: Theory and applications*. Cambridge, Massachusetts: MIT Press.
9. Hastie, T., Tibshirani, R., & Friedman, J. (2001). *The elements of statistical learning*. New York: Springer.
10. Todorovski, L., Ljubič, P., & Džeroski, S. (2004). *Inducing polynomial equations for regression. Proceedings of the Fifteenth International Conference on Machine Learning* (pp. 441–452). Banff, Alberta, Canada: ACM.
11. Witten, I. H., & Frank, E. (Eds.). (2005). *Data mining: Practical machine learning tools and techniques*. San Francisco: Morgan Kaufmann.
12. Platt, J. (1998). *Machines using Sequential Minimal Optimization.* B. Schoelkopf and C. Burges and A. Smola, editors, *Advances in Kernel Methods - Support Vector Learning.*
13. *George H. John, Pat Langley, (1995).* Estimating Continuous Distributions in Bayesian Classifiers Eleventh Conference on Uncertainty in Artificial Intelligence, *San Mateo, (pp. 338-345)*
14. *Ross Quinlan (1993).* C4.5: Programs for Machine Learning. *Morgan Kaufmann Publishers, San Mateo, CA.*
15. *E. Frank, Y. Wang, S. Inglis, G. Holmes, I.H. Witten (1998).* Using model trees for classification. Machine Learning. *32(1): (pp 63-76)*
16. *Y. Wang, I. H. Witten, 1997* Induction of model trees for predicting continuous classes. Poster papers of the 9th European Conference on Machine Learning.

3-Axial Accelerometers Activity Recognition

Hristijan Gjoreski[1], Matjaz Gams[2], Ivan Chorbev[1]

[1] Department of Informatics and Computer Engineering, Faculty of Electrical Engineering and Information Technologies, PO BOX 574, Skopje, R. of Macedonia
[2] Jožef Stefan Institute, Dept. of Intelligent Systems, Jamova 39, 1000 Ljubljana, Slovenia
{hristijan.gjoreski,matjaz.gams}@ijs.si,
ivan@feit.ukim.edu.mk

Abstract. Activity Recognition is a typical classification problem. The goal is to detect and recognize everyday activities of a person. This paper presents our approach to measurements and classification of a person's movements. This is done by using two 3-axis radio accelerometers attached to the person's body and by reconstruction and interpretation of the user's behavior. We compared two machine learning algorithms (J48 and Random Forest) and various attributes characterizing the user's behavior in order to obtain accurate classification of the behavior into predefined classes - activities.

Keywords: Activity Recognition, 3-axial Accelerometers, Attributes, Activities, Instances, Machine Learning, Classifiers.

1 Introduction

Activity Recognition is becoming a very popular research topic. There are a lot of projects and useful applications being build, aiming at nursing older people [6, 8]. Most of them belong to the area of ambient assisted living and are trying to make the everyday life easier, simpler and safer. The essential part in these applications is the activity recognition module. There are applications that aim to recognize activities and find macro and micro level irregularities. Macro level covers more general and longer periods of activities like recognizing the pattern of walking and alarming if something is wrong with the walking signature. On the other hand, micro level activities represent every single movement and gesture of the person. Recognizing these activities is a very interesting and challenging field, and therefore they are the goal of our research.

Activity recognition can be formulated as machine learning, classification problem. In the narrow sense it can be seen as a pattern recognition problem [1, 2, 5]. Therefore good attributes that describe the person's behavior are crucial for the classification techniques to be applied. There are lots of projects using different attributes

(features), starting from traditional attributes like: mean, variance, correlation [1, 3, 4] to more complex that transform the data in frequency domain [2].

Our goal was to detect and recognize every single movement the person makes. Other researchers are using windows, window sizes and overlapping windows of the data [1, 2, 5]. With this approach, every instance cannot be classified, but a group of instances (one window) are classified together. The disadvantage of this approach is that short activities like going down and standing up cannot be recognized.

For the needs of the research described in this paper, two 3D accelerometers were used. A 3-axial accelerometer is a sensor that provides the projections of the acceleration vector along three axes: x, y and z. The 3 projections and the timestamp are the raw data that we get from the sensors. Using the raw data we extracted some features that are later used for the machine learning and the classification techniques. Some of the attributes are the traditional features that are used for acceleration activity recognition like: mean, standard deviation, root mean square, etc., but there are also some features that measure the direction and the movement of the person.

We tried to distinguish and recognize 7 activities: *Standing, Sitting, Lying, On all fours, Sitting on the ground, Standing up,* and *Going down*. We decided to include walking into *standing*, so *standing* can be a static or a dynamic activity. Some very quick activities like *going down* and *standing up* were also included. These two activities are very difficult to separate and recognize using just acceleration sensors, because they happen very fast and we don't have as many instances as we have for *lying* for example.

The classification techniques used for this research are: J48 (Weka's implementation of C4.5) and Random Forest. These were the techniques yielding the best results, after analyzing several options in the open source library Weka.

2 Data Collection

The data needed for training the classifiers was retrieved using two 3-axial accelerometers. One of them was positioned on the chest and the other on the right ankle of a person. It is important to state that the same results would have been achieved even if the second accelerometer was placed at the left ankle instead of the right ankle.

The raw data consisted of the following attributes: timestamp, acceleration along x axis, acceleration along y axis and acceleration along z axis. The data in the raw state is insufficient in the recognition process, so we calculated and extracted other attributes that can help in the classification.

The accelerometers frequency of sending data was 5Hz, so 5 measurements per second were obtained. Other researchers try to combine the data and use windows to recognize the activities [1, 2, 5]. With this approach every single data (instance) measured cannot be recognized and classified. Instead, this approach classifies bigger windows of instances (ex. 50 instances together as one action). The advantage is that noise is eliminated, but on the other hand, information is lost. From our point of view this method can be very useful when the aim of the project is recognizing longer activities. We couldn't use it because we are also interested in short activities like

going down and *standing up*. With our approach we tried to recognize every instance that was coming from the accelerometer. Perhaps this is a more difficult task, because we tried to detect every single instance, even if they are very quick and do not last long (ex. *going down*). In fact, that was the aim of this research. We collected the data in order to recognize these 7 activities:

1. *Standing*
2. *Lying*
3. *Sitting*
4. *On all fours*
5. *Sitting on the ground*
6. *Going down*
7. *Standing up*

The data that we collected was separated in two scenarios. Each of them contained all seven activities. The first scenario was longer and had around 7000 instances and was used for building the classification model. The second one was shorter and was used for testing and analyzing the results. The process of collecting the data, building the model and classification is presented in Figure 1.

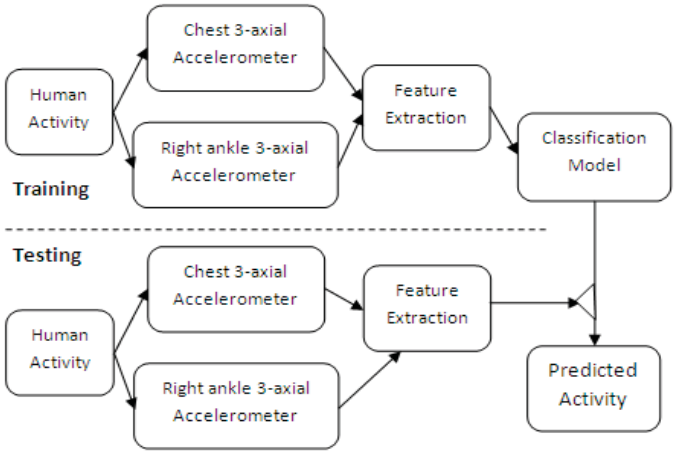


Figure 1. The activity recognition process used in this research.

3 Feature Extraction

Once raw data was obtained, 14 features (attributes) were extracted from it. We used both the 6 raw data attributes (3 attributes per accelerometer) and the extracted features in our classification procedure. Eventually, before building the model, we summed up to 20 features that describe each activity. Some of the features that we extracted from the raw data are the traditional features used in activity recognition, like: mean [1, 7], standard deviation [1, 3, 7], root mean square.

When calculating the mean we used 7 acceleration values (instances). Since from the accelerometers we obtained 5 acceleration values in one second (5 Hz), 7 values corresponds to 1.4 seconds worth of activities. The value 7 was chosen so that the

delay in the data is minimized, but the necessary information is still present. The attribute is good for distinguishing long lasting static activities from fast and transitional activities. We used four mean attributes in the model: mean value of the acceleration along 3 axes (x, y, and z) and mean value of the length of the acceleration vector. Very similarly, root mean square and standard deviation are calculated too. The only difference is in calculating the standard deviation attribute. After analyzing this feature we decided just to use the standard deviation for the length of the acceleration vector.

The next two attributes that we used are the lengths of the acceleration vectors for the two accelerometers. This is an attribute very simple for computation, but very useful for machine learning. It is used in the computations of some of the other attributes.

Another very useful feature is the direction of the acceleration. In fact, the direction of the acceleration is the angle between the acceleration vector and one of the axes.

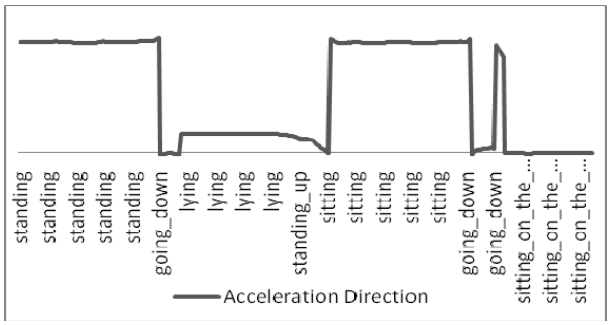


Figure 2. The direction of the acceleration vector

Using this attribute it is very easy to distinguish between activities that have different direction of acceleration (lying on the bed and standing, or lying and sitting on the ground). The attribute and its values on different activities are shown in Figure 2.This attribute is simply computed as an angle between two vectors:

$$\cos\alpha = \frac{acc\_y}{\sqrt{acc\_x^2 + acc\_y^2 + acc\_z^2}} \quad (1)$$

Another feature that is used only for the chest accelerometer is the *change of differences of the lengths of the acceleration vectors*. The mathematical definition of this attribute is:

$$\sum_{i=0}^{n} |len_{i+1} - len_i| \quad (2)$$

This attribute is good in detecting the movement of the person. The value of *n* is chosen to be 15. That means that 15 past instances are used when we decide if the person is moving or not. If we take in consideration that the hardware sending data frequency is 5 Hz (5 instances per second), the conclusion is that we use the last 3 seconds of the person's measured activities. The value is chosen to be 15 after series of tests and analyzing the results.

We also calculated an additional boolean (true/false) attribute that completely uses the value of the *change of differences of the lengths of the acceleration vectors* attribute. Actually we just compared a threshold value to the value of the previous attribute and if it is above the threshold, the boolean attribute will be *true*, otherwise *false.*

4 Building the Models

The activity recognition classification algorithm should recognize the accelerometer signal pattern corresponding to every activity. We should emphasize that every activity has its own pattern that is more or less different than the others [1]. We formulate the activity recognition problem as a classification problem where classes correspond to activities. A test instance is a set of acceleration attributes collected over a time interval and post-processed into a single example instance. The decision, which classifiers to use in our research was made after evaluating the results in the Weka toolkit. We managed to achieve the best results with J48 and the Random Forest classifier. J48 is an open source Java implementation of the C4.5 algorithm in the Weka data mining tool.

When building the model we used a very long scenario containing approximately 7000 instances. We tried to have an equivalent distribution of the class activities. However, some activities are very quick compared to others (ex. standing up vs. lying) and do not happen as often in real life. Therefore it was not possible to have equal number of instances of every activity.

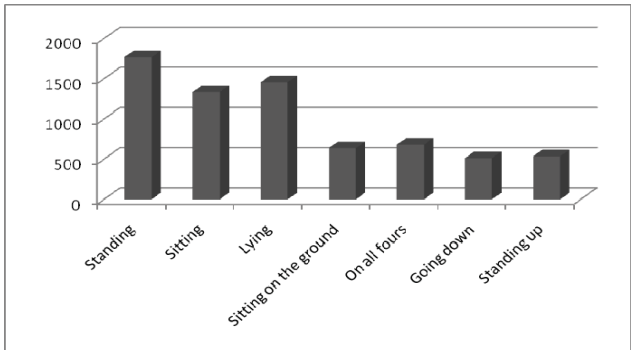


Figure 3. The distribution of the activities instances used in building the classification model.

Naturally, we used a different scenario for testing the model. The difference in the testing scenario was that the activities were shorter and contained fewer instances. The final distribution of the instances is shown in Figure 3.

5 Tests and Results

In the models that we used, there are 20 attributes (6 raw data, 14 extracted) and one class attribute. We should also mention that the class attribute was labeled manually by watching the video/scenario and analyzing the data. That caused the accuracy percentage to have a variation of 2-3%, especially for short activities. Therefore, successive activities (ex. *standing->going down->lying*) probably have some instances misclassified at the beginning and at the end of each activity. We assumed that the best way to present our results is by using confusion matrix and percentage accuracy for each classifier. The results that we got by building the model and testing the data in Weka are shown in *Table 1* and *Table 2*:

Table 1. Confusion Matrix for J48 Model and percentage accuracy

J48		**Predicted Class**							**%**
		Standing	Going d.	Sitting	Stan. up	Lying	Sit on g.	On all 4's	
Actual Class	Standing	**772**	3	2	31	0	0	0	95,5
	Going d.	28	**30**	1	15	9	0	3	34,9
	Sitting	30	6	**42**	42	0	0	0	35
	Stan. up	16	18	3	**75**	10	2	46	44,1
	Lying	0	1	0	0	**442**	0	70	86,2
	Sit on g.	1	0	32	1	0	**170**	1	82,9
	On all 4's	0	2	0	0	0	0	**67**	97,1
Overall accuracy									**81**

Table 2. Confusion Matrix for Random Forest Model and percentage accuracy

Random Forest		**Predicted Class**							**%**
		Standing	Going d.	Sitting	Stan. up	Lying	Sit on g.	On all 4's	
Actual Class	Standing	**786**	0	0	22	0	0	0	97,3
	Going d.	37	**18**	1	16	2	0	15	20,2
	Sitting	26	9	**79**	0	0	0	0	69,3
	Stan. up	10	16	4	**70**	7	3	61	40,9
	Lying	0	4	0	0	**471**	0	40	91,5
	Sit on g.	0	0	16	4	1	**184**	0	89,8
	On all 4's	0	1	0	0	0	0	**68**	98,6
Overall accuracy									**85**

It is obvious that Random Forest gives better results in almost every activity. Actually it is worse only in short and transitional activities like *going down* and *standing up*. Another interesting fact that we can notice is that sitting is very often misclassified with standing, which is because the direction of the accelerometers is almost the same in both activities. As a result 23 % of the sitting instances are misclassified with standing. Similarly, *lying* is misclassified with *on all fours* because when the person is lying on the stomach, the directions of the accelerometers are the same as when he is *on all fours*.

In conclusion we can state that better results are achieved on longer activities (*standing*, *lying*, *sitting on the ground, on all fours*) and worse on transitional, short activities (*going down, standing up*). That is because the accelerations are not that drastic ex. when the person is going down, so it is easy for machine learning to misclassify with other activity (ex. standing - walking). The overall accuracy is 81% with J48 and 85 % with Random Forest and that is a good start for a research in this field, where we try to classify every single instance that we get from the sensors.

6 Future Work

The first issue that needs to be addressed in our future work is the problem that we have while distinguishing *standing* and *sitting*. This problem can be easily solved by putting one more accelerometer on the thigh. By doing this, *standing* and *sitting* will have completely different direction of the acceleration vector and can be easily recognized.

Also adding one more hardware component (not necessarily accelerometer) can improve the results and make the system more efficient.

Another aspect that could be improved is adding detection of falling among the classes being classified. This can be an excellent improvement if our activity recognition is used as a part in an e-health application.

The machine learning algorithms used can be improved, particularly the classifiers. One approach that could be used is combing the best classifiers, in a voting process. Several meta-level classifiers and techniques can be used [1]: bagging, boosting etc.

7 References

1. Nishkam Ravi, Nikhil Dandekar, Preetham Mysore and Michael L. Littman: “Activity Recognition from Accelerometer Data”, USA (2005)
2. Zhenyu He, Lianwen Jin: “Activity Recognition from acceleration data Based on Discrete Consine Transform and SVM”, San Antonio, TX, USA - October (2009).
3. Jennifer R. Kwapisz, Gary M. Weiss, Samuel A. Moore: “Activity Recognition using Cell Phone Accelerometers”, (2010).
4. Cliff Randell Henk Muller: “Context Awareness by Analysing Accelerometer Data”
5. Paula Martiskainen, Mikko JaËrvinen, Jukka-Pekka SkoËn, Jarkko Tiirikainen, Mikko Kolehmainen, Jaakko Mononen: “Cow behaviour pattern recognition using a three-

dimensional accelerometer and support vector machines", Applied Animal Behaviour Science 119 (2009) 32–38.
6. Mitja LUŠTREK, Matjaž GAMS, Igone VÉLEZ: "Posture and movement monitoring for ambient assisted living" (2009)
7. S. Wang, J. Yang, N. Chen, X. Chen and Q. Zhang: "Human Activity Recognition with User-Free Accelerometers in the Sensor Networks", IEEE Int. Conf. Neural Networks and Brain, vol. 2, pp. 1212–1217, (2005).
8. Hein, Albert; Kirste, Thomas: "Activity Recognition for Ambient Assisted Living: Potential and Challenges", Berlin (2008)

A Comparison of Models for Gene Regulatory Networks Inference

Blagoj Ristevski[1], Suzana Loskovska[2]

[1]Department of Information Systems Management
Faculty of Administration and Information Systems Management
St. Kliment Ohridski University – Bitola, Republic of Macedonia
blagoj.ristevski@uklo.edu.mk
[2]Department of Computer Science and Informatics
Faculty of Electrical Engineering and Information Technologies
Ss. Cyril and Methodius University – Skopje, Republic of Macedonia
suze@feit.ukim.edu.mk

Abstract. Gene regulatory networks are complex networks composed of nodes representing genes, transcription factors, microRNAs and other components or modules and their mutual interactions represented by edges. These networks can reveal and depict the fundamental gene regulatory mechanisms in the cells. In this paper we compare the obtained results of gene regulatory networks inference from gene expression microarray data. We have used dynamic Bayesian networks, Boolean networks and graphical Gaussian models as models for network inference. We applied three different size gene expression datasets simulated using a simple autoregressive process. After network inference, we compared the values of the area under ROC curve (AUC) as a validation measure. Some directions for further improved approach for GRNs reconstruction which will include prior knowledge are proposed at the end of this paper.

Keywords: bioinformatics, gene regulatory networks, Bayesian networks, graphical Gaussian models, Boolean networks, area under ROC curve.

1 Introduction

The complex networks composed of genes, proteins and other components regulate the functions and development of the cells through their interactions. The gene regulatory networks (GRNs) provide an understandable view for gene regulatory mechanisms and can uncover the reasons for many diseases. GRNs components are nodes which represent the genes, metabolites, proteins or modules, and edges which correspond to the direct and indirect interactions between nodes. Genes as key components in the GRNs are DNA segments which present fundamental heredity units of every living organism.

The central dogma in molecular biology is presented by two processes: *transcription* and *translation*. In the process of transcription a gene is transcribed into mRNA and after that proteins are produced by translation. When the protein is

synthesized the corresponding coding gene is expressed. The gene expression levels correspond to the approximate number of produced RNA copies from the corresponding gene, which means that gene expression is related to the amount of produced proteins. The microarray technology provides gene expression data as an observation of gene expression under specific experimental conditions or different time points [8].

The inferring or reconstructing of gene regulatory interactions from experimental data is called GRNs inference. There are many models for inferring of GRNs such as Boolean networks, Bayesian networks, dynamic Bayesian networks, graphical Gaussian models, Petri networks, linear and nonlinear differential and difference equations systems, information theory approach, state space models and fuzzy logic models.

The remainder of this paper is organized as follows. In Section 2 we describe Boolean networks. In the third section we present the graphical Gaussian models and their assumptions and usage in the networks inference. We also describe the partial correlation coefficients and their significance for network inference. Bayesian Networks and dynamic Bayesian networks (DBNs) are presented in the following section. The area under ROC curve as a validation measure is described in Section 5. In Section 6 we describe the simulation of artificial gene expression data used for GRNs inference and the obtained inferred networks and the AUC values are shown, too. The concluding remarks are given in the last section.

2 Boolean Networks

Boolean Networks model is a simple model for GRNs inference, consisting of a set of nodes and edges. The nodes represent genes whereas the edges between the nodes correspond to the gene interactions. In Boolean networks, gene expression levels are discretized and presented by two levels states. The genes which have expression levels above a certain threshold are represented by state 1 and the other genes by state 0.

The graph representing a Boolean network gives information about the connection between genes, but it is not sufficient for understanding the all dependencies between genes. The main goal of the reverse engineering in Boolean networks is finding a Boolean function of every gene in the network, so that discretized values of gene expression can be explained by the model. But, the small changes in the gene expression levels cannot be covered by two levels discretization, which leads to information loss. Another shortcoming of Boolean networks is the super-exponential number of all possible networks depending on the number of genes n and it is equal to 2^{2^n}.

REVerse Engineering Algorithm (REVEAL) based on Boolean networks has been introduced by Liang et *al*. (1998) [9]. This algorithm constructs a Boolean network of given expressed gene data by setting the gene in-deegree. If n is the number of nodes and k is the value of in-degree of the genes, then the number of all possible networks can be computed by the Eq. 1:

$$\left(2^{2^k}\frac{n!}{(n-k)!}\right)^n \quad (1)$$

REVEAL extracts minimal network structures using the mutual information approach from the state transition tables of the Boolean network.

3 Graphical Gaussian Models

Graphical Gaussian models (GGMs) are commonly used as a method for GRNs reconstruction based on gene expression data and they are very computationally efficient [3]. GGMs as graphical probabilistic models can identify conditional independence relations among the nodes. They make an assumption that the input gene expression data follow a multivariate Gaussian distribution [6].

The nodes represent genes, and the edges represent conditional dependence relations between nodes. The absence of an edge between two genes means that the corresponding genes are conditionally independent given other genes in the model.

Let Y be the input gene expression data matrix with *G* columns, corresponding to the number of genes, and with *N* rows which correspond to the number of samples (time series data points or other experimental conditions) [3]. It is supposed that matrix Y follows a multivariate normal distribution $N_G(\mu,\Sigma)$, where $\mu=(\mu_1,...,\mu_G)'$ is the mean vector, and $\Sigma=(\sigma_{ij})_{1\le i,j\le G}$ is the positive definite covariance matrix. $\sigma_{ij}=\sigma_i\sigma_j$ are covariance parameters between genes *i* and *j*, and σ_i^2 are related to the variance terms for gene *i*. The estimation of the covariance matrix of the data distribution is a base for the GGMs inference.

First, in the GGM inferencel, to make a reliable estimation of the partial correlation matrix $\widetilde{\mathrm{P}}=(\widetilde{\rho}_{ij})$ is required [4]. This matrix is related to the inverse matrix of the covariance matrix Σ. The straightforward estimator is given by the following Eq. 2:

$$\widetilde{r}_{ij}=-\frac{\hat{\omega}_{ij}}{\sqrt{\hat{\omega}_{ii}\hat{\omega}_{jj}}} \quad (2)$$

where

$$\hat{\Omega}=(\hat{\omega}_{ij})=\hat{\Sigma}^{-1} \quad (3)$$

The partial correlation coefficients $\widetilde{r}_{ij}$, which describe the correlation between nodes/genes Y_i and Y_j conditional dependent on all other nodes in the network, are measures of the direct interactions among nodes/genes [5]. Partial correlation between two genes measures the degree of correlation remaining after removing the effects of the other genes which differs from Pearson correlation coefficients [1] [6].

The above mentioned procedure is appropriate when *N* is larger than the number of genes *G*, otherwise the covariance matrix is not positive-definite and its inverse matrix cannot be found. In microarray data, the sample size *N* is much smaller than the number of genes *G*. For that reason it is suggested to use a shrunk estimate of the covariance matrix. The goal is to construct well conditioned positive definite matrix,

so that the matrix can be inverted. If λ is a shrinkage coefficient so that $0 \le \lambda \le 1$, then shrunk covariance matrix Σ^* is computed by following Eq. 4

$$\Sigma^* = \lambda T + (1-\lambda)S \tag{4}$$

where $\hat{S}$ is the estimated empirical covariance matrix. The shrinkage parameter λ is chosen to minimize the mean-square error and it is determined analytically given by Eq. 5.

$$\lambda^* = \frac{\sum_{i \neq j} \mathrm{var}(r_{ij})}{\sum_{i \neq j} r_{ij}^{\ 2}} \tag{5}$$

After computing the partial correlation coefficients $\tilde{r}_{ij}$, the distribution of $|\tilde{r}_{ij}|$ is checked and the edges with significantly small values of $|\tilde{r}_{ij}|$ are removed from the network [2].

The second stage of the GRNs inference is model selection – assigning statistical significance to the edges from the GGM network.

4 Bayesian Networks

Bayesian networks (BNs) are a special type of graph model defined as a triple (G, F, θ), where G denotes the graph structure, F is the set of conditional probability distributions, and θ is the set of parameters for the graph structure [10]. The structure of the graph G consists of a set of n nodes x_1, x_2, ..., x_n and a set of directed edges between the nodes. The nodes correspond to the random variables and the directed edges show the conditional dependences between the variables (genes).

A directed edge from the node X to the node Y is denoted as $X \rightarrow Y$ which means that X is a parent of Y denoted as $pa(Y)$. Edges and nodes and edges together have to create a directed acyclic graph (DAG).

The joint probability distribution is given by Eq. 6:

$$p(x) = \prod_{i=1}^{n} p\left(x_i \middle| x_{\{1,\cdots,i-1\}}, \theta, G\right) \tag{6}$$

If pa_i denotes the parent nodes of the node x_i which means that the state of each variable x_i depends on the states of its parent pa_i:

$$p(x) = \prod_{i=1}^{n} p\left(x_i \middle| pa_i, \theta, G\right) \tag{7}$$

BNs can deal with noisy and stochastic nature of gene expression data and with incomplete knowledge about the system. The small number of data points (samples) and large number of genes are common problems for BNs learning. Another disadvantage is that feedback loops cannot be captured, although they exist in the GRNs. BNs represent probabilistic relations between genes at the same time and they cannot represent the time relationships between variables

To overcome these drawbacks of BNs, dynamic Bayesian networks (DBNs) are used to model gene regulations. DBNs can deal with stochastic variables, time series gene expression data, feedback loops, missing values, hidden variables and can include prior knowledge [11]. The hidden nodes (variables) can capture effects that cannot be directly measured in a microarray experiment.

If x_t^i represents the i-th node at time point t, the joint probability distribution is given by Eq. 8:

$$p\left(x_t \middle| x_{t-1}\right) = \prod_{i=1}^{n} p\left(x_t^i \middle| pa(x_t^i), \theta, G\right) \tag{8}$$

The GRNs inference is followed by structure and parameter learning of the BNs from training data D [7]. For given data D, the aim is to find posterior distribution of the network structure M, and then from this distribution the structure M^* which best fits the data should be found according to Eq. 9:

$$M^* = \text{argmax}_M \{P(M \mid D)\} \tag{9}$$

For an optimal network structure M^* and given data, it is required to find posterior distribution of parameters q by Eq. 10:

$$q^* = \arg\max_q \{P(q \mid M^*, D)\} \tag{10}$$

The BNs learning is NP-hard task and thus BNs and DBNs are appropriate for inference of small networks [12] because the number of DAGs $G(n)$ super-exponential depends on the number of nodes n and it is given by the Eq. 11:

$$G(n) = \sum_{k=1}^{n} (-1)^{k+1} \binom{n}{k} 2^{k(n-k)} G(n-k) \tag{11}$$

In Table 1 the number of all possible DAGs as a dependence on the number of the graph nodes is shown.

Table 1. A table survey of the number of all possible DAGs depending on the number of nodes.

number of nodes	number of all possible DAGs
4	543
5	29 281
6	3 781 503
8	783 702 329 343
10	$4{,}175099 \cdot 10^{18}$
15	$2{,}377253 \cdot 10^{41}$
20	$2{,}344880 \cdot 10^{72}$
22	$1{,}075823 \cdot 10^{87}$
24	$9{,}435783 \cdot 10^{102}$

5 Validation of Inferred GRNs

To validate obtained results, the inferred network should be assessed in comparison with the referent network. Commonly used criteria for validation are the Receiver Operator Characteristics (ROC) curves and the area under ROC curve (AUC). The ROC curve is a chart of the ratio between sensitivity and (1-specificity), where sensitivity corresponds to a proportion of the actual positives edges which are correctly identified whereas specificity is proportion of negatives edges which are correctly identified [13] [15].

To facilitate the model validation , instead of ROC curve the AUC can be used. The AUC is the area covered by the ROC curve with the x-axis. Bigger value of AUC means better inferred network. The AUC is calculated by integrating the area bounded by the ROC curve and the x-axis [14].

6 Simulated Data and Results

To infer GRNs and then to validate above described models: Boolean networks, GGMs and DBNs, we have simulated artificial gene expression data by a simple first order autoregressive process given by Eq. 12:

$$X(t) = Ax(t-1) + B + \varepsilon(t) \quad (12)$$

where $\varepsilon(t)$ is a vector distributed by zero-centered multivariate Gaussian distribution with diagonal variance matrix.

We have obtained three different size datasets. The first dataset Data1 consists of simulated gene expression data for 5 genes and 50 time points. The dataset Data2 corresponds to 10 genes and 50 time points, and the number of genes in the third dataset Data3 is 15 measured in 100 time points.

The true referent networks and the inferred networks for the three datasets Data1, Data2 and Data3 are illustrated on Fig. 1-Fig.3. The values for AUC as validation criteria are shown tabular on Table 2. These AUC values show that for smaller datasets, Boolean networks model has the best performance in comparison to the GGMs and DBNs. For larger datasets GRNs inference performed by GGMs overcomes the other models.

Table 2. A comparison of the AUC values for three different inference models: GGMs, Boolean network and DBNs.

network inference model	Data1	Data2	Data3
GGMs	0.65	**0.63**	**0.57**
Boolean networks	**0.94**	0.56	0.46
DBNs	0.29	0.15	0.51

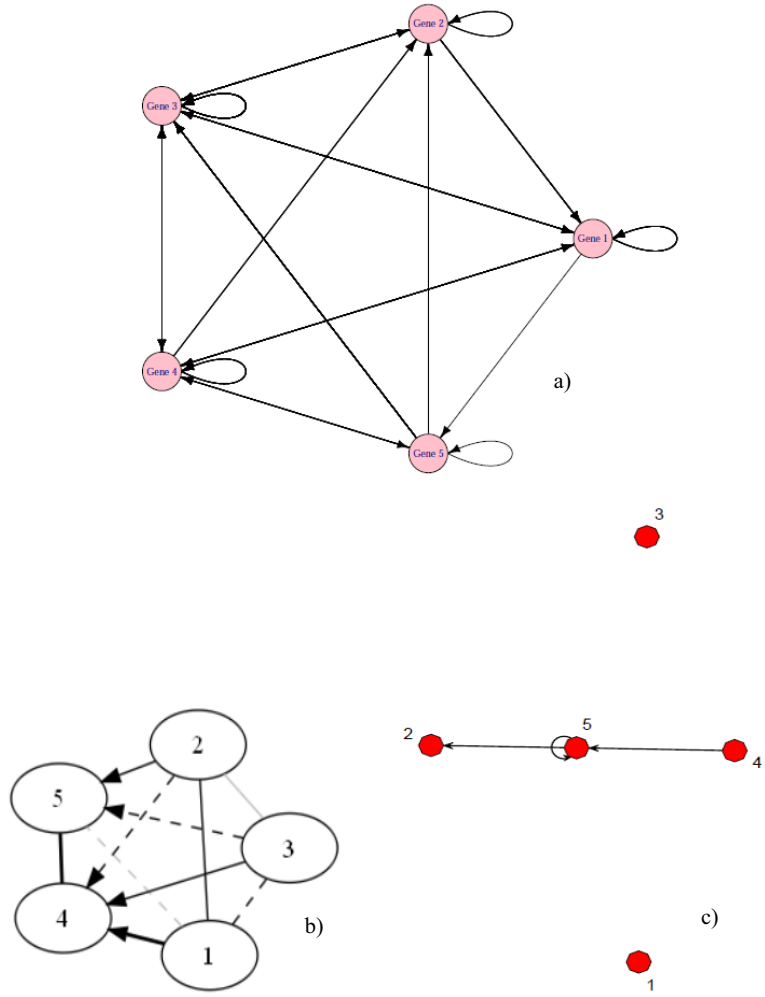


Fig. 1. True and inferred networks a) inferred Boolean network from Data1 b) reconstructed GRNs by GGMs and c) the true referent network corresponds to the Data1.

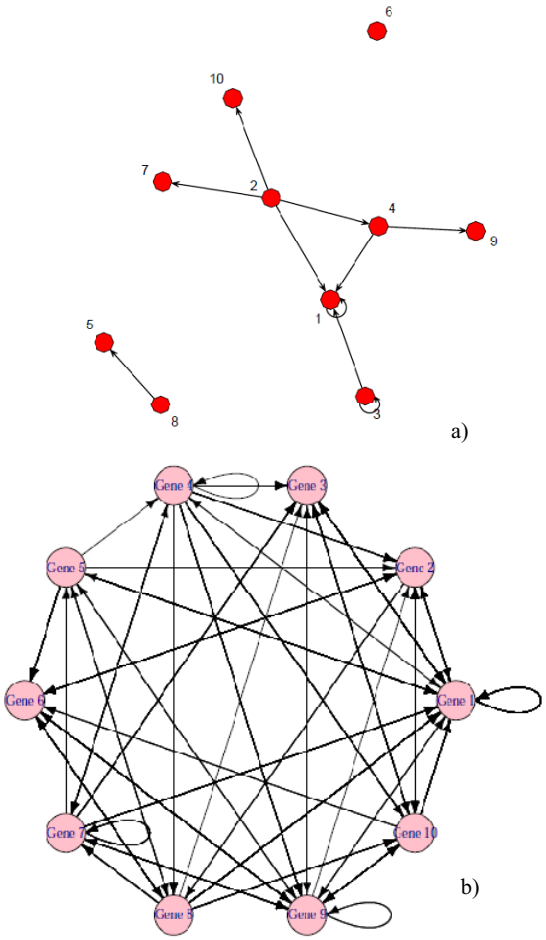


Fig. 2. True and inferred network a) the true referent network corresponds to the Data2 b) inferred Boolean network from Data2.

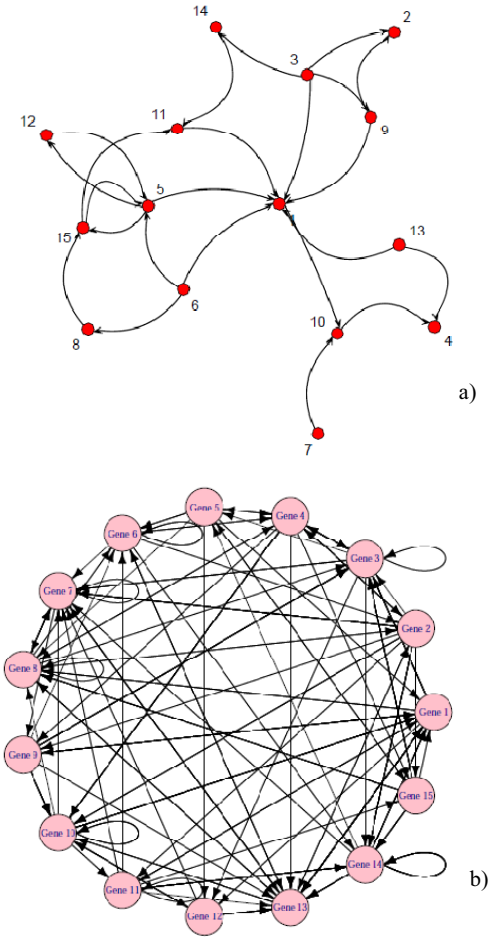


Fig. 3. True and inferred network a) the true referent network corresponds to the Data3 b) inferred Boolean network from Data3.

7 Conclusions

The presented AUC values obtained by GRNs inference using different models and datasets have shown that for datasets containing time series for larger number of genes, GGMs surpass the other network inference models: Boolean networks and DBNs. Only in the case where time series are for small number of genes (in our dataset - 5 genes) the Boolean network model has better inference performance compared to GGMs and DBNs, whereas DBNs model has shown worst inference properties. In accordance to these results we suggest using of GGMs results as prior

knowledge for improved approach for GRNs inference whereas the second inference stage is Markov Chain Monte Carlo (MCMC) simulation method to reconstruct more reliable GRNs.

Reference

1. N. Kraemer, J. Schaefer, A.-L. Boulesteix, *Regularized Estimation of Large-scale Gene Association Networks Using Graphical Gaussian Models*, Technical Report, Department of Statistics, University of Munich, 2009.
2. A. V. Werhli, M. Grzegorczyk, D. Husmeier, *Comparative Evaluation of Reverse Engineering Gene Regulatory Networks with Relevance Networks, Graphical Gaussian Models and Bayesian Networks*, Bioinformatics, Vol.22, no. 20, 2006.
3. F. Jaffrezic, G. Tosser-Klopp, *Gene Network Reconstruction from Microarray Data*, BMC Proceedings, 2009.
4. J. Schaefer, R. Opgen-Rhein, K. Korbinian, *Reverse Engineering Genetic Networks using GeneNet Package*, R-News 6/5:50-53, 2006.
5. S. Ma, Q. Gong, H. J. Bohnert, *An Arabidopsis Gene Network Based on the Graphical Gaussian Model*, Genome Research, 2009.

J. Schaefer, K. Strimmer, *Learning Large-Scale Graphical Gaussian Models from Genomic Data,* American Institute of Physics, Vol. 776, 2005.

6. B. Ristevski, S. Loskovska, *Bayesian Networks Application for Representation and Structure Learning of Gene Regulatory Networks*, 12th International Conference on Computers and Information Technology ICCIT '09. , Dhaka, Bangladesh; 2009.
7. B. Ristevski, S. Loshkovska, S. Dzeroski, I. Slavkov, *A Comparison of Validation Indices for Evaluation of Clustering Results of DNA Microarray Data*, Bioinformatics and Biomedical Engineering, 2008. ICBBE 2008, China.
8. S. Liang, S. Fuhrman and R. Somogyi, *REVEAL, a general reverse engineering algorithm for inference of genetic network architectures*, Pacific Symposium on Biocomputing 3, 1998, pp. 18-19.
9. N. Friedman and M. Goldszmidt, *Learning Bayesian Networks with Local Structure*, Proceedings of the NATO Advanced Study Institute on Learning in graphical models, 1998.
10. R. E. Neapolitan, *Learning Bayesian Networks*, Prentice Hall, 2003.
11. M. Grzegorczyk and D. Husmeier, *Improving the structure MCMC sampler for Bayesian networks by introducing a new edge reversal move*, Machine Learning, 71: 265-305, 2008.
12. L. M. de Campos, *A Scoring Function for Learning Bayesian Networks based on Mutual Information and Conditional Independence Tests*, Journal of Machine Learning Researc 7, 2006.
13. T. Fawcett, *An introduction to ROC analysis*, Pattern Recognition Letters 27, 2006, pp. 861-874.
14. C. T. Le, *Introductory Biostatistics*, John Wiley & Sons, Inc., New Jersey 2003.

Digital Tapeless Solution for HD/SD TV Stations and It's Workflow Automation

Gjorgji Manceski[1] and Goran Ambardziev[2]

[1]University St.Clement Ohridski, Bitola, Republic of Macedonia
gjorgji.manceski@uklo.edu.mk
[2]VerteX Broadcast Solution, Prilep, Republic of Macedonia
goran@vertex.com.mk

Abstract. This paper tries to present a digital solution for television channels, which can use HD/SD media for HD/SD broadcast with a possibility for automation of the whole business process. This solution, with its functionality and efficiency, can gain huge economic advantages for television companies. In this paper, we try to present the products with its functions and internal integration, and openness for integration in existing, third party solutions. This digital solution is based on a Windows platform. Supported Windows versions are "Windows XP Professional" and "Windows 7". Integration is implemented through database and storage servers – where the broadcast material, with all of its different media types, is stored. This solution covers all the critical points of working with digital media formats and of the business process and its automation, too. The design of the system architecture is plug-in oriented, using the most modern object oriented concepts of software development. Components are developed using "Microsoft Visual Studio.NET 2003", using C# and C++. Plug-in orientation enables easy component replacement and developing plug-ins for any specific hardware, without the need to intervene in the rest of the system. The system implements client/server architecture.
The plug-ins, provide the main functionality in the server. This plug-in based model imposed itself as modern and flexible way for the needs that this solution required:
Hardware independence of the server – the server is not coupled to a specific hardware (like a video card for ex.), new hardware implementation is just a matter of implementing a new plug-in.
Modern, object – oriented programming approach – the plug-in architecture abstracts specific hardware problems and completely isolates the server from it, leaving this task to the specified plug-in.
Media format independence – because the media formats area (video or audio) is developing in fast pace these last years, the server should be able to adopt them and adopt them fast. Isolating such format problems in a plug-in enabled us to catch with the pace efficiently.

Keywords: PlayOUT, tapeless HD/SD broadcast solution, PlayOUT automation.

A Case for Decision Support Systems on Project Management

Sinem Besir, Kökten Ulas Birant

Dokuz Eylül University, Computer Engineering Department, Buca,Izmir, Turkey
sinembesir@gmail.com, ulas@cs.deu.edu.tr

Abstract. Companies use project management systems in order to manage and plan their processes. The use of these systems provide functionalities such as generation of time charts, definition of milestones, adaption of time charts relative to risk analysis and more. The data used in these systems is entered by the system user. The validity and accuracy of this data entered is relative to the experience, knowledge and prediction skills of the person who entered the data. Because of this, results can be erroneous. To avoid this problem, an automatic calculation for input values could be used resulting in a more valid and accurate planning. Using the system developed, companies can perform project planning and management functions easier and through the parameters and statistical data used and more accurate time charts can be generated.

Keywords: Decision Support Systems, Data Mining, Project Management

1 Introduction

Development of software-based computerized solutions requires creativity, time, money and labor effort which are used in the various phases of development. One of the most important factors that should be taken into consideration in the process of development is planning and management of the resources available. Resource management is the efficient and effective deployment for an organization's resources when they are needed. Such resources may include financial resources, inventory, human skills, production resources or information technology. Project management systems are used for this purpose.

Project management is the planning, organizing, directing, and controlling of company resources for a relatively short-term objective that has been established to complete specific goals and objectives. Furthermore, project management utilizes the systems approach to management by having functional personnel assigned to a specific project. Classical project management includes a number of elements: initiation, planning or development, production or execution, monitoring and controlling, closing. So, the initiation phase contains too many estimations and assumptions. For this phase, the amount of the real and proven information is the most important data for a good estimation and therefore for a successful project. A

decision support system may be a supporter for the experts with a synthesis of the information.

2 A Simple Decision Support System

Project management systems which are currently in use operate by relying on the data gathered through the user. Projects are planned and managed according to the start and end dates provided by the user as it is decided by the project manager. The defined start and end dates may not be the optimal dates depending on the experience level of the project manager. In order to avoid the use of experience-based data, a standardized system which provides ease in data gathering and calculating can be developed. By generating a project archive with the parameters effective in the management process and applying data mining techniques to this archive, the effective weights of parameters can be found and used in order to obtain a standardized system.

2.1 Motivation

Project management systems which are currently in use operate by relying on the data gathered through the user. Projects are planned and managed according to the start and end dates provided by the user as it is decided by the project manager. The defined start and end dates may not be the optimal dates depending on the experience level of the project manager. In order to avoid the use of experience-based data, a standardized system which provides ease in data gathering and calculating can be developed. By generating a project archive with the parameters effective in the management process and applying data mining techniques to this archive, the effective weights of parameters can be found and used in order to obtain a standardized system.

2.2 Application

The standardized system would have the following roles in its structure:

- Projects, which are made up of tasks
- Tasks, which are performed by employees
- Employees, who are assigned to tasks

Considering the system as a whole, it is possible to group the parameters which would be used in the system in three groups: project-based parameters, employeebased parameters, task based parameters. In order to be able to determine the weights of the effective parameters, a sample data set must be prepared to train the system and determine the weight values.

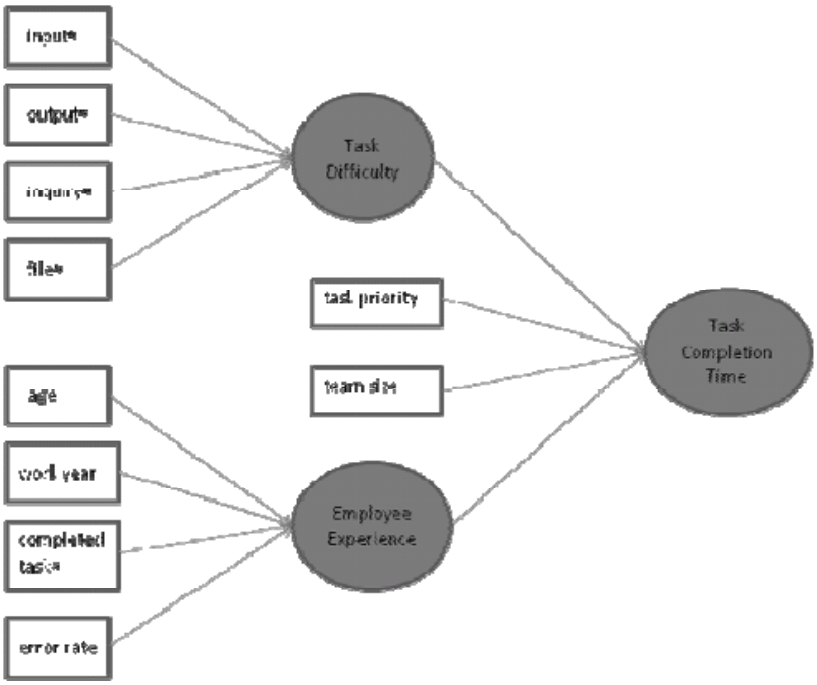


Fig. 1. Calculation of parameter weights and generation of the results

In the process of preparing the sample data set, the first step is determining the parameters that would be effective in training the system. Effective parameters are defined by considering the variability of the parameters. If a change in a parameter's value affect the result that is demanded to be calculated, that parameter must also be used in the process of finding the weights. The second step is to define the range of possible values that the parameter can have. The minimum and maximum range values would be defined by considering the worst and best cases for a parameter. The third and last step is determining the range intervals for the range found in the previous step.

Fixed k-interval discretization method would be appropriate to determine the range intervals. Fixed k-interval discretization method divides the sorted values of a numeric attribute into k intervals, where each interval contains (n/k) possibly duplicated adjacent values, where n is the number of observed values. "k" is determined without reference to the properties of the sample data set. The method originally ignores relationships among different attributes. In order to provide a relationship among the parameters determined at the first step, the intervals determined for all parameters needs to be combined in a form that there exists a value for each possible interval combination.

In order to calculate the parameter weight values which are used to calculate task completion time, employee experience and task difficulty, neural network algorithm is used by using the tables which contain sample datasets for the relative calculations. Neural network works by processing the sample dataset by recalculating weights

while processing each record in the table. When the error value becomes stable, it is possible to stop the processing. As the dataset size is quite large and at least one sample record is provided for all possible combinations of parameter range values, the processing is stopped after processing all records in the sample dataset twice.
The algorithm used for the neural network application in the thesis is as follows:
1. Initialize weight values (symbolized with w), learning rate (symbolized with lr) and threshold (symbolized with t)
2. Load the sample dataset into the memory
3. For each record in the dataset:
    a. Gather parameter values (symbolized with x) and result value (symbolized with z) and assign the values gathered into relative variables
    b. Calculate the output product (x * w) for all parameters and their corresponding weights (symbolized with c)
    c. Calculate the sum of all outputs (symbolized with s)
    d. If (s > t) then value assigned for network is 1, otherwise 0 (symbolized with n)
    e. Calculate the error value with the subtraction (z - n) (symbolized with e)
    f. Calculate the correction value with the production (lr x e) (symbolized with r)
    g. Calculate new weight values by the formula (w + (r * x))

2.3 Risks

The developed system is run on the sample dataset that is generated by producing random values according to predefined criteria. This would affect the reliability of the results as the dataset is not consisted of actual test data.

Relatively to this risk, the parameters are not sufficient; they are defined according to basic requirements. More efficient results would be gathered by using a more extensive list of parameters.

3 Further Work

When this system is expanded and worked on in the future, two possible upgrades can be possible. One of the improvements is providing a web-based system rather than a standalone system. By converting the current system into a web-based system, the system would be available to anyone from anywhere without an installation.

Developing a better formula can be achieved by having a more reliable dataset which consists of actual task results instead of a randomly generated dataset. The system can be expanded and run with a new dataset by adapting the parameters relatively to the new dataset. This would provide better parameter weight values. Additionally, a deeper inspection with real data would reveal other parameters that would be effective in the weight calculation.

4 Conclusion

Statistical Timeline Development System is designed to provide efficient and reliable results that are used for project management purposes with the usage of neural network algorithm and the formulas generated. The system is extendable for new parameters and new dataset formats. With self-learning logic, the system is capable of perfecting the weights used in the formulas which are used in generating the results. The purpose of the system is generating the most accurate results.

References

1. Kerzner, H.: Project Management: A Systems Approach to Planning, Scheduling and Controlling. John Wiley & Sons (2003)
2. Yang Y., Webb, G.I.: Non-Disjoint Discretization for Naïve-Bayes Classifiers, ICML'02, 2002

Selection of Mobile Base Station Location Using the Speech Quality as Primary Factor

Zoran Gacovski, Ivan Kraljevski, Biljana Spireva, Sime Arsenovski

FON University, Bul. Vojvodina, bb, Skopje, Macedonia
{zoran.gacovski,ivan.kraljevski,biljana.spireva,
sime.arsenovski}@fon.edu.mk

Abstract. The selection of base station location is one of the most important decision issues for mobile operators. We propose in this paper a new multiple criteria decision-making method in order to solve the location of base station problem under fuzzy environment. In the proposed method, the ratings of each alternative and the weight of each criterion are described by linguistic variables which can be expressed by triangular fuzzy numbers. The final evaluation value of each base station (BS) location is also expressed by a triangular fuzzy number. By calculating the difference of final evaluation value between each pair of BS locations, a fuzzy preference relation matrix is constructed to represent the intensity of the preferences of one location over another. We have simulated the measurement of speech quality and the BER, and took them as decision factors; then we solved different examples of decision-making for selecting the base station location.

Keywords: Fuzzy decision making, selection of the best alternative, speech quality, bit error rate.

1 Introduction

Base station (BS) location is a common problem faced by mobile operators, in terms of coverage and better signal. In recent years, increased use of cell phones has focused attention on base stations location. Base station is viewed as a tool for gaining more customers and increasing the quality of service. In order to cover larger area, and to strengthen the signal, selecting a suitable BS location has become one of the most important decision issues for mobile operators. In the process of selection it is necessary first and foremost to identify the set of influential factors relevant to the BS location selection. Many influential factors are considered for the selection of a particular BS location, e.g. investment cost, area coverage, covered population, speech quality, etc. [1]. Multiple criteria decision-making (MCDM) methods were provided to deal with the problem of ranking and selecting the BS location under multiple criteria [2]. In order to estimate the speech quality, test measurement of the transferred signal quality must be performed in aspect of BER and RxQual parameters. The parameters have to be measured on specific locations in the range of

the each BS that is subject of evaluation using specialized equipment (mobile BTS) or radio network planning software with detailed digital maps of the area.

In general, the selection of a best BS location from among two or more alternative locations on the basis of two or more factors is a multi-criteria decision-making problem. Under many situations, the values for qualitative criteria are often imprecisely defined for the decision-makers. Besides, the desired value and importance weight of criteria are usually described in linguistic terms, e.g. "very low", "medium", "high", "fair", etc. It is not easy to precisely quantify the rating of each alternative location and the precision-based methods as stated above are not adequate to deal with the plant location selection problem. This fuzziness in the BS location selection process motivated us to develop a fuzzy decision-making method.

By using the pair-wise preference relations, we present a new fuzzy decision-making method to deal with BS location selection problem in this paper [6]. The decision-making criteria are divided into quantitative and qualitative criteria in our method. The importance weights of decision criteria and the ratings of qualitative criteria are assessed in linguistic variables which are described by triangular fuzzy numbers. In the proposed method, we aggregate the ratings (fuzzy and crisp) and fuzzy weights to calculate the final fuzzy evaluation values of all candidate locations. A preference relation is defined to indicate the over degree of preference of each pair of BS locations by comparing the difference between their final fuzzy evaluation values for all possibly occurring combinations. According to the preference relations, we construct a fuzzy preference relation matrix and use a stepwise ranking procedure to determine the ranking order of a large number of plant locations.

The organization of this paper is as follows. First, we introduce the basic definitions and notations of fuzzy numbers and linguistic variables. Next, we define a fuzzy preference relation to derive the fuzzy preference relation matrix, and propose a stepwise ranking procedure to determine the ranking order of all BS locations. Then we give an overview of speech quality measurement, and an example is solved in Matlab to illustrate the working of the proposed method. Finally, we give some conclusions at the end of this paper.

2 Fuzzy Decision Making

In this section, a systematic approach to the BS location selection problem by using the concepts of fuzzy set theory and multiple-criteria decision analysis is proposed. This method is very suitable for decision-making under fuzzy environment. Knowing the fuzziness of the BS location selection problem, the importance weights of various criteria and the ratings of qualitative criteria are considered as linguistic (fuzzy) variables in this paper.

The linguistic variables can be expressed as triangular fuzzy numbers – given in Tables 1 and 2. We suggested that the decision maker easily uses the linguistic variables (shown in Table 1 and 2) to evaluate the importance of the criteria and the ratings of alternatives with respect to various subjective criteria.

Let $A_1,\ldots, A_m$ be possible alternatives (number of feasible BS locations) and $C_1,\ldots,$ Cn be criteria with which alternative performances are measured. As stated above, a

fuzzy multi-criteria decision-making method for the selection of BS location problem can be concisely expressed in matrix format as:

$$\underline{D} = \begin{bmatrix} \underline{x}_{11} & \underline{x}_{12} & \cdots & \underline{x}_{1n} \\ \underline{x}_{21} & \underline{x}_{22} & \cdots & \underline{x}_{2n} \\ \vdots & \vdots & \cdots & \\ \underline{x}_{m1} & \underline{x}_{m2} & \cdots & \underline{x}_{mn} \end{bmatrix} \quad \underline{W}=[\underline{w}_1, \underline{w}_2, \ldots, \underline{w}_n]$$

where $\underline{x}_{ij}$, $\forall i,j$ is the fuzzy rating of alternative A_i $(i=1, 2, \ldots,m)$ with respect to criterion C_j and $\underline{w}_j$ $(j=1, 2,\ldots,n)$ is the weight of criterion C_j. These fuzzy ratings and the weights of each criterion are linguistic variables which can be described by triangular fuzzy numbers, $\underline{x}_{ij}=(a_{ij}, b_{ij}, c_{ij})$ and $\underline{w}_j=(w_{j1}, w_{j2}, w_{j3})$.

Therefore, we can obtain the normalized fuzzy decision matrix denoted by $\underline{\boldsymbol{R}}$ as:

$$\begin{aligned} &\underline{\boldsymbol{R}}=[\underline{r}_{ij}]_{m \times n} \\ &\underline{r}_{ij} = (a_{ij}/c_j^*, b_{ij}/c_j^*, c_{ij}/c_j^*), j \in B \\ &\underline{r}_{ij} = (a_j/c_{ij}, a_j/b_{ij}, a_j/a_{ij}), j \in C \\ &c_j^*=max_i\, c_{ij}, \text{ if } j \in B, \quad a_j=min_i\, a_{ij}, \text{ if } j \in C \end{aligned} \tag{1}$$

where B and C are the set of benefit criteria and cost criteria, respectively.

The normalization method mentioned above is to preserve the property that the ranges of normalized fuzzy numbers belong to [0, 1].

Considering the different importance of each criterion, we calculate the final fuzzy evaluation value of each alternative as:

$$\underline{P}_i = \sum_{j=1}^{n} \underline{r}_{ij}(\cdot)\underline{w}_j, \; i=1, 2, \ldots, m \tag{2}$$

where $\underline{P}_i$ is the final fuzzy evaluation value of alternative A_i. After the calculation of the final fuzzy evaluation value of each alternative, the pair wise comparison of the preference relationship between the alternatives A_i and A_j can be established as stated in the following section.

To define a preference relation of alternative A_i over alternative A_j we do not directly compare the membership function of $\underline{P}_i$ and $\underline{P}_j$. Instead, we use the membership function of $\underline{P}_i(-)\underline{P}_j$ to indicate the preferability of alternative A_i over alternative A_j and then compare $\underline{P}_i(-)\underline{P}_j$ with zero.

Here, the final fuzzy evaluation values $\underline{P}_i$ and $\underline{P}_j$ are triangular fuzzy numbers. The difference between $\underline{P}_i$ and $\underline{P}_j$ is also a triangular fuzzy number and can be calculated as:

$$\underline{Z}_{ij}=\underline{P}_i(-)\underline{P}_j \tag{3}$$

$$\underline{Z}^{\alpha}_{ij}=[\, z^{\alpha}_{ijl}, z^{\alpha}_{iju}] \tag{4}$$

$$\underline{P}_i^{\alpha}=[\, p^{\alpha}_{il}, p^{\alpha}_{iu}], \; \underline{P}_j^{\alpha}=[\, p^{\alpha}_{jl}, p^{\alpha}_{ju}], \; z^{\alpha}_{ijl}=p^{\alpha}_{il}- p^{\alpha}_{ju}, \qquad z^{\alpha}_{iju}=p^{\alpha}_{iu}- p^{\alpha}_{jl}$$

If $z^{\alpha}_{ijl}>0$ for $\alpha\in[0, 1]$, then alternative A_i is absolutely preferred to A_j. If $z^{\alpha}_{ijl}<0$ for $\alpha\in[0, 1]$, then alternative A_i is not absolutely preferred to A_j. If $z^{\alpha}_{ijl}<0$ and $z^{\alpha}_{ijl}>0$ for some α values, we define e_{ij} as a fuzzy preference relation between alternatives A_i and A_j to represent the degree of preference of alternative A_i over alternative A_j. The e_{ij} is defined as:

$$e_{ij}= S_1/S,\ S>0,\ S_1= \int_{x>0} \mu_{\underline{Z}ij}(x)dx,\ S_2= \int_{x<0} \mu_{\underline{Z}ij}(x)dx,\ S=S_1+S_2 \qquad (5)$$

The value of e_{ij} is the degree of preference of alternative A_i over alternative A_j and $\mu_{\underline{Z}ij}(x)$ is the membership function of $\underline{P}_i(-)\underline{P}_j$.

Intuitively, S_1 indicates the portion where alternative A_i is preferred to alternative A_j in the most favorable situation. The e_{ij} indicates the over degree of preference of alternative A_i over alternative A_j. An illustration of calculating e_{ij} is shown on Fig.1. Therefore, $e_{ij}>0.5$ indicates the alternative A_i is preferred to alternative A_j. If $e_{ij}=0.5$ then there is no difference between alternatives A_i and A_j. If $e_{ij}<0.5$ then alternative A_j is preferred to alternative A_i.

Table 1. Linguistic variables for the importance weight of criterions

Very low (VL)	(0, 0, 0.1)
Low (L)	(0, 0.1, 0.3)
Medium low (ML)	(0.1, 0.3, 0.5)
Medium (M)	(0.3, 0.5, 0.7)
Medium high (MH)	(0.5, 0.7, 0.9)
High (H)	(0.7, 0.9, 1.0)
Very high (VH)	(0.9, 1.0, 1.0)

Table 2. Linguistic variables for the ratings

Very poor (VP)	(0, 0, 1)
Poor (P)	(0, 1, 3)
Medium poor (MP)	(1, 3, 5)
Fair (F)	(3, 5, 7)
Medium good (MG)	(5, 7, 9)
Good (G)	(7, 9, 10)
Very good (VG)	(9, 10, 10)

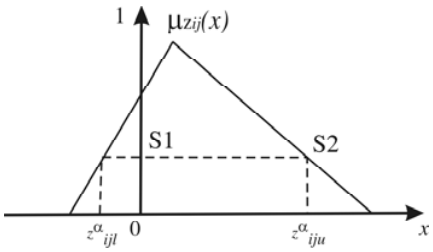


Fig. 1. An illustration of calculating e_{ij}

Using the fuzzy preference relation, we can construct a fuzzy preference relation matrix as:

$$\boldsymbol{E} = [e_{ij}]_{\mathrm{m\ x\ m}} \tag{6}$$

The fuzzy preference relation matrix represents the degree of preference of each pair alternatives. According to the fuzzy preference relation matrix ***E***, the fuzzy strict preference relation matrix can be defined as:

$$\boldsymbol{E}^{\mathrm{S}} = [e^{S}_{ij}]_{\mathrm{m\ x\ m}} \tag{7}$$

$$e^{\mathrm{S}}_{\mathrm{ij}} = \begin{cases} e_{ij} - e_{ji}, when\ \ e_{ij} \geq e_{ji} \\ 0, \quad otherwise \end{cases} \tag{8}$$

The value of e^{S}_{ij} is a degree of strict dominance of alternative A_i over alternative A_j. Then, the non-dominated degree of each alternative A_i *(i=1, 2, ...,m),* can be determined by using the fuzzy strict preference relation matrix as:

$$\mu^{ND}(A_i) = min_{j \in \Omega} \{1 - e^{S}_{ji}\} = 1 - max_{j \in \Omega} e^{S}_{ji} \tag{9}$$

where $\mu^{ND}(A_i)$ is the non-dominated degree of each alternative A_i and Ω is a set of alternatives. Therefore, we can use the $\mu^{ND}(A_i)$ values to rank a set of alternatives. The ranking procedure is described as follows:

(i) Set K=0 and Ω=$\{A_1,...,A_m\}$.

(ii) Select the alternatives which have the highest non-dominated degree, say A_h, $\mu^{ND}(A_h) = max_i\{\mu^{ND}(A_i)\}$. Set the ranking for A_h as $r(A_h) = K$+1.

(iii) Delete the alternatives A_h from Ω, i.e. Ω=$\Omega \backslash A_h$. The corresponding row and column of A_h are deleted from the fuzzy strict preference relation matrix [7].

(iv) Recalculate the non-dominated degree for each alternative A_i, $A_i \in \Omega$. If Ω=∅, then stop. Otherwise, set K=K+1, and return to step (ii).

3 Speech Quality as a Decision Factor

In order to estimate the speech quality regarding BER (bit error rate) and RxQual parameters, on-field measurement using specialized equipment (mobile or moveable BTS) could be performed, or simulation of the radio signal propagation with appropriate model for urban / rural areas as we have done.

In GSM, bit error rate (BER) measurements are used to decide whether transmitter power should be changed and in deciding whether a call should be attached to another base station. A single BER measurement in GSM is reported as one of the eight quality levels (RXQUAL_0...7) which is estimated by backward coding of the

decoded bit sequence and comparing it to the received bit sequence. This is a measure of the raw bit error rate, and does not take into consideration channel coding and the used speech codec. In this simulation - BER is estimated using AMR-NB codec with appropriate channel coding [3].
The received speech is compared with the test sequence transmitted between the BS and the receiver (downlink) in a similar way as the human speech perception and the quality is graded, (the listeners should do it in traditional subjective tests, like MOS).
Example of one of the most popular used algorithms for intrusive tests in packet switched and mobile networks is PESQ, defined in P.862 ITU-T [4]. PESQ is capable to predict subjective quality expressed by MOS values with good correlation in a very wide range of conditions, which may include coding distortions, errors, noise, filtering, delay and variable delay.

To justify the use of these two criteria several tests were performed with simulation testbed of the GSM transmission path. On Fig. 2 is presented BER plot of the simulated transmission path and measured MOS score. It could be noticed that these two parameters are loosely correlated, there is no linear relationship between them, and sometime good BER could produce unrecognizable decoded speech and vice versa. The location of the errors within speech frame has influence on the perceived speech quality as well.

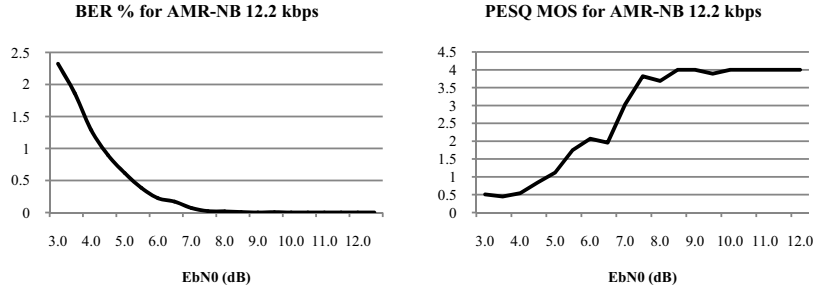


Fig. 2. BER and PESQ plot of the simulated GSM transmission path

AMR-NB (Adaptive Multi-Rate – Narrow Band, ACELP) is used for speech codec in these tests. It is standardized by ETSI for GSM applications, and it is chosen as mandatory for 3GPP networks [5]. AMR-NB is a speech codec with 8 different narrowband modes of operation and data rates between of 4.75 and 12.2 Kb/s. In the simulations, the rate of 12.2 Kb/s which is compatible with GSM-EFR codec (3GPP TS 26.071) is used. This speech codec is mainly used for toll quality speech compression in the 2nd and 3rd generation mobile telephony applications.

The system that is used for simulation is designed and coded in Matlab, it allows simulation of reference sequence transmission over the communication link with packet or frame loss events at the receiver side. Then comparison between the reference and received speech sequence is done and MOS score is evaluated.

Packet or frame loss is a major source of speech impairment and GSM applications. The impact of packet loss in perceived speech quality depends on

several factors, including loss pattern, codec type, and packet loss size. It may also depend on the location of loss within the speech. Even more - as most real communication channels exhibit burstiness of packet loss, occurrence of burst of lost packets has significant impact over speech perception.

The resulting data stream is then protected by an error-control coding scheme according technical specification [3]. On the radio transmission path, various sources of errors can disturb the transmitted data and at the receiver, the channel decoder attempts to recover from these errors and delivers a "cleaned up" version of the received data. Finally, speech is reconstructed in the speech decompression block.

In Table 3, BER and MOS values are shown, produced by simulation of transmission path (with simplified Rayleigh fading channel) with 3 different stations on 4 measurement points for given different EbN0 - energy per bit to noise power spectral density ratio.

Table 3. Simulated measurements on 4 points per BS location (makro-cells)

	BS1		BS2		BS3	
Eb/N0	RxQ (BER)	PESQ	RxQ (BER)	PESQ	RxQ (BER)	PESQ
6 dB	1,69%	2,10	2,52%	1,56	2,90%	1,36
7 dB	1,07%	2,08	0,96%	1,37	1,30%	2,14
8 dB	0,37%	2,33	0,48%	1,94	0,71%	2,40
9 dB	0,09%	3,02	0,22%	1,36	0,44%	2,49

4 Simulation Results

We have developed application in Matlab that enables selection of a location for establishing a new base station. A graphical output will be obtained in order to enable easy and effective application of our work for end-users. We will illustrate a problem with three decision-makers D_1, D_2 and D_3, three alternative locations, and five decision criteria. After preliminary screening, three candidate-sites A_1, A_2 and A_3 remain for further evaluation. The company considers the following five criteria to select the most suitable location:

(1) investment cost (C_1),
(2) area coverage (C_2),
(3) population covered (C_3),
(4) BER-RxQual (C_4),
(5) Speech quality - PESQ (C_5),

The benefit and cost criteria set are $B=\{C_2, C_3, C_4, C_5\}$ and $C=\{C_1\}$, respectively. The hierarchical structure of this decision problem is shown in Fig. 3.

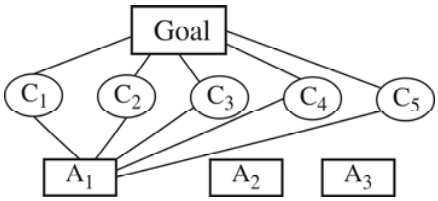


Fig. 3. The hierarchical decision structure.

The proposed method is currently applied to solve this problem. The computational procedure is summarized as follows:

Step 1: The decision-makers use the linguistic weighting variables (from Table 1) to assess the importance of the criteria and present it in Table 4. The fuzzy weight of each criterion calculated is given in Table 5.

Step 2: The decision-makers use the linguistic rating variables (shown in Table 2) to evaluate the rating of alternatives with respect to each criterion and reset in Table 6.

Step 3: According to Table 5, the fuzzy decision matrix is constructed as shown in Table 7.

Step 4: Construct the fuzzy normalized decision matrix as shown in Table 8.

Step 5: The final fuzzy evaluation values of three alternatives are calculated as:

$\underline{P}_1$=(3.03, 4.17, 4.74)
$\underline{P}_2$=(2.60, 3.83, 4.50)
$\underline{P}_3$=(2.57, 3.67, 4.37)

Step 6: The difference between two final fuzzy evaluation values are calculated as:

$\underline{P}_1$ (-) $\underline{P}_2$=(-1.47, 0.34, 2.14)
$\underline{P}_1$ (-) $\underline{P}_3$=(-1.34, 0.50, 2.17)
$\underline{P}_2$ (-) $\underline{P}_3$=(-1.77, 0.16, 1.93)

Step 7: Construct the fuzzy preference relation matrix:

$$\boldsymbol{E}=\begin{bmatrix} 0.50 & 0.66 & 0.72 \\ 0.33 & 0.50 & 0.56 \\ 0.28 & 0.44 & 0.50 \end{bmatrix}$$

Step 8: Construct the fuzzy strict preference relation matrix as:

$$\boldsymbol{E}^{\boldsymbol{S}}=\begin{bmatrix} 0 & 0.33 & 0.44 \\ 0 & 0 & 0.13 \\ 0 & 0 & 0 \end{bmatrix}$$

Step 9: Compute the non-dominated degree of each alternative A_i (i=1, 2, 3) as:
$\mu^{ND}(A_1) = 1.00$; $\mu^{ND}(A_2) = 0.67$; $\mu^{ND}(A_3) = 0.56$;

Table 4. The importance weight of the criteria

	$\mathbf{D_1}$	$\mathbf{D_2}$	$\mathbf{D_3}$
$\mathbf{C_1}$	H	VH	VH
$\mathbf{C_2}$	MH	H	MH
$\mathbf{C_3}$	H	H	H
$\mathbf{C_4}$	H	H	H
$\mathbf{C_5}$	VH	VH	VH

Table 5. The fuzzy weights of the criteria

	C_1	C_2	C_3	C_4	C_5
Weight	(0.8,0.9, 1.0)	(0.57,0.77,0.93)	(0.7, 0.9, 1.0)	(0.7,0.9,1.0)	(0.9,1.0,1.0)

Table 6. The ratings of the three candidates by decision-makers under all criteria

Criteria	Candid.	D_1	D_2	D_3
C_1	A_1, A_2, A_3	3, 6, 4 [mil]	5, 8, 6 [mil]	4, 7, 5 [mil]
C_2	A_1, A_2, A_3	G, VG, MG	VG, VG, G	F, VG, VG
C_3	A_1, A_2, A_3	F, G, G	G, G, MG	G, G, VG
C_4	A_1, A_2, A_3	VG, G, F	VG, G, F	VG, G, F
C_5	A_1, A_2, A_3	VG, G, VG	VG, G, VG	VG, G, VG

Table 7. The fuzzy decision matrix

	C_1	C_2	C_3	C_4	C_5
A_1	4 m.	(6.3, 9, 7)	(5.7, 7, 7)	(9, 7, 3)	(9, 7, 9)
A_2	7 m.	(8, 10, 8.7)	(7.7, 9, 8.7)	(10, 9, 5)	(10,9,10)
A_3	5 m.	(9, 10,9.7)	(9, 10,9.7)	(10, 10, 7)	(10,10,10)

Table 8. The fuzzy normalized decision matrix

	C_1	C_2	C_3	C_4	C_5
A_1	1	(0.6,.9,.7)	(0.6,0.7,0.7)	(0.9,0.7,0.3)	(0.9,0.7, 0.9)
A_2	0.6	(0.8,1,.9)	(0.8,0.9,0.9)	(1,0.9,0.5)	(1, 0.9, 1)
A_3	0.8	(0.9, 1, 1)	(0.9, 1, 1)	(1, 1, 0.7)	(1, 1, 1)

Step 10: The alternative A_1 has the highest non-dominated degree and set $r(A_1)$=1.

Step 11: Delete the alternative A_1 from the fuzzy strict preference relation matrix.

Step 12: After deleting the alternative A_1, the new fuzzy strict preference relation matrix is:

$$\boldsymbol{E}^S = \begin{array}{c} \\ A_2 \\ A_3 \end{array} \begin{array}{c} \begin{array}{cc} A_2 & A_3 \end{array} \\ \begin{bmatrix} 0 & 0.13 \\ 0 & 0 \end{bmatrix} \end{array}$$

The non-dominated degree of alternatives A_2 and A_3 are 1.0 and 0.87 respectively. Therefore, $r(A_2)=2$ and $r(A_3)=3$. The ranking order of the three alternatives is $\{A_1\}>\{A_2\}>\{A_3\}$. Therefore, the site A_1 is the best location to establish a new base station. We can see that the proposed method not only allows decision-makers to determine the ranking order of alternative alternatives but also can indicate the degree of preference of each pair of alternatives. Therefore, it is more suitable and effective in dealing with subjective judgments in an imprecise environment.

5 Conclusion

In this paper we have proposed a new fuzzy multiple criteria group decision-making method for solving the problem of base station (BS) location. In BS location

selection, very often, the assessment of alternatives with respect to criteria and the importance weight of criteria are given in linguistic variables. We have presented a stepwise and objective method to determine the ranking order of fuzzy numbers.

In this paper – a systematic and objective method is proposed to deal with BS location selection problem. The proposed method can help the decision-maker to make a suitable decision under fuzzy environment. In order to estimate the speech quality, on-field measurement of the transferred signal quality must be performed in aspect of BER and MOS parameters. The parameters have to be measured on specific locations in the range of the each BS that is subject of evaluation.

We have realized simulation in Matlab, and solved different examples of decision-making tasks for selecting the base station location. We've illustrated a problem with three decision-makers, three alternative locations, and five decision criteria. A graphical output is obtained in order to enable easy and effective application of our work for end-users.

6 References

1. G. A. Spohrer, T. R. Kmak, Qualitative analysis used in evaluating alternative plant location scenarios, Industrial Engineering, pp. 52-56, August 1984.
2. S. J. Chen, C. L. Hwang, F. P. Hwang, Fuzzy Multiple Attributes Decision-making Methods and Applications, Springer, Berlin, 1992.
3. 3GPP TS 05.03 V8.9.0 (2005-01) 3rd Generation Partnership Project; Technical Specification Group GSM/EDGE Radio Access Network; Channel coding, (Release 1999)
4. ITU-T Rec. P.862, "Perceptual Evaluation of Speech Quality (PESQ), an Objective Method for End-to-end Speech Quality Assessment of Narrowband Telephone Networks and Speech Codecs", International Telecommunication Union, Geneva, Switzerland (2001 Feb.)
5. Digital Cellular Telecommunications System (Phase 2+), Universal Mobile Telecom System (UMTS), AMR Speech Codec, 3GPP TS 26.071 Version 6.0.0 R6.
6. T.Y. Chou, C.L. Hsu, M. C. Chen, A Fuzzy Multi-criteria Decision model for International tourist hotel selection, International Journal of Hospitality management, vol. 27, 2008.
7. H. K. Alfares, S. O. Duffua, Assigning cardinal weights in multi-criteria decision making based on ordinal ranking, Journal of Multi-criteria Analisys, vol. 15 (5), pp. 125-133, 2008.
8. C. T. Chen, A fuzzy approach to select the location of the base station, Fuzzy Sets and Systems, vol. 118, pp. 65-73, 2001.
9. Z. M. Gacovski, S. Deskovski, Expert Decision and Control: Fuzzy Petri net Reasoning Simulator (FUPERS), IO on Automatic Control, Winning Award, St. Petersburg, 2002.
10. H. M. Hsu, C. T. Chen, Aggregation of fuzzy opinions under group decision-making, Fuzzy Sets and Systems, vol. 79, pp. 279-285, 1996.
11. A. Kaufmann, M. M. Gupta, Introduction to Fuzzy Arithmetic Theory and Applications, Van Nostrand Reinhold, New York, 1985.

A new design of a system for contest management and grading in informatics competitions

Bojan Kostadinov[1], Mile Jovanov[2] and Emil Stankov[2]

[1] Faculty of Electrical Engineering and IT, Karpos II bb, 1000 Skopje
bojankostadinov@gmail.com
[2] Institute of Informatics, FNSM, Gazi Baba b.b., 1000 Skopje
{mile, emil}@ii.edu.mk

Abstract. Competitions in informatics are usually synonyms for algorithmic programming contests. Many of these competitions use automatic grading of the contestants' solutions. This is done by running them on batches of input data and testing the correctness of the output. In this paper we will introduce a newly developed system called "MENDO", which is used by the Macedonian Computer Society in the organization of national informatics competitions. We will outline the main modules every grading system should support (sandbox, grader, controller, auxiliary modules) and present how each one of them is implemented in our system. We will show that MENDO is comprehensive and superior in a number of functionalities as compared to the other presently existing systems.

Keywords: programming contests, algorithmic problems, grading systems, automatic grading, task evaluation, contest management system

1 Introduction

Competitions in informatics were introduced about forty years ago, with the idea of attracting talented young people to the science of computer programming. These competitions are usually synonyms for algorithmic programming contests (other types include architecture, design, development, specification, assembly, testing scenarios, etc). The closest thing to these algorithmic programming contests from the point of view of competition problems are the Olympiads in Mathematics.

Competitions in informatics have a long tradition in Macedonia. There were 20 national contest cycles till the end of 2009. After many competitions on national level, the best contestants represent themselves and Macedonia at IOI – International Olympiad in Informatics, BOI – Balkan Olympiad in Informatics (for high school students), and JBOI – Junior Balkan Olympiad in Informatics (for primary school students).

Usually, these programming competitions require students to submit programs which are then run through a variety of test scenarios and judged accordingly. The difficulty however, lays not so much in the programming but rather the design of the underlying algorithms [1]. A sample task can be seen in the appendix at the end of

this paper. More often than not, these contests are based on automatic grading of the submitted solutions. This is accomplished by running them on batches of input data and testing the correctness of the output. Time and space limits are usually enforced during the process, which allows to judge not only by the (approximation of) correctness of the solution, but also by its time and space complexity [2]. Besides the automatic grading capability, a contemporary programming contest system requires some other functionalities that would facilitate the preparation of the students for competitions, like some kind of a communication page (or a forum), a page for sharing appropriate learning materials, an info page, etc.

2 Contest management and grading systems

Historically, in Macedonia for example, a few separated systems (evaluation program, contest website, forum page called "Communication window", etc) supported the organization of the competitions. However, a lot has changed since the first competitions – the computer equipment has improved, new information technologies have been introduced, and at the same time, the technique and technology of development and judging competition tasks has improved.

At the International Olympiads in Informatics a use of a grading system is an absolute necessity. Let us reflect on IOI'1998 in Setubal, Portugal, with about 250 contestants, 3 tasks and 20 test cases average for a task. For each test case 2 executions have to be performed: execution of the solution with the corresponding test case input and execution of the checker (typing the output of the contestant on the screen and comparing it, visually, with the judge output is even more time consuming). That means a total amount of about 30000 executions. This is why the grading process in Setubal started at 3 P.M. and finished at 6 A.M. on the next day. The solution of the problem outlined above is obvious – use a software system for organizing and implementing the evaluation process [3]. The improvement is best seen in Fig. 1.

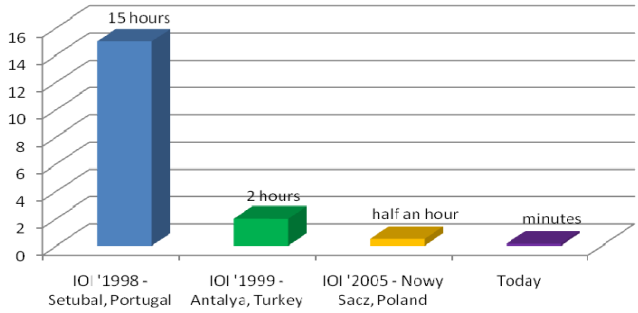


Fig. 1. Bar chart showing the time required for grading around 1000 algorithmic programming solutions on the IOI competitions (in chronological order)

The most suitable approach, from a practical point of view, certainly is to integrate all the required functionalities mentioned above in a single automated system for

complete contest management. This system would provide all the necessary contents for the students to successfully prepare and participate in the programming contests, and also all the necessary capabilities for the organization of programming contests (including submission of tasks and automatic grading).

3 Present contest management systems

There are few existing systems for contest management and grading. Traditionally, we still refer to them as grading systems. Today, however, we can talk about self managed systems that completely administer contests (including gathering and grading of submissions, managing competitor's questions and clarifications, publishing results, providing statistics, etc). Some systems even offer a continuous on-line training process.

Whatever the system offers, it should be designed such that it is simple, robust, secure and flexible. Standard modules for almost all grading systems are:

- **Sandbox** – ensures that the execution of a submission will not harm the system or the host computer; enforces time, memory and network restrictions. The sandbox is really the "heart" of every grading system.
- **Grader** – does compilation of contestant solutions, management of the sandbox, comparison of user output to correct output, and grading of a submission.
- **Controller** – handles the communication with the judges and competitions, and executes database operations.
- **Auxiliary modules** – handle printing, backups, storage, etc.

Table 1 summarizes some popular systems with their significant characteristics. As it is shown, American and Moe do not implement ACM style grading, while PC^2 does not implement IOI style grading. With the exception of PC^2, all the systems are dependent on the operating system that is used.

Table 1. Comparison between some of the popular grading systems

	pluggable graders	IOI style grading	ACM style grading	operating system	front end
American	Yes	Yes	No	Linux	web browser
Moe	Yes	Yes	No	Linux	command line
SIO	Yes	Yes	Yes	Windows	desktop app
PC^2	Yes	No	Yes	Win, Linux	java app

4 Architecture and analysis of the proposed system

Here, we describe the newly developed system called "MENDO", which is currently used by the Macedonian Computer Society in the organization of national informatics

competitions. MENDO was developed following the goal of integration of all previously used module functionalities in one compact environment. It was enriched with new features endorsed by the new technologies.

4.1 Architectural design

As shown in Fig. 2, MENDO includes all the modules that are standard for the contest management and grading systems. The implementation specifics for each of these MENDO's modules are outlined below:

- MENDO's **sandbox** uses P/Invoke (Platform Invoke) signatures and Win32 functions to create processes and group them in jobs. Jobs have the ability of limiting process privileges and resources.
  As a job object allows groups of processes to be managed as a unit, we use them to limit processes count, time and memory usage, security, etc. Job objects are namable, securable, sharable objects that control attributes of the processes associated with them. Operations performed on the job object affect all processes associated with the job object.
  The Linux variant of the sandbox is based on the Moe-Eval system, and uses Linux kernel calls (ptrace, ulimit, etc) to run the contestant programs in a controlled environment [4].

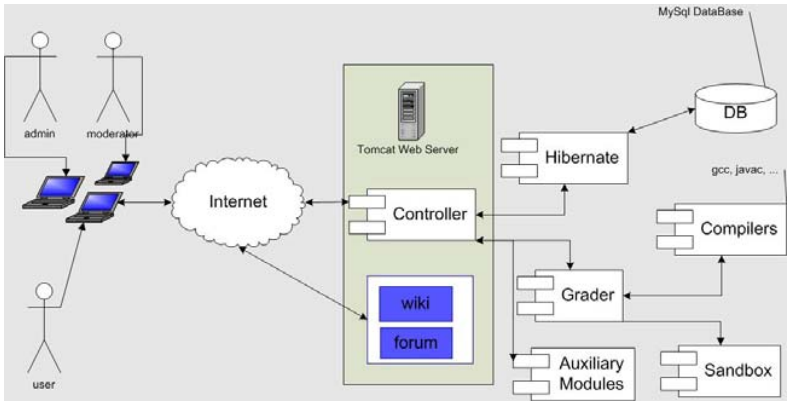


Fig. 2. MENDO's architectural design – the communication between the controller, sandbox, the graders and the auxiliary modules

- The **grader** is written in Java, and can be run on every popular operating system (like MS Windows and Linux)
- Judges and competitors use their web browsers to communicate with the system through a web application written in Java – the **controller.** For database operations we use Hibernate and C3P0 thread management to connect to a database and execute queries.

- In our implementation, the **auxiliary modules** are part of the web application and are also written in Java and run on an Apache Tomcat server. Besides the main system, MENDO also contains a public forum and a wiki.
  Every user is allowed to post content on the wiki, and to upload materials, images and other files. Only administrators and moderators are allowed to delete content. The same approach applies to the forum. As a wiki engine, we use JSPWiki, which is a feature-rich and extensible WikiWiki engine built around the standard J2EE components. JForum is used as a discussion board system (forum).

4.2 Specific MENDO features

An important feature of MENDO is that it uses cookies to support SSO – Single Sign On across multiple applications (the main system, the wiki, and the forum). All applications are running on the same server. Since all subsystems use Single Sign On and automatic language detection, there is no need to change the language or login every time you switch from one subsystem to the other, since this is automatically done by the underlying system.

MENDO uses 3 levels of caching (database caching – as provided by Hibernate, JCS – "Java Caching System" as secondary cache, and its own submission caching system).

Other MENDO's specific features, can be summarized as follows:

- MENDO is one of the rare systems that can operate on Microsoft Windows. Besides the fact that only a small number of grading systems operate on MS Windows, we found the OS to be very stable, reliable and easily controllable. The system can also be run on any popular Linux distribution;
- MENDO easily distributes load. Plugging more graders is easy, thanks to the modular architectural design of the system;
- It controls the entire system of the Computer Society of Macedonia, by providing automatic backups for itself and the other applications, self-tests of the application, the server and the operating system;
- The system has multilingual support (currently Macedonian and English, but we are planning to add more languages in the near future);
- MENDO is managed by several administrators and moderators, each with his own privileges and responsibilities. Every moderator and administrator can add tasks, create competitions, generate reports for each task and competition, and initiate system backups;
- There is heavy use of AJAX to simplify user interface operations (during registration, training sessions, competitions, etc). We use the jQuery 'write less - do more' javascript library to implement most of the event handling and AJAX operations;
- The system is designed and created by using free software (Java, Apache Commons, Hibernate, Struts, Java Caching System - JCS), and runs "entirely" on free software (Tomcat, MySQL, JForum, JSPWiki);

- MENDO was designed to be an entire gateway for algorithmic related topics, by including a news page, forum and an open wiki for publishing results, solutions and programming related materials.

4.3 Employment and performance of MENDO

In practice, MENDO is used as:

- a training system (contains tasks from past contests, both national and international);
- a contest management system (for organizing official national competitions and open online tournaments);
- Macedonian algorithmic programming gateway, containing a news page, a lot of programming related materials (organized in a wiki), and a public forum.

The training system ***can be used 24/7***. Every time a user logs in to our web site, he can view all the tasks that are available for training, and possibly submit a solution. After a solution has been submitted, the submission is added to a queue and judged as early as possible (this is ***no longer than 20 seconds***, even during heavy-load competitions). After a submission has been judged, the results of every test case are shown to the user. There is no limit to the number of submissions a user can make during a time period, but the system does ***support a couple of defensive mechanisms to prevent Denial of Service attacks***. A screenshot of the MENDO training system is shown in Fig. 3.

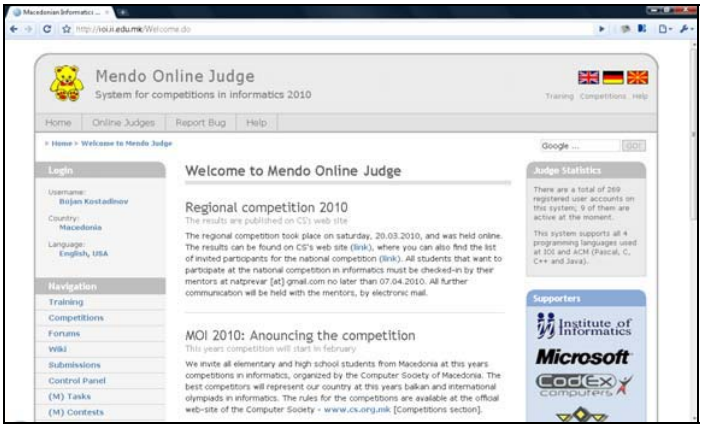


Fig. 3. Screenshot of the MENDO training system (made with Google Chrome). You can visit the system online at http://ioi.ii.edu.mk

As a contest management system, we have used MENDO in the 2010 contest cycle (more than 10 competitions, including both online and onsite rounds), and the system

proved to be quite stable, fast and robust. However, we plan to add a lot more features for the 2011 contest cycle.

Looking at the official rules of the International Olympiads in Informatics (IOI), we can conclude that MENDO ***supports every IOI rule and requirement***. It contains modules for detailed feedback, group testing, automatic competition management, statistics, backups, printing, and a lot more.

One other thing that is worth mentioning here is the ability of the system to produce reports and statistics. After the grading of each competition, the results are automatically published (this can be disabled, if necessary), and a report is generated. This report contains details and statistics for every competitor, every programming language, every task and even every test case that is part of that competition.

We tested the system in great detail, carried out many experiments and compared the system with the most popular grading systems. Table 2 is an extension of Table 1 and shows a comparison between MENDO and the grading systems that were previously analyzed.

Table 2. Comparison between MENDO and other popular grading systems

	pluggable graders	**IOI style grading**	**ACM style grading**	**operating system**	**front end**
American	Yes	Yes	No	Linux	web browser
Moe	Yes	Yes	No	Linux	command line
SIO	Yes	Yes	Yes	Windows	desktop app
PC^2	Yes	No	Yes	Win, Linux	java app
MENDO	Yes	Yes	Yes	Win, Linux	web browser

Using the open source software Apache JMeter, which is a 100% pure Java desktop application designed to load test functional behavior and measure performance, we have tested MENDO's 3 levels of caching architecture. Our architecture ***passed the test of over 5000 concurrent users***. The support of so many concurrent users is more than enough because one can rarely find a competition with so many participants. At the International Olympiads in Informatics there are usually around 300 competitors, and at ACM competitions even less (and they compete in teams, so in reality there is only one connected user per team).

We also made an experiment and ran about 2000 submissions through the system (we used a task given at our national competitions and the contestant's solutions to the task – executed enough times). The results we got from running the submissions on just 5 machines (at IOI, more than 30 machines are used per competition day) showed that the system can easily handle the load and that the time it took to test the solutions, generate the results and the statistics is comparable to all other grading systems outlined above. By supporting virtualization and/or implementing several graders on one machine, it is possible to lower the time required to test these solutions even more.

5 Future improvements of MENDO

In the near future we plan to address several remaining issues, including time measurement, reactive tasks, hacks, java security and large input/output. We also plan to completely rewrite the sandbox and add a new Rich Internet Application front-end to the system, based on Adobe Flex.

Time measurement is one of the most problematic issues that arise in grading algorithmic solutions, and is something that has been discussed in several papers. The current implementation of time measurement, based on the 'time' command, which can either return the real (or wall clock) time, or the user+sys (or instructions+system calls) time, can be improved by measuring the time of the solution programs several times and taking the minimum of those values.

Reactive tasks are a relatively new addition to programming contests. MENDO currently supports three types of tasks: reactive tasks, standard/batch tasks - or tasks for which the contestant needs to submit a source code of a program that reads the input data from a file or keyboard, and writes the results to a file or the monitor; and output-only tasks - or tasks for which the contestant already has the input data and just needs to submit the output.

Restricting java applications is another big issue, because the Sun JVM (Java Virtual Machine) performs a lot of operations that would normally be blocked by the kernel module [5]. The option of using virtualization to protect the machines on which we execute solution programs could be additionally considered. With the introduction of new multi-core processors, it can also be worthwhile to setup several graders on just one machine. This will greatly lower the time required to test solution programs, but serious testing has to be done before this can be used in a real contest environment.

A chat module is something that probably will greatly improve the system (a lot of popular systems, like TopCoder's Arena already allow communication). Implementing such a module can even ease the communication between contestants and organizers, and even between the organizers themselves.

Finally, a support for more competition modes can be added. Currently, only IOI and ACM grading styles are supported, but there are other popular grading options that can be used in real competitions (like TopCoder's grading based on submission speed). Also, another option that can be considered is allowing solutions in more programming languages, even though the system already supports all languages allowed at IOI and ACM competitions.

6 Conclusion

In this paper we discuss the need for contest management and grading systems. We have developed and presented a contest management and grading system called “MENDO”. We outlined the main modules every grading system should support (sandbox, grader, controller, auxiliary modules) and showed how every one of them is implemented in the MENDO system. This system is comprehensive and superior in a number of functionalities as compared to the other presently existing systems.

MENDO is currently used by the Macedonian Computer Society in the organization of national informatics competitions. Other contest organizers can smoothly adopt this system for use in their national Olympiads in informatics as it provides all the necessary functionalities for various forms of IOI style competitions. MENDO supports every IOI rule and requirement. Additionally, as it is web oriented, it can be used for organization of online world wide competitions.

Appendix: Sample task

Fence (ACM ICPC 2001. Northeastern European Region)

Workers are going to enclose a new working region with a fence. For their convenience the enclosed area has to be as large as possible. They have N rectangular blocks to build the fence.

The length of the i-th block is L_i meters. All blocks have the same height of 1 meter. The workers are not allowed to break blocks into parts. All blocks must be used to build the fence.

Input: The first line contains one integer N ($3 \leq N \leq 100$). The following N lines describe fence blocks. Each block is represented by its length in meters (integer number, $1 \leq L_i \leq 100$).

Output: Write one non-negative number S - maximal possible area of the working region (in square meters).

Sample Input	Sample Output
4	28.00
10	
5	
5	
4	

References

1. Burton, B.A.: Informatics Olympiads: Challenges in Programming and Algorithm Design. In Proc. Thirty-First Australasian Computer Science Conference (ACSC 2008), Wollongong, NSW, Australia. CRPIT, 74. Dobbie, G. and Mans, B., Eds. ACS. 9-13.
2. Mares M.: Perspectives on Grading Systems. Olympiads in Informatics 2007 Vol. 1. pp. 124-130 (2007)
3. Manev K., Sredkov M., Bogdanov T.: Grading Systems for competitions in programming. Thirty-eight spring conference of the Union of Bulgarian Mathematicians (2009)
4. Mares M.: Moe - Design of a Modular Grading System. Olympiads in Informatics 2009 Vol. 3. pp. 60-66 (2009)
5. Merry B.: Using a Linux Security Module for Contest Security. Olympiads in Informatics 2009 Vol. 3. pp. 67–73 (2009)

CiGLA: Context aware mobile service based on SMS and MMS for use in tourism

Jovan Kostovski

Faculty of Electrical Engineering and Information Technologies
Ruger Boskovic bb, Skopje, Republic of Macedonia
jovan.kostovski@gmail.com

Abstract. This paper describes a context aware mobile service based on Short Message Service, SMS, and Multimedia Message Service, MMS, for use in tourism. The goal of this research was to determine the contextual information which defines the tourists while they go sightseeing in some area and to use that information, to present them some useful information about the places they are visiting. As a result from this research a prototype system was build and the defined use cases were tested. The built system is a good base for developing various services on top of it which will use the collected contextual information.
Keywords: mobile context, mobile service, mobile computing, sms, mms, tourism.

1 Introduction

With the rapid growth of technology, mobile phones and mobile devices in general, have become widespread, powerful communication tools able to present information in various different formats such as: text, pictures, audio and video. Because of such capabilities, they are great platform for presentation of information. People are in constant movement and search for information, therefore getting the right information in the right time at the right place is very important. Tourists as a category of people who are always on the move and with restricted amount and weight of things they can carry would be glad if they have a tourist guide in their phone. So instead of carrying lots of paper based tourist guides they will use a device which they always have and get prompt information for the place where they are located. The system I build, CiGLA, City Guide Location Assistant, locates the tourists and sends them some relevant information based on their location and the time when they sent the request. They can get information for historical sites, restaurants and other predefined places in their surroundings.

1.1 Context Aware Mobile Services

The context aware mobile services are based on context, information which can be used to determine the state in which some entity is. The entity can be a human, a place or an object which is relevant for the interaction between the system and its users. In my case the interaction is between the system and the tourist. The state of the tourist can be defined by the following parameters: access time, location, mobile device capabilities, type of access network and some personal settings [3][4]. From the mentioned parameters, CiGLA uses the location, access time, type of requested service and chosen language. Those are the parameters which are used for message generation.

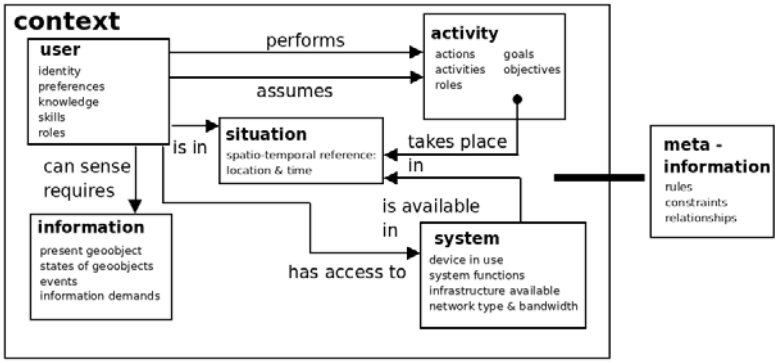


Fig. 1. Context model for mobile cartography.

Figure 1 represents the context model for mobile cartography and the interaction between the user and the system. It can be clearly seen that the user in need of information performs some activities and the system reacts accordingly to these activities and the state of the user.

1.2 Location Based Services

The Location Based Services, LBS, are services available to the users of mobile devices. The location of the users is determined by the systems of the mobile operators GSM network or GPS. There are two types of positioning in GSM network, cell based (the area covered by one cell) less precise and triangulation based, the most precise method [4].

Figure 2 represents Location interoperability form (LIF) [5], the flow of the positioning data when the user accesses the resources via web browser and a mobile device. The most frequently used medium for LBS is the SMS message. The common usage of the location based services is giving information how to get somewhere (in

stages), location based marketing, tracking people or goods, locating a nearby object such as bank, restaurant, shop etc.

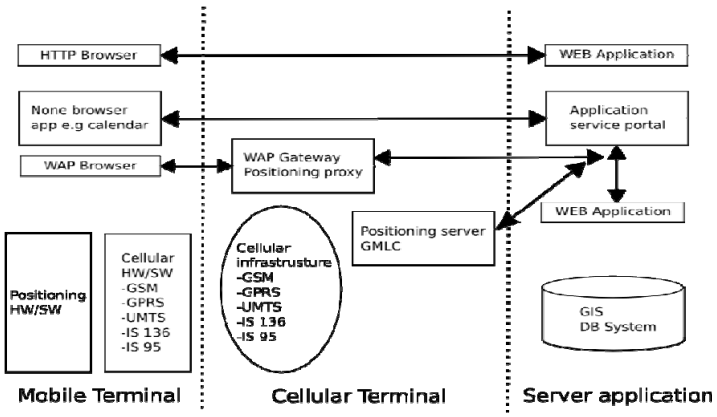


Fig. 2. Location interoperability form (LIF).

2 Related work

There are different types of mobile tourist guides available, which are based on different technologies and users are accessing the information in different ways. Some of them are installed as applications on the phone (they don't use context) and some of them have interaction, track the user and send information. The most of them are client/server web applications, which are sending web pages adapted for the user's device [7][8]. Some of them enable personal customization of the data presentation. This method of data presentation is bad because the charge rates for internet access while the user is abroad are very high. The other drawback is that the content must be adapted for many different types of devices, which means more processing on the server side. From the applications which I found during my research the once worth to mention are: COMPASS - COntext-aware Mobile Personal ASSistant, CRUMPET - Creation of User-friendly Mobile Services Personalized for Tourism , GUIDE, Lol@ - Local Location assistant, MobiDENK, m-ToGuide.

3 System overview

CiGLA is a system that tracks the tourist location in some defined area and based on it sends information of interest to the tourists. In the area where the tourists are moving, some locations are marked because they are considered to be of tourist interest. These locations are called Points of Interest, POI and can include: historical

sites, restaurants, cash machines theatres, museums etc. The system can autonomously track the users and send relevant information or the users can request information by sending Unstructured Supplementary Data, USSD, requests or SMS messages. These requests have predefined format which is shown in Table 1.

Table 1. USSD and SMS request format.

Type of request	Format
USSD	*USSD_CODE*REQUEST_CODE*TYPE_OF_DATA# example: *156*250*1#
SMS	KEYWORD POI_ID TYPE_OF_DATA example: INFO 250 MMS

CiGLA supports sending of relevant information as SMS, picture MMS or video MMS. For automatic receiving of messages the tourist has to subscribe to a sightseeing tour. The sightseeing tours are groups of indexed POIs. If the user is located at a specific POI he/she will automatically receive message with directions how to get to the next POI in the tour. The sending of the messages with directions how to get from a specific POI to another is strictly defined. If we want the system to send directions to a specific POI from a random location, a Geographic Information System, GIS, must be used. The messages are sent in the user's native language or a predefined language. The user's native language is determined by the user's phone number.

From technical point of view the system has common architecture for systems for Mobile Value Added Services, VAS [10][11][12].

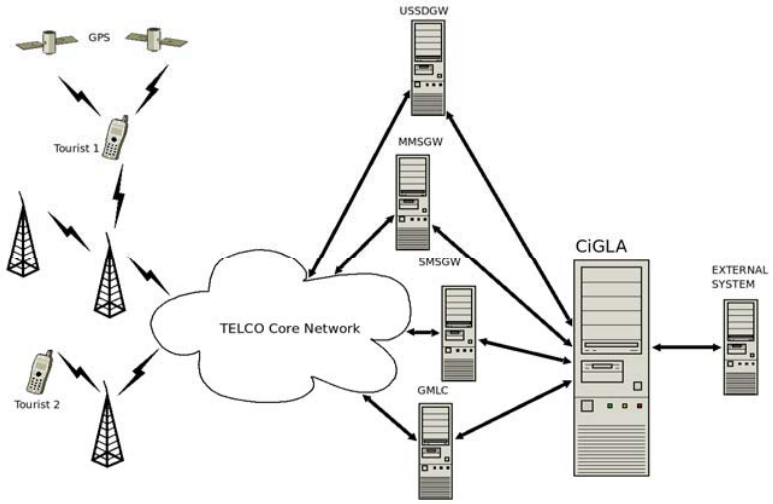


Fig. 3. System architecture

As it can be seen on Figure 3, the system is interconnected with the core systems from the mobile operator network with Gateways, GW. They are transforming the communication from the elements of the core network of the Telecom operator to IP communication. Through this elements CiGLA sends SMS and MMS messages, gets the location of the users and requests services from other external systems. The system doesn't do any content adaptation for MMS messages, that operation should be implemented in the Multimedia Message Center, MMSC. It doesn't enable any billing functionality as well. The only statistics that can be gathered from the system is the amount of a specific message type which is send to some user or group of users. From user point of view the good characteristics of the system are: fully user oriented solution, sending requests via USSD which is free of charge and automatic detection of the user's language. From technical point of view the good thing is that the system is based on GSM network with standard technologies such as SMS and MMS and the server side logic is based on well proven architecture which enables high availability and reliability. The bad thing about this solution is that sending SMS and MMS works with store and forward mechanism. The messages are first stored in the appropriate system in the mobile operator's core network and then are sent. This means that the message delivery might be delayed due to high system load. The other problem is that because the messages are send to the user's native language they are using UTF-8 encoding which means that the SMS messages will have smaller payload in number of characters and more than one message will be sent per request. The other thing is that the size of the MMS messages is usually limited to few hundred kilo bytes and the expensive GPRS roaming.

4 System implementation

During the development of the system I used Free/Open Source Software, FOSS, technologies only. Every software component was carefully chosen and because the whole system works in a GSM network, some GSM standards are satisfied. The goal of the system development was to build secure, stable, reliable, modular system which can be easily customised to the customer needs and in the same time to have low price. The FOSS licences enable the software to be used for whatever we want and the access to the source code and the big user community assures us that we can easily tailor the whole solution to our needs. From other point of view, usually the FOSS has lower cost or we can get it free of charge. To assure big modularity and scalability the whole system has layered design and all the software components are written using object-oriented concepts. The system has three layers: Operating System and System for Monitoring and Clustering, Application Server and Application Layer. The administration interface of the system was build and tested on Linux platform using the LAMP (Linux-Apache-MySql-Php) architecture as application server. The back-end server was build using the QT C++ Framework, which assures speed and easy extensibility with additional modules. The operating system is GNU/Linux – Ubuntu server with all necessary OS system services for clustering, monitoring and high-availability. The prototype application is written in PHP. The whole system is shown on Figure 4. The communication protocols were chosen in a way to enable easy integration with other systems. For communication between other systems HTTP communication is used. The generation of SMS messages complies the GSM 03.38 specification and SMIL 2.0 (Synchronized Multimedia Integration Language) for generation of MMS messages. The application for management uses HTML 4.0 and JavaScript.

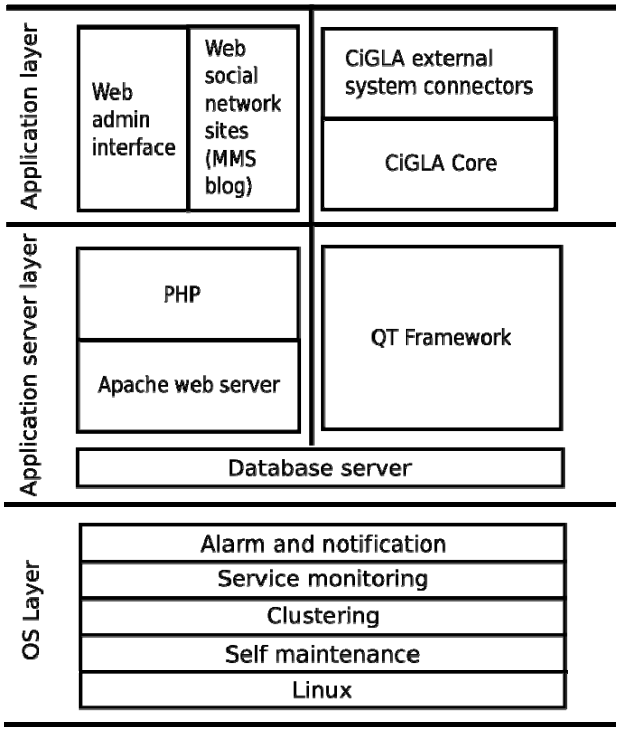


Fig. 4. The layered design of the system.

CiGLA is deployed as cluster server with two nodes which work in active – standby mode. The clustering and monitoring software monitors the state of the running services and in case of failure, it transfers the whole workload from the active to the standby node. The whole cluster is accessible through virtual IP address. This deployment strategy enables continuous work without service outage. The system can be deployed as a part of mobile operator's VAS platforms or as a third-party VAS provider.

5 Test results

Because of the specific architecture of the system, the tests were made in laboratory. The main reason for doing this is that in Macedonia there is no operator which has system for fine tracking of the subscribers and unavailability USSD gateway access. Complete functionality of all software components was tested each one on its own and the interaction between them. Tests were made in two separate stages: the core functionality and the web interface for administration. The location of the users was simulated by manually sending location information to the system. The message delivery was tested with Kannel SMSGW and Mbuni MMSGW connected to a mobile phone which was used for sending and receiving messages. The service requests were made via SMS and the content was delivered via SMS and MMS. The tests have shown that the system requirements were completely satisfied and the system works correctly in all specified circumstances.

6 Conclusion

The goal of the research to determine the tourist behavior and to design and build a mobile tourist guide is completely accomplished. If the requirements, expected results and actual results are compared it can be seen that the task is completed successfully. If the system is deployed in some mobile network there is a possibility that some interfaces to external system will have to be reconfigured or new interfaces to be developed. The built system is a good base for building new mobile services and extending its functionalities. This means extension of the context which characterizes the tourists: for example detection of the type of network the user is connected to 2G, 2.5G or 3G and sending richer content, audio or video streaming; Adding GIS server would enable locating and routing the user more precisely without the need users to be at predefined POI; The database of user actions, movement in the area or taken photos can be used for building online communities (social-networking web sites) where users can share photos and impressions of the places they visit.

References

1. B'Far, R.: Mobile Computing Principles – Designing and Developing Mobile Applications with UML and XML, 2005 Cambridge University Press, ISBN: 0-521-81733-1.
2. Meng, L., Zipf, A., and Reichenbacher T.: Map-based Mobile Services Theories, Methods and Implementations, 2005, Springer, ISBN 3-540-23055-6.
3. Loke, S.: Context-Aware Pervasive Systems - Architectures for a New Breed of Applications, 2007, Auerbach, ISBN 0-8493-7255-0.
4. Schiller, J., and Voisard, A.: Location-Based Services, 2004, Elsevier Inc,ISBN: 1-55860-929-6 .

5. Jagoe, A.: Mobile Location Services: The Definitive Guide, 2002, Prentice Hall PTR ISBN 0-13-008456-5
6. Hela, A., Haskell, B., Carter, J. L., Brice, R., Woelk, D., and Rusinkiewicz, M.: Anytime, anywhere computing - Mobile Computing Concepts and Technology, 2002, Kluwer Academic Publishers, ISBN: 0-306-47301-1
7. Huijinen, C.: Mobile Tourism and Mobile Government - An Inventory of European Project, 2006, European Centre for Digital Communication,ISBN: 9077743049
8. Sharda, N., Georgievski, M., Ahmed, I., Armstrong, L. J., Brogan, M., Woodward, A., Kohli, G., and Clark, M.: Leading-Edge Developments in Tourism ICT and Related Underlaying Technologies, 2006, CRC for Sustainable Tourism Pty Ltd, ISBN: 1920704841
9. Schwinger, W., Grün, Ch., Pröll, B., Retschitzegger, W., and Schauerhuber, A.: Context-awareness in Mobile Tourism Guides – A Comprehensive Survey,
10. Ralph, D., and Graham, P.: MMS Technologies, Usage and Business Models, 2004 John Wiley & Sons , ISBN 0-470-86116-9 .
11. Le Bodic, G.: Mobile Messaging Technologies and Services SMS, EMS and MMS, 2005 John Wiley & Sons, ISBN 0-470-01143-2
12. Retford, B., and Schwartz, J.: How to Build an SMS Service, 2007, O'Reilly, ISBN-10: 0-596-51513-8

Forecasting stock market prices

Miroslav Janeski, Slobodan Kalajdziski

Faculty of Electrical Engineering and Information Technologies, Skopje, Macedonia
miroslav.janeski@gmail.com,skalaj@feit.ukim.edu.mk

Abstract. In recent years, use of data mining and machine learning techniques in finance for such tasks as pattern recognition, classification, and time series forecasting have dramatically increased. However, the large numbers of parameters that must be selected to develop a good forecasting model have meant that the design process still involves much trial and error. The objective of this paper is to select the optimal parameters for designing of a neural network model for forecasting economic time series data. There is proposed a neural network based forecasting model for forecasting the stock market price movement. The system is tested with data from one Macedonian Stock, the NLB Tutunska Banka stock. The system is shown to achieve an overall prediction rate of over 60%. A number of difficulties encountered when modeling such forecasting model are discussed.

Keywords: neural network, backpropagation, financial forecasting, time series, stock forecasting

1 Introduction

Forecasting is a process that produces a set of outputs by given a set of variables. It assumes that some aspects of the past patterns will continue into the future. In general, forecasting process is working with historical data. Main idea is to discover patterns in the past and to use them in current moment. If we have a function which can map t-2 data to forecast t-1 data, then we can use it to predict t+1 data with t data. Forecasting financial time series is even more complicated task because there are a various sets of input data that can be analyzed. Methods for forecasting financial markets, especially stock market prices can be grouped into two types: technical and fundamental analysis.

Fundamental analysis of a market is consisting of analyzing overall state of the economy and overall state of the market where the observed companies are operating. With this analysis the stock price can be predicted reasonably. Fundamental analysis has very good performance for long term prediction, but there is also a huge disadvantage. This analysis is difficult to formalize for automated decision support because very often is highly subjective.

Technical analysis is a process for forecasting the future direction of prices of a stock market through the analysis of past market data, especially stock prices and volume. Unlike the fundamental analysis, technical analysis is used for short term

prediction. Although this analysis is widely used in forecasting it is not recommended because is highly subjective and statistically not valid.

Despite the above two analysis, there are several data mining techniques that offer very good performance in forecasting financial time series. In recent years, artificial neural networks [13] (ANN's) have become another important technique for predicting stock prices. A NN is an interconnected network of simple processing nodes with a different weight associated with each connection. With a proper network topology and appropriate weights between the connections, a NN can be trained to approximate any function mapping between its input(s) and output(s) by using an appropriate learning algorithm such as backpropagation (BP). In [5] is shown how a forecasting model can produce buy/sell signals, in [6,7 and 8] are shown best practices for data preprocessing, especially time series data, and in [2 and 4] are shown best practices in designing a good forecasting model.

2 Designing a neural network based forecasting system

The object of this paper is to summarize recommendations from the recent research [2 and 14] and to experimentally select the optimal parameters for designing a neural network based model for forecasting time series, especially predicting the daily closing price from one Macedonian Stock, the NLB Tutunska Banka stock. The model shall provide framework for testing different input parameters and different network topology parameters. These variations will be tested with training data, and then the optimal parameters will be selected for the testing data.

Fig.1 illustrates the proposed architecture of neural network. On the left (input) side of the neural network are different time series which will be input variables in the forecasting model. On the output side is one output node which represent the predicted time series, in this case closing price of the above stock.

2.1 Neural network overview

Recent research [1 and 2] activities in artificial neural networks have shown that ANN's have powerful capabilities in classification and regression problems. There are several advantages [2] that make them attractive in forecasting financial time series. First of all they are data – driven self adaptive methods, which means that ANN's can be used in field where we know a little or nothing about the relationships along the analyzed data. Second, they can generalize. ANN's can often correctly infer the unseen part of the data even if the data has a high noise coefficient. As forecasting is performed via prediction of future behavior (unseen part) from examples of past behavior, it is an ideal application for neural networks. And the third advantage is that the ANN's are universal approximators. They can approximate any kind of function, regardless it is linear or non linear. Despite the advantages there are disadvantages in the using of ANN's. One of the main disadvantages is the black box nature of their solutions, excessive training times, difficulty in replicating stable solutions, over

fitting and the large number of parameters that must be experimentally selected to generate a good forecast.

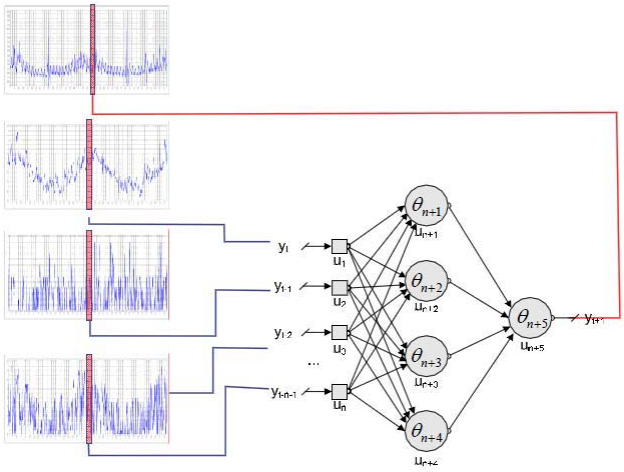


Fig. 1. Neural network architecture [14]

According to a recent survey [12] made by Crone/Kourentzes and [13] there is a decade's long research interest in artificial neural networks (ANN) that has lead to several successful applications. Besides their success, both in theoretical and practical works, there are still a lot of modeling issues [9 and 10] that are critical to design a reliable forecasting model. Kaastra [4] and [3 and 5] are proposing similar procedures for designing a neural network forecasting model which are consisted of several main tasks: variable selection, data preprocessing ,training and testing sets, neural network paradigm and neural network training. Every step has its own meaning and own part in the overall performance of the forecasting system.

2.2 Variable selection

Success in designing a neural network depends on a clear understanding of the problem. Our problem is to predict the daily closing price, so the traders can use it to make their trading more successful. Same as the mentioned type of analysis: fundamental and technical, there are fundamental and technical data that can be used as input variables. Fundamental data are mainly represented via fundamental parameters, like local and global economy and market parameters, which despite technical data are harder to interpret as an input variables for short term prediction. Technical data are in general prices and volume from historical trading on the stock market. In the [5] is proposed "keep it simple and short" KISS method which uses only technical data from every day trading and some derived technical data. In our approach we will use the same input variables which are:

O_i the opening price of day i
O_{i-1} the opening price of day i–1
H_{i-1} the daily high price of day i–1
L_{i-1} the daily low price of day i–1
C_{i-1} the closing price of day i–1
C_{i-2} the closing price of day i–2
V_i the trading volume of day i
M_{i-1} the 5-day momentum of day i-1(C_{i-1}–C_{i-6})

The last one, the 5 day momentum is used as a derived technical attribute, which is the difference between C_{i-1} and C_{i-6}. The output of the system will be the daily closing price of the stock. The historical data can be found on Macedonian Stock Exchange web site (http://www.mse.org.mk).

2.3 Data Preprocessing

The next step in designing a neural network is data preprocessing. Once the attributes are selected they should be adequately preprocessed in order to achieve higher quality and reliable predictions. Data preprocessing refers to analyzing and transforming the input and output attributes to minimize noise, highlight important relationships, and detect trends and seasons of the variables to assist the neural network in learning the relevant patterns. In recent research by Zhang [7 and 8] there are several techniques for detrending and deseasonalization the time series. It is shown that these techniques can influents on the overall forecasting performance of the system. In our approach these techniques will be left out of the analysis. Also, Sven Crone in his research [6] is experimenting with various scaling methods. His results show that scaling offers better performance than normalization. In our approach attributes from the data sets are scaled in interval [-1, 1], appropriate to the thresholds of the activation function which will be explained latter.

2.4 Training and testing sets

Defining the training and testing sets and strategy is probably the most important issue in designing a neural network. Training set is usually larger than the testing set and is used to train the neural network. The neural network model that has the best results [4] on the training set shall be used on testing set. In the nature of the time series is the time as a dimension which must be considered in the analysis. Therefore, a moving-windows approach [11] was adopted for training and testing.
Defining the appropriate size of the training set is the main issue in the moving-windows approach. If the training set is too small, neural network will convergence easily, but it will not predict with acceptable accuracy, on the other side if the training set is too large it will be more difficult for the network to convergence. Because time series are time depended, if the training set is too large (in meaning of time period) it may have learned some patterns that may no longer be effective because the market conditions are have already changed.

The benefit of the moving windows approach is that the observed market and stock may change over the time, so the constant moving the training and testing sets can avoid these changes. That means the trained network that was optimal in the past may not be optimal any more. Besides the recommendations [4 and 5] for using the moving-windows approach there are no defined rules how to define the training set size. The testing period was from May 2009 to May 2010. Because the period has 13 months the training set sizes with which was tested were less than a13 months. We tested with 12 months, 6 months and 3 months period, like a training set sizes. The appropriate training sets are described in [tables1, 2 and 3].

Table 1. 12 + 1 month training and testing sets

Cycle	*Training set*		*Testing set*	
	Period	No. of patterns	Period	No. of patterns
1	May 09 - Apr 10	201	May 10	16

Table 2. 6 + 1 month training and testing sets

Cycle	*Training set*		*Testing set*	
	Period	No. of patterns	Period	No. of patterns
1	May 09 - Oct 09	86	Noe 09	20
2	Jun 09 - Noe 09	98	Dec 09	19
3	Jul 09 - Dec 09	100	Jan 10	16
4	Aug 09 - Jan 10	104	Fev 10	20
5	Sept 09 -Fev 10	109	Mar 10	21
6	Oct 09 - Mar 10	115	Apr 10	19
7	Noe 09 - Apr 10	115	May 10	16

After preliminary experiments we have concluded that best performance is achieved when training set size is three month period, so three month period training set size is used for further experiments. On fig.3 are illustrated experimental results.

Table 3. 3 + 1 month training and testing sets

Cycle	*Training set*		*Testing set*	
	Period	No. of patterns	Period	No. of patterns
1	May 09 - Jul 09	37	Aug 09	15
2	Jun 09 - Aug 09	44	Sept 09	15
3	Jul 09 - Sept 09	42	Oct 09	19
4	Aug 09 - Oct 09	49	Noe 09	20
5	Sept 09 -Noe 09	54	Dec 09	19
6	Oct 09 - Dec 09	58	Jan 10	16
7	Noe 09 - Jan 10	55	Fev 10	20
8	Dec 09 - Feb 10	55	Mar 10	21
9	Jan 10 - Mar 10	57	Apr 10	19
10	Feb 10 - Apr 10	60	May 10	16

Another issue in the training process is the overtraining of the network which means that the network is trained too much, or it is too complex. Overtraining can be avoided if the mean square error is minimized during training. That means that the number of epochs should be decreased when MSE is increasing, which is done during the experiments.

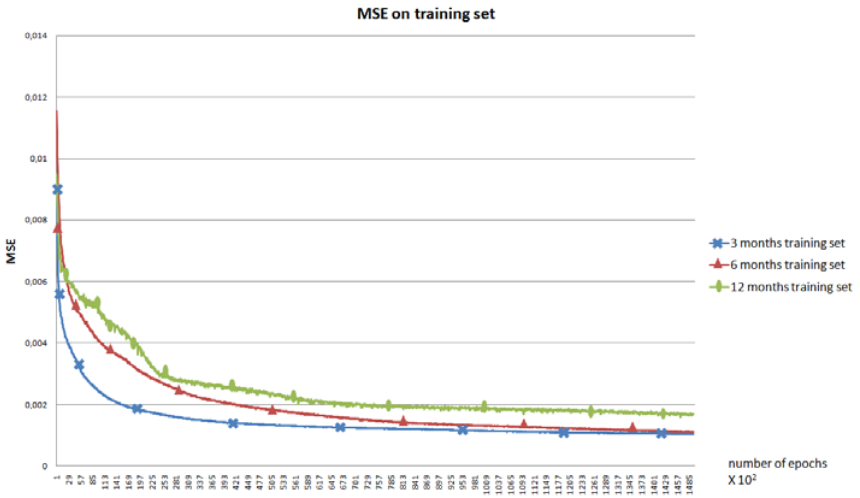


Fig.3. MSE on different training set sizes

2.5 Neural network paradigms

Defining neural network parameters is related with the problem of the observed domain. In general it is consist of defining the number of hidden layer, number of hidden neutrons, number of output neutrons, which are architectural considerations, and the dynamics of the neural network which is a transfer or activation function.

2.5.1 Neural network topology

Because the number of input and output variables are already defined with the attribute selection process the only thing remaining is the number of hidden layers and the number of hidden nodes. In theory, a neural network with one hidden layer with a sufficient number of hidden neurons is capable of approximating any continuous function. In practice [1 and 12], networks with one or two hidden layers are used in 90% of the cases. Large number of hidden layers than two can make the neural network time consuming and over fitted, which means that the network has too few degrees of freedom and memorize general patterns. In our system we used one hidden layer with different number of hidden neurons which was part of the experiments. So we have the following topology: 8 – k -1; k – number of hidden neutrons.

Besides the large number of examples of neural network models in the recent research there is no formula for determination the number of hidden neurons. Some recommendations [4] are that the number has to be from one half to three times the number of input nodes. Also in [5] is provided the following formula (1) which was used in our experiments

$$k = i \, x \, N - 1; \; i = 1, 2, 3... \qquad \textbf{(1)}$$

However, selecting the best number of hidden neurons involves experimentation. According to this we have tested with different number of hidden neurons, 7, 10, 14 and 21.

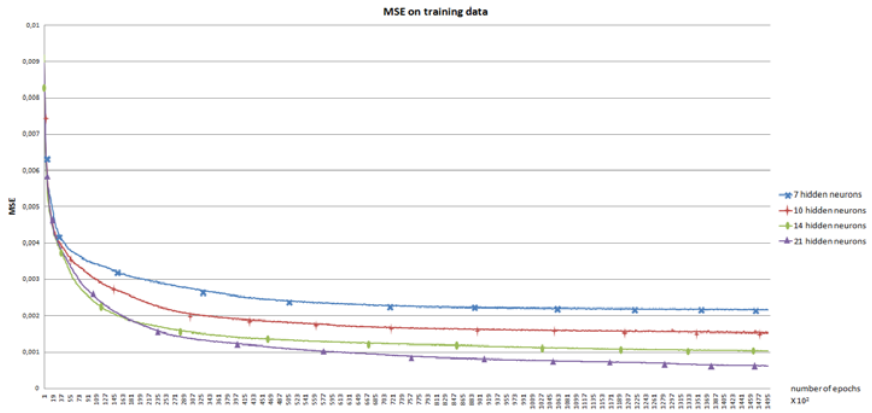


Fig. 4. MSE on different number of hidden neurons

Fig. 4 illustrates that the number of hidden neurons seriously has very high impact on the error rate of the neural network. The experiments have shown that neural network with 21 hidden neurons has lower error rate than the networks with less number of hidden neurons.

2.5.2 Transfer function

Transfer functions are mathematical formulas that determinate the output of a processing neuron. The large amount of neural networks models uses the sigmoid (S) function, besides the others like hyperbolic tangent, step, ramping, etc. In our system sigmoid function is used as a transfer function. Selection of the transfer function is often related with data scaling. To provide consistent system, data scaling in interval [-1, 1] was used same as the thresholds of the selected transfer function [-1, 1].

2.6 Training

Number of training iterations is another important issue in designing a neural network. There are two approaches in the literature [4] about selecting the right

number of training iterations. First school says that researcher should only stop training until there is no improvement in the error function. The second school says that there should be series of training test iterations, where neural network is trained on different number of iterations and tested on test data. In our system we use the first approach. Firstly we started with 150,000 epochs, but on fig.4 we have shown that there is no big improvement in MSE after 80,000 epochs, which was therefore used for the testing set.

Other important issues in designing a neural network are connected with the backpropagation algorithm. These issues are learning rate and momentum. The recommendations are that the learning rate should be <0, 5 and momentum = 0, 4.

3 Experimental results and discussion

In general the performance of the neural network is measured by its MSE error, because is one of the most used in recent research [12]. MSE error of the neural network was especially important in the training phase where the optimal parameters were selected according to the MSE error movement. From the results of the training phase (fig.4) we tested with two neural network architectures. One is 8 – 14 – 1 and the second is 8 – 21 – 1. MSE errors on different neural network architectures are shown in table 4. The results was that 8-14-1 architecture has minimum MSE error, therefore was selected for further testing.

Table 4. MSE errors on testing phase

# of hidden neutrons	*14*	*21*
MSE	0,046	0,086

However in the testing phase the actual value or usefulness of the neural network is measured by its ability to make accurate predictions of the future market prices. In our approach we use an average prediction rate. We recorded the predictions of the daily closing price of NLB Tutunska Banka stock over the 10 months test period (table 3) and compared them with the actual values. For each testing cycle we generated average prediction rate, and also an average prediction rate of the whole testing period.

Table 5 depicts the experimental results. Column labeled “Average Prediction Rate” show the results for neural network accuracy in different cycles. The average prediction rate of the system is over 60%, which is fairly good. In the cycles we have ones with very high accuracy and with very low accuracy. Main reason for these oscillations is the difference in the training set size. Besides the fact that training size set was limited to three month period, obviously it was not enough. In training sets we have sets that are two times larger than the other ones. In fact, that is from the nature of the time series. In some periods time series are more frequent, in other they are not.

This approach was effective [5] in stock exchange where stock price values were more constant during the time (in a manner of frequency).

Table 5. System accuracy

	Testing set			
Cycle	Period	# of patterns	Average Prediction Rate	MSE
1	Aug 09	15	37,69	0,0141
2	Sept 09	15	34,46	0,0733
3	Oct 09	19	66,34	0,1019
4	Noe 09	20	53,54	0,0887
5	Dec 09	19	58,05	0,0722
6	Jan 10	16	78,02	0,0351
7	Fev 10	20	84,39	0,0166
8	Mar 10	21	91,03	0,0075
9	Apr 10	19	70,75	0,0338
10	May 10	16	16,72	0,0102
Average Prediction Rate & MSE			60,98	0,0460

We have shown that for Tutunska Banka stock, we need more constant training set size than the proposed one. In future work we will test with same parameters for others stocks from Macedonian stock exchange and from other Balkan stock exchanges to check our hypothesis and to find the optimal mowing windows approach parameters.

Also, it is reasonable to expect that results can be improved in the following ways: including more technical, derived technical or even fundamental input attributes and fine tuning on network topology.

Finally, in this work we have used best practices in recent research and keep the simplicity in mind. Our approach uses the simple methods and recommendations to build neural network forecasting system which has an acceptable performance in accuracy and cost.

4 Conclusions

In this paper was elaborated a theoretical and practical analysis on the design and implementation of a neural based system for forecasting stock market prices, in detail the NLB Tutunska Banka Stock. Main objective was to determine the optimal parameters, especially the ones which still are generated via trial and error process. In this phase we have experimental results in defining the training set size in moving-windows approach, number of training epochs and number of hidden neurons. The design issues of neural networks for prediction tasks were discussed in detail and the design of forecasting system was presented. The empirical results showed that the proposed system was able to predict short-term price movement directions with an

overall accuracy of over 60%. The system demonstrated that fairly good prediction accuracy can be achieved using a backpropagation network without the use of extensive market data or knowledge.

References

1. Zhang G.P., B.E. Patuwo, M.Y. Hu, "Forecasting with artificial neural networks: the state of the art", Int. J. Forecast. Vol.14 pp.35–62, 1998
2. Adya, M. & Collopy, F. "How effective are neural networks at forecasting and prediction? A review and evaluation". Journal of Forecasting, Vol.17, pp.481-495. 1998
3. JingTao Yao and Chew Lim Tan. "Guidelines for financial forecasting with neural networks". In Proceedings of International Conference on Neural Information Processing, pages 757-761, Shanghai, China, November 2001
4. Kaastra L. and M. Boyd. "Designing a neural network for forecasting financial and economic time series". Neurocomputing, Vol 10, Issue 3, pp 215-236 April 1996,Financial Applications, Part II
5. Philip M. Tsang, Paul Kwok, S.O. Choy, Reggie Kwan, S.C. Ng, Jacky Mak, Jonathan Tsang, Kai Koong, Tak-Lam Wong," Design and implementation of NN5 for Hong Kong stock price forecasting", Engineering Applications of Artificial Intelligence Vol.20 pp.453–461 (2007)
6. Crone S., Lessmann S., Stahlbock R.. "The Impact of Preprocessing on Data Mining: An Evaluation of Support Vector Machines and Artificial Neural Networks", European Journal of Operations Research, Vol. 173, No.3, pp. 781-800, 2005
7. Zhang, G. P. & QI, M. "Neural network forecasting for seasonal and trend time series". European Journal Of Operational Research, Vol.160, pp. 501-514. 2005
8. Zhang G. P. , D. M. Kline, "Quarterly time-series forecasting with neural networks", IEEE Transactions on Neural Networks, Vol.18/6, pp.1800-1814, 2007
9. Crone S.. "Stepwise Selection of Artificial Neural Network Models for Time Series Prediction", Journal of Intelligent Systems, Vol. 14, No. 2-3,pp. 99-122, 2005
10. Crone S., R. Dahwan. "Forecasting Seasonal Time Series with Neural Networks: A Sensitivity Analysis of Architecture Parameters", Proceedings of the International Joint Conference on Neural Networks, IJCNN'07, Orlando, USA, IEEE: New York, 2007
11. Skabar, A., Cloete, I., 2002. Neural networks, financial trading and the efficient market hypothesis. Australian Computer Science Communications 24 (1), 241−249
12. Kourentzes, N. Crone S. "Advances in forecasting with artificial neural networks", working paper, Lancaster University Management School, Lancaster UK 2010
13. Dorffner, G, "Neural Networks for Time Series Processing". *Neural Network World* Vol.4/96, pp. 447-468, 1996.
14. Crone S.F.. "Forecasting with Artificial Neural Networks", EVIC 2005 Tutorial Santiago de Chile, 15 December 2005

Measuring Teaching Quality in Higher Education: Instrument For Collecting Student Feedback

Danijel Mijic

University of East Sarajevo, Faculty of Electrical Engineering, Vuka Karadzica 30, 71123 East Sarajevo, Bosnia and Herzegovina
danijel.mijic@etf.unssa.rs.ba

Abstract. Quality assurance and continuous evaluation of higher education study programmes is one of the major tasks set for the higher education institutions. European standards for quality assurance in higher education define important role of students in quality assurance process. One way of getting feedback about teaching process quality is evaluation using students' questionnaires. This paper describes some previous experiences in students' evaluation of teaching quality at University of East Sarajevo, Faculty of Electrical Engineering and also gives a brief overview of web-based application that is currently in use as an instrument for getting feedback from students at the university level.

Keywords: quality assurance, higher education, evaluation, teaching, information system.

1 Introduction

Quality assurance plays important role in higher education. In order to retain and improve their position at the educational services market, institutions of higher education need to pay close attention to the quality of their services. One of the processes that is crucial for improving overall quality is a teaching process. Measuring teaching process quality involves getting information about various aspects of teaching process and usually includes measuring quality of teachers, quality of individual course units, and quality of complete academic programmes.

There are many instruments used for measuring quality indicators in higher education, as described in [1], [2], [3]. Some of them are focused on measuring quality of individual course units and teachers, while others tend to provide information about quality of complete academic programmes or institutions. One common thing for both of the mentioned types of instruments is that they are mainly used for getting feedback from students in order to get information about students' perception of various aspects of teaching process. As students are directly involved in teaching process, their evaluation of teaching quality plays very important role in measuring overall teaching quality.

This paper describes an instrument for measuring teaching quality that is developed and used at University of East Sarajevo, Faculty of Electrical Engineering,

from the academic year 2006/07 to the present day. The paper presents some experiences and results using the instrument at Faculty of Electrical Engineering and gives a brief overview of a improved web-based application that is currently in use at the university level.

2 Instruments For Measuring Teaching Quality

Students' role in measuring teaching quality is well known and established in some regions in the world. This is especially true for North America and Australia, where standardized and proven instruments are used for getting students' feedback, like SEEQ (Student Evaluation of Educational Quality) and CEQ (Course Experience Questionnaire).

The most widely used student questionnaire in the USA is the SEEQ. This instrument is not based on student learning research but on psychometric analysis [4]. Apart from the USA, the instrument is also used in many other countries with some modifications reflecting the education system or only like a translated version.

The SEEQ is developed by Professor Herbert W. Marsh of the University of Western Sydney, Macarthur. There is a series of publications describing research, methodologies and the SEEQ instrument [5], [6], [7], [8].

The CEQ instrument is developed by Professor Paul Ramsden, the Graduate Careers Council of Australia and the Australian Commonwealth Department of Education, Training and Youth Affairs. The CEQ instrument is widely accepted at Australian universities [1]. It is also adopted for the other purposes different from the original one, with some modifications and additions according to requirements and the application, as described in [9], [10].

In the European higher education area students' role in quality assurance is defined in Standards and guidelines for QA, published by European Association for Quality Assurance in Higher Education (ENQA). Many higher education institutions have developed their own instruments for getting feedback from students, but there is not much published research evidence on proven reliability and validity of results of using these instruments like it is the case with the SEEQ and CEQ.

As noted by Marsh (1987), students' role in measuring teaching quality is important for the several reasons:

- "Diagnostic feedback to teachers about the effectiveness of their teaching,
- A measure of teaching effectiveness to be used in administrative decision making,
- Information for students to use in the selection of course units and teachers,
- An outcome or process description for use in research on teaching."

Traditional way of getting students' feedback is by using paper forms with a list of statements to which students indicate their level of agreement. The statements are chosen to appropriately reflect various aspects of teaching. The level of agreement is mostly indicated using five point scale, called Likert scale, with descriptive optional answers, like "Strongly agree" to "Strongly disagree", or similar scales that enable

paper forms to be scanned and processed for getting results. The paper forms are distributed to students participating the evaluation usually in two ways. One is the class environment where students are asked to fill out the questionnaires and return them immediately. The other way is sending questionnaires to students by classical postal system and expecting them to return them in the same way. However, when larger groups of students are involved in the evaluation, this means of getting students' feedback becomes time-consuming and inefficient when it comes to collecting and processing results.

An alternative solution to traditional one mentioned earlier became available with the increased use of information technologies in higher education. The paper form questionnaires are transformed into the online forms with the similar content. In this way results are automatically stored into electronic form suitable for further processing and analysis. The online questionnaires have advantages over the traditional ones, the main advantage is improved efficiency. A few examples of instruments for online evaluation are described in [9], [10], [11].

3 Method and the Evaluation Instrument

Students' evaluation of teachers and course units at University of East Sarajevo, Faculty of Electrical Engineering (FEE), was formally introduced in 2002. At first, the traditional way of getting students' feedback was used. At the end of each semester, during the evaluation period of two weeks, students filled out a printed questionnaire which contained statements and questions for evaluation of teaching quality of individual teachers and quality of individual course units in the current semester. Students are required to respond to statements and questions using predefined answers in the form of rates from 5 to 10, or using Yes/No answers. It was also possible to write a free comment for every teacher or course unit in the questionnaire. The ratings 5 to 10 are selected because the same rating system is used in local higher educational system for rating students at exams and assessments.

This form of students' evaluation was used for several years and achieved positive results with respect to getting feedback from students about their perception of teaching process quality. However, this means of getting students' feedback shown itself as inefficient and very time-consuming when it comes to processing and analyzing results. Since all of the active students are required to fill out the questionnaire at the end of each semester, there was quite a lot of paper forms to be manually sorted and processed in order to get the results. The sample size was usually about 200 to 250 active students involved in the four year undergraduate programme. Every student had to evaluate teaching performance of each teacher and for each course unit in the semester. Data from the returned paper forms was entered manually into Microsoft Excel sheets in order to perform statistical analysis and generate reports. Textual comments were also entered manually for each paper form, and appended to the summary results for each teacher and course unit. Even with the relatively small sample, the work that had to be done for processing the results was significant and could not be done in the short time. This was the main reason for initiating development of application for online evaluation of teaching quality.

The application for online evaluation of teaching quality was developed in the form of multi-user web application. It enables students to fill out online questionnaires having similar content to the traditional ones, but it also enables students, teachers and administration to access the results of students' feedback. The application can be accessed from any place having Internet connection by using standard web browser.

The online questionnaires uses four scales for evaluation of course units and teachers, overall students' course satisfaction and evaluation of institution's resources used as a support for teaching. Scale for evaluation of course units and teachers consists of three questions with Yes/No answers and a free comment field, while scale for evaluation of teachers consists of eight statements which are rated on the scale from 5 to 10 and one free comment field. Scale for rating overall students' course satisfaction and scales for rating each of the institutional resources used for supporting teaching consist of one statement with rates from 5 to 10 and a free comment field.

A screenshot of students' interface for evaluation of teachers is shown on Figure 1.

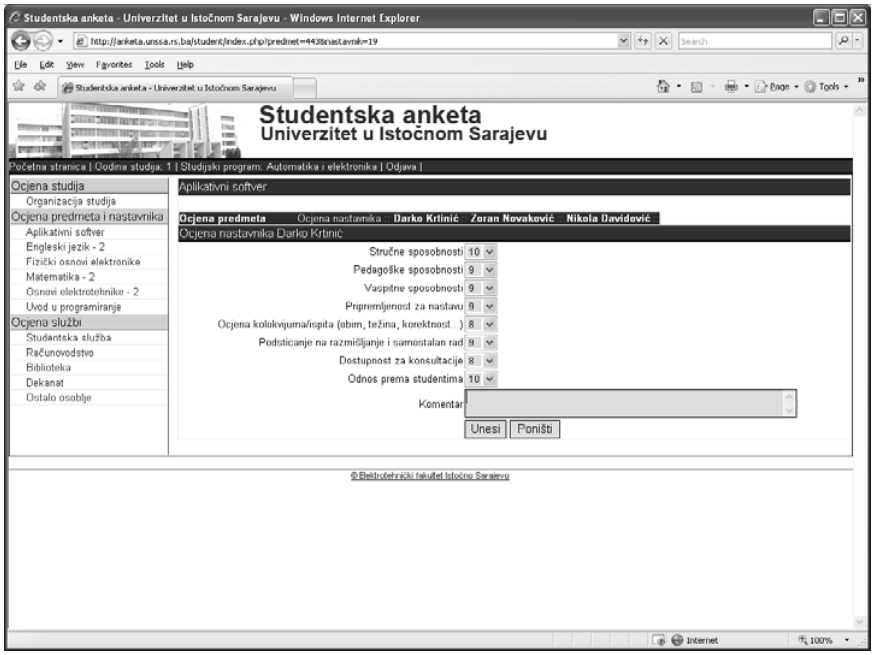


Fig. 1. Students' interface for evaluation of teacher

In the left part of the screen shown on Figure 1 students can see all course units available in the current semester. By clicking the course unit name, it is possible to rate the quality of the course unit itself and all the teachers for that course unit. Questions and statements for rating the course units and teachers are shown in the central part of the screen. All of the questions and statements, except the comment text field, are required for input. Rates are chosen on the scale from 5 to 10 from the

select list box. Except rating the course units and teachers, students can rate their overall satisfaction with the study programme and some institutional resources used for teaching purposes. For each element of a questionnaire (e.g. rating course unit or individual teacher on the course unit) it is possible to provide only one rating.

The application was developed using open source technologies on the LAMP platform (Linux, Apache, MySQL, PHP). One of the main reasons for choosing these technologies was their free availability, with no additional financial expenses for software licenses. This is very important factor for universities in Bosnia and Herzegovina, which have limited financial resources when compared to other universities in developed European countries. The other important reason for choosing these technologies was the existing information system that is used at University of East Sarajevo (UoES) and at all of its organizational units. The DBMS that is used in UoES information system is also MySQL. The application uses much of the existing data from the UoES information system, especially data about students, teachers and course units.

3.1 Students Evaluation at the University Level

UoES was transformed to an integrated university in 2008. The term "integrated" here means that the University is the only legal body that represents all of its organizational units. It consists of 15 faculties and 2 academies that are geographically located mainly in the Eastern part of Bosnia and Herzegovina. In the new organizational scheme quality assurance is governed by a body constituted at the University level. This body is called Committee for Quality Assurance (CQA) and it is responsible for all activities regarding quality assurance at the University and all of its organizational units – faculties and academies. CQA activities are coordinated to the organizational units by university-level Quality Assurance Coordinator. Each organizational unit also has a local Quality Assurance Associate which is a person responsible for coordination of QA activities at organizational unit level.

CQA coordinates students' evaluation of teaching quality at the University level. The instrument that was used at the UoES for students' evaluation was a traditional one, using paper forms for getting feedback from students and scanning returned forms to process and analyze results. The students' evaluation was performed once in academic year, at the end of the second semester. The questionnaires used for getting students' feedback included statements and questions for evaluation of teachers and course units in general, not on the individual basis. Using this kind of questionnaires it was not possible to conclude about teaching quality of individual teachers nor the quality of individual course units, and consequently it was not possible to act in the right direction in order to improve quality. This is why the CQA decided to change methodology and the instrument for obtaining students' feedback wanting to get more detailed feedback about individual teachers and course units. However, using traditional instrument in this sense was not appropriate since the level of expected information was multiplied. Instead of one general questionnaire for all teachers and course units, now it was needed to cover all course units and all teachers involved in those course units. The easiest and most cost-effective solution to the problem was to accept the instrument that was developed and used for several years at FEE, but with

some necessary modifications and improvements required to apply the existing system to the UoES and all of its organizational units. The required modifications are made and the system is currently in use at the University level. In the time of writing this paper students' evaluation of teaching is in progress at 14 organizational units. Expected response rate for each organizational unit is minimum 50%. Results of the evaluation will be available at the end of the evaluation period, which is the end of June, 2010 at most of the organizational units.

A screenshot of interface for application administrator, displaying response statistics for selected organizational unit and the number of responses from all organizational units, is shown on Figure 2.

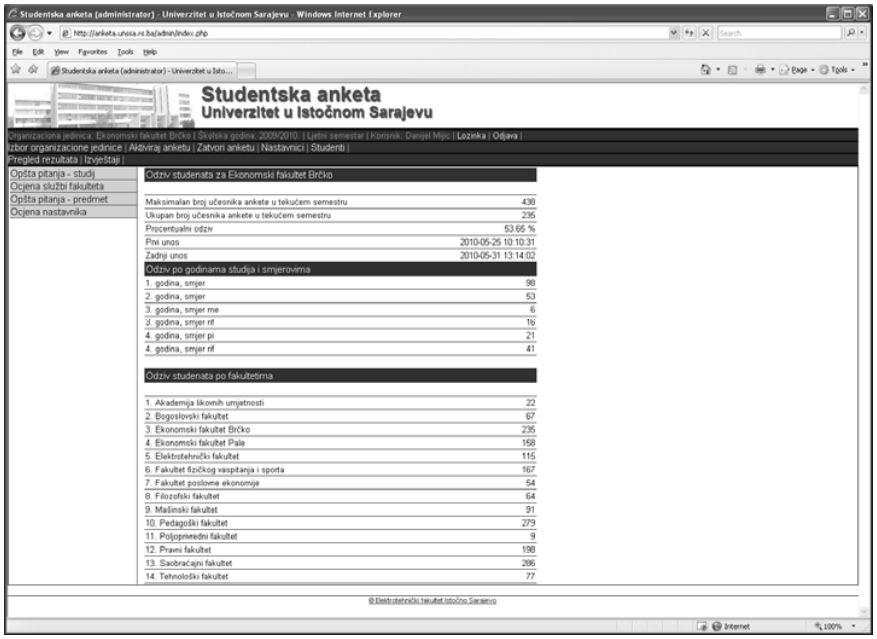


Fig. 2. Administrator's interface

3.2 Users

Users of the application are students, teachers, quality assurance coordinators and management staff at university and organizational unit levels. Access to the application is protected using username and password. After authentication and authorization, users are allowed to use specific application functionalities and to access results according to their privilege level. The important note is that application ensures anonymity for students participating the evaluation. The application use-case diagram is shown on Figure 3.

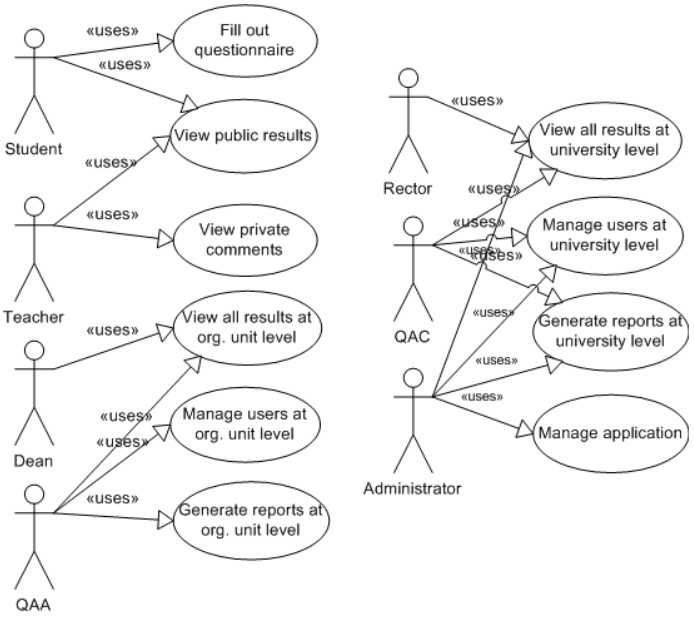


Fig. 3. The application use-case diagram

4 Results

Since the beginning of use at FEE in second semester of academic year 2006/07, the application collected feedback from 1015 students about 86 individual teachers and 172 individual course units. The overall sample size was 1226 students and overall response rate is 82.79%. Response rates in specific academic years and semesters are given in Table 1.

Table 1. Response rates through academic years and semesters.

Academic year	Semester	Sample	Response	Response rate
2006/07	2	208	186	89.42%
2007/08	1	236	215	91.10%
2007/08	2	170	115	67.65%
2008/09	1	186	155	83.33%
2008/09	2	186	144	77.42%
2009/10	1	240	200	83.33%

4.1 Students' Reflections on Using the Instrument

Using the instrument was considered convenient by most of the students. They were not limited to time and place for filling out the questionnaire, instead they could do it any time during the evaluation period from the privacy of their homes or other places having Internet connectivity. They were not under pressure to fill out questionnaires in the class environment, so they felt more comfortably. Response rates shown in Table 1 are well above response rate of 50% that is mostly considered to be a minimum value for accepting results as representative from the given sample [1].

4.2 Teachers' Reflections on Using the Instrument

Teachers' reflections on the instrument were generally positive. In some minor number of cases they greatly depend on the results obtained from the students' feedback. Some teachers complained that students are not competent to evaluate their work, or some said that teachers' rates depend on the course unit complexity.

One of the most important advantages of using this instrument, as with the other similar instruments for getting students' feedback about teaching quality, is the process of self-regulation. Students' perception of teaching quality helps teachers to identify potential problems and improve quality of their teaching. Since the teachers are able to access results of students' evaluation for the other teachers, they are additionally motivated to give their best in teaching process and compete with other teachers in order to get higher ratings.

4.3 Influence on Quality

Students' evaluation of teaching quality is not by itself enough to improve teaching quality. Having a collection of students' evaluations of teaching quality doesn't guarantee the quality improvement. This could be true for the several reasons, but one of the most important reason, especially in Bosnia and Herzegovina higher education, is lack of institutional policies that regulate interpretation, analysis of results and necessary actions based upon results of evaluation in order to improve quality.

In the case of UoES, in the present situation there is no procedures defined for acting upon results of students' evaluation of teaching quality. The only influence on quality is done by the process of self-regulation. Teachers with lower ratings can identify segments of their work that are to be improved in order to improve teaching quality. The same applies to individual course units and the other elements of teaching process that are subject to evaluation. As quality assurance in higher education is becoming more and more important in Bosnia and Herzegovina and the region, it is likely that the institutional policies will very soon incorporate appropriate mechanisms for taking into account results obtained by students' evaluations of teaching quality.

Figure 4 shows trend of overall course satisfaction of students over the several years and semesters. The ratings are displayed on a scale from 5 to 10.

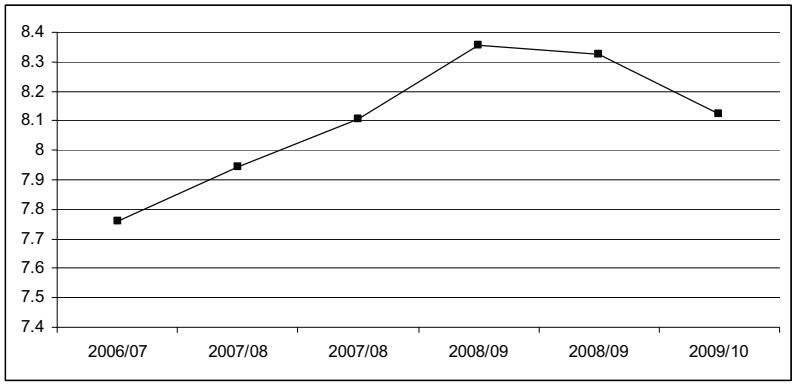


Fig. 4. Trend of students' overall course satisfaction

Based on results shown in Figure 4 it could be concluded that the students' overall course satisfaction has positive trend from second semester of academic year 2006/07 to second semester of academic year 2008/09. Next two evaluations resulted in small negative trends. The results shown could be interpreted as a quality improvement, but there is no evidence that this is due to some actions taken in order to improve quality.

Results of all students' evaluations are stored in a database and can be accessed in order to analyze trends of specific quality indicators. Since the application is in function from the second semester of academic year 2006/07 there is not enough data to establish long term trends of quality indicators. This is especially true for evaluation of teachers. Figure 5 shows trends of overall teaching performance of randomly selected teacher on two course units which are in a study programme for the second semester in academic years 2006/07, 2007/08 and 2008/09.

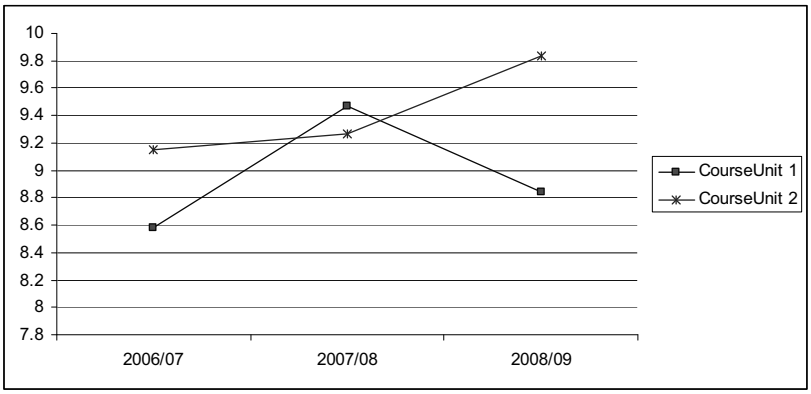


Fig. 5. Trend of teacher's ratings for two course units in the second semester

Figure 6 shows trends of overall teaching performance of the same teacher on two course units which are in programme for the first semester in academic years 2007/08, 2008/09 and 2009/10.

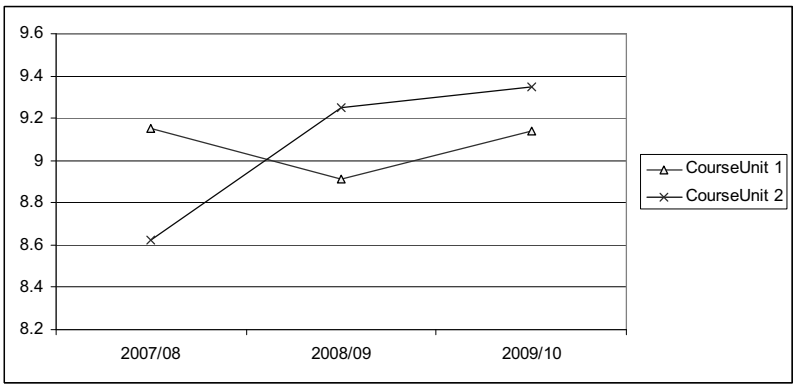


Fig. 6. Trend of teacher's ratings for two course units in the first semester

5 Summary and discussion

The instrument used for measuring teaching quality at FEE has many benefits when compared to the previous methods of obtaining students' feedback. The main benefit is it's efficiency in getting students' responses and processing results. In the previous solution using traditional means of getting feedback from students it took weeks to collect, process and publish results of the students' evaluation of teaching quality. Using the instrument for online evaluation this time is greatly reduced and completely eliminated. At the end of evaluation period results are already available and ready for publishing and analysis. Partial results are available even during the evaluation period, in the real time.

Another important benefit is the availability of results of evaluations in electronic form which is convenient for further processing and analysis. Based on the results from the previous evaluations one could conclude about trends of specific quality indicators. This way it could be detected if the evaluation is having impact on the quality improvement. The results could also be aggregated and transformed in form convenient for advanced reporting and analysis which is required for administration staff as a support for decision making. Using OLAP tools data could be analyzed through different dimensions by persons with limited IT expertise. The application usage on the university level will dramatically increase the amount of data that needs to be processed and analyzed. This is the reason for further developments and improvement of the application in order to incorporate some business intelligence elements.

By using the application the University management will have more detailed insight in teaching quality indicators when compared to previous solution which was very limited. This model of the application should offer enough flexibility in order to apply it on the other similar institutions having different organizational scheme. It should also offer integration with other existing proprietary systems for the purpose of using the existing data where the required data is already available.

For the better success of the students' evaluation of teaching quality it is necessary to motivate relevant population of students and to achieve high response rates. Students should be aware that their response will be considered and taken into account and they can improve their courses by being more involved in the process of quality assurance.

Teachers also have to be motivated to react in response to students' evaluation. Self-regulation is not always sufficient and higher education institutions should take more care about defining relevant policies and procedures in order to improve overall teaching quality.

References

1. Richardson, J. T. E.: Instruments for obtaining student feedback: a review of the literature, Assessment & Evaluation in Higher Education Vol. 30, No. 4, pp. 387--415 (2005)
2. McInnis, C., Griffin, P., James, R., Coates, H.: Development of the Course Experience Questionnaire (CEQ), Centre for the Study of Higher Education and Assessment Research Centre, Faculty of Education, The University of Melbourne (2001)
3. Hung, W., Smith, T. J., Harris, M. S., Lockard, J.: Development research of a teachers' educational performance support system: the practices of design, development, and evaluation, Association for Educational Communications and Technology (2007)
4. Coffey, M., Gibbs, G.: The Evaluation of the Student Evaluation of Educational Quality Questionnaire (SEEQ) in UK Higher Education, Assessment & Evaluation in Higher Education, Vol. 26, No. 1 (2001)
5. Marsh, H, W.: A Longitudinal Perspective of Students' Evaluations of University Teaching: Ratings of the Same Teachers over a 13-Year Period, Paper presented at the Annual meeting of the American Educational Research Association, San Francisco, CA (1992)
6. Marsh, H, W.: Students' Evaluations of University Teaching: Research Findings, Methodological Issues, and Directions for Future Research, International Journal of Educational Research, 11, 253--387 (1987)
7. Marsh, H, W.: Students' Evaluations of College/University Teaching: A Description of Research and an Instrument, Annual Meeting of the Australian Association for Research in Educational, Sydney, Australia (1980)
8. Beatty, B., Marsh, H. W.: Students' Evaluations of Instructional Effectiveness: Research and a Survey Instrument, University of California (1974)
9. Tucker, B., Jones, S., Straker, L., Cole, J.: Course Evaluation on the Web: Facilitating Student and Teacher Reflection to Improve Learning, New Directions for Teaching & Learning, Issue 96, p81--93 (2003)
10. Tucker, B., Jones, S., Straker, L.: Online student evaluation improves Course Experience Questionnaire results in a physiotherapy program, Higher Education Research & Development, Vol. 27 Issue 3, p281--296 (2008)
11. Harrington, C. F., Reasons, S. G.: Online Student Evaluation of Teaching for Distance Education: A Perfect Match? The Journal of Educators Online, Volume 2, Number 1 (2005)

Evaluation methodology for national enterprise architecture frameworks

Magdalena Kostoska, Marjan Gusev and Kiril Kiroski

Institute of Informatics, Faculty of Natural Sciences and Mathematics,
St. Cyril and Methodious University,
Skopje, Republic of Macedonia
{magi,kiril}@ii.edu.mk, marjangusev@gmail.com

Abstract. This article introduces a new methodology for NEA (National Enterprise Architecture) evaluation based on two sets of indicators. The first set consists of quantitative indicators regarding to the e-government initiatives and their characteristics, including interoperability strategy, NEA strategy, legal framework and legislatives, sustainability, quality control, dissemination and predicted goals. The second set is based on the definition of NEA and NEA characteristics, consisting of the following quantitative indicators: methodology and models, model interfaces, architectural standards and principles, repository and use of development indicator. The final result of the evaluation is calculated as a sum of the results of the both sets of indicators.
In this paper we analyze how similar efforts resulted in smaller methodology and evaluate several NEA initiatives with comparison of results. Further we comment how our approach results in a new sophistication methodology that can be used to evaluate NEA initiatives and help others in giving directions to build new sophisticated solutions.

Keywords: National enterprise architecture evaluation framework, quantitative evaluation framework, e-Government, e-Services benchmarking

1 Introduction

E-government has been studied from many perspectives. OECD has published several evaluations for distinct e-governments and country comparison reports. [1][2][3][4] [5][6][7][8][9][10][11] Surveys of national enterprise architectures (NEA) have been scattered and unrelated. Most detail study of NEA activity has been published by Christiansen and Gotze [4]. The goal of their research is to gain a detailed report of the NEA activities, but the results in their report are summarized for all the countries.

There have also been other studies, but most of the reports of NEA that have been produced use descriptive evaluation framework.

Leist and Zellner have conducted a study that evaluates part of the well-known, reputable and often used frameworks (regarding the elements of a method), including few national architecture frameworks. [5] Methodology introduced in their report

consists of five constitutive elements: 1) Meta model, 2) Procedure model, 3) Technique/modeling technique, 4) Role and 5) Specification document.

Each of the elements is evaluated as fully accomplished, partly accomplished or not accomplished. A summary of their results is given in Figure 1.

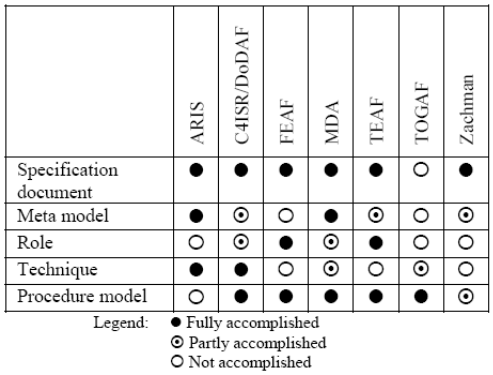


Fig. 1. Evaluation of architecture frameworks according to the study of Leist and Zellner [5].

Another detailed study of the national enterprise architecture models is conducted by the Finnish Enterprise Architecture Research Project. The research report is an overview of enterprise architecture work in 15 countries. The goal was to support the development of the Finnish state IT function and, in particular, the enterprise architecture work of public administration carried out in the Interoperability Development Programme by evaluating this work in relation to foreign development. [2] The overview describes and compares different countries' enterprise architecture programmes on a descriptive evaluation framework which helps to evaluate the contents, focus areas and perspectives of the enterprise architecture programmes in different countries. [2] The framework is based on Janssen and Hjort-Madsen's framework [6] of five viewpoints: 1) Policies, actors and structures; 2) Governance; 3) Architecture frameworks and methodologies; 4) Architecture principles and standards, and 5) Implementations.

Also two new viewpoints are added: 1) Reported benefits of NEA work and 2) Evaluation of national EA work based on observations. The extended framework used in the report is presented in Table 1.

2 A new evaluation model

Creation of a new NEA methodology is a challenging process. Main reasons can be found due to the fact different plans, goals, expected benefits and different governance models of each country.

Our survey introduces a new methodology for NEA evaluation. We have created new evaluation framework that consists of two sets of quantitative indicators. The

first set consists of quantitative indicators regarding the e-government initiatives and their characteristics. The second set consists of the NEA definition and characteristics.

Table 1. FEAR-project evaluation framework [2].

Viewpoint	Explanation
1. Policies, actors and structures	Political and environmental drivers for NEA. The strategic objectives for architecture are provided by political actors and constrained by democratic structures.
2. Governance	NEA's governance model and practices that are needed for keeping the architecture up-to-date. Governance guidelines also encourage desired behavior.
3. Architecture frameworks and methodologies	Definition of the NEA, framework used and architecture process.
4. Architecture principles and standards	Standards, principles and guidelines used for implementation, and change management. Compatibility with international models (e.g. EIF and FEA)
5. Implementations	NEA implementations and cross-public sector projects.
6. Benefits	Benefits of the NEA work and their measurement, experiences from NEA work and its usefulness.
7. Evaluation	Special characteristics and advantages/disadvantages of the NEA work.

2.1 E-government initiatives indicators

We introduce the following seven indicators in the first set, and define quantitative and quality measure:

1. **(IS1) Interoperability strategy**
   - 0 – have not been published and it is not planned at all
   - 1 – published, but yet uncompleted strategy
   - 2 – long-term plan for strategy publishing
   - 3 – short-term plan for strategy publishing
   - 4 – published complete strategy
2. **(IS2) NEA strategy**
   - 0 – have not been published and it is not planned at all
   - 1 – published, but yet uncompleted strategy
   - 2 – long-term plan for strategy publishing
   - 3 – short-term plan for strategy publishing
   - 4 – published complete strategy
3. **(IS3) Legal framework and legislatives**
   - 0 – no mandatory use
   - 1 – suggested as common practice
   - 2 – only some parts of the NEA are mandatory
   - 3 – the usage of NEA is still not mandatory, expected in the future
   - 4 – the usage of NEA is mandatory

4. **(IS4) Sustainability**
   - 0 – no defined model
   - 1 – governance model is only defined, but not used yet
   - 2 – decentralized governance,(administrative bodies are responsible)
   - 3 – insufficient budget and centralized model defined
   - 4 – centralized model is defined and used
5. **(IS5) Quality control**
   - 0 – no defined quality control model
   - 1 – quality control model is only defined, but not used yet
   - 2 – decentralized governance of quality control model, (administrative bodies are responsible)
   - 3 – insufficient budget and centralized quality control model defined
   - 4 – centralized quality control model is defined and used
6. **(IS6) Dissemination**
   - 0 – no defined model nor budget
   - 1 – governance model and budget is only defined, but not used yet
   - 2 – each administrative body defines model and budget
   - 3 – insufficient budget and centralized model defined
   - 4 – centralized model defined and used
7. **(IS7) Predicted goals** – for each of the listed goals 1 point is gain
   - Resource management
   - Improved service delivery
   - Infrastructure renewal
   - Improve cross governmental interoperability
   - Improve process effectiveness
   - Reduce time to deliver IT projects
   - Enable greater flexibility in business processes
   - Reduce IT cost
   - Improved IT security

A further description of the used indicators is given in Table 2.

Table 2. E-government initiatives indicators description.

	Indicator	Description
IS1	**Interoperability strategy**	Evaluates the national interoperability framework, the way that it is composed into the national strategy, so as the standards that includes.
IS2	**NEA strategy**	Evaluates the NEA framework and the way that it is composed into the national strategy.
IS3	**Legal framework and legislatives**	Evaluates the existence or type of legal model of conducting the published NEA framework.
IS4	**Sustainability**	Evaluates the existence or type of governance of the published NEA model.
IS5	**Quality control**	Evaluates the existence or type of quality control model of the published NEA performance.
IS6	**Dissemination**	Evaluates the existence or type of marketing budget model for national strategy propagation like awareness rising, EA staff training etc…
IS7	**Predicted goals**	Evaluates the predicted goals and motivations of the published NEA framework.

2.2 NEA indicators

We introduce the following five indicators in the second set and define quantitative and quality measure:

1. **(IA1) Methodology and models**
   - 0 – there is no published methodology for NEA usage
   - 1 – publishing of NEA usage methodology is planned
   - 2 – published NEA usage methodology, but only for a part of the development process
   - 3 – published complete NEA usage methodology
   - 4 – published several complete NEA usage methodologies
2. **(IA2) Model development interfaces**
   - 0 – no defined tools
   - 1 – usage of graphical model development tools with no automatic translation between models
   - 2 – usage of graphical model development tools with partial automatic translation between models
   - 3 – usage of graphical model development tools with complete automatic translation between models
3. **(IA3) Architectural standards and principles** – for each of the listed standards and principle 1 point is gain
   - Technical standards and principles
   - Conceptual standards and principles
   - Organizational standards and principles
   - Procedural standards and principles
4. **(IA4) IT Repository**
   - 0 – no defined repository
   - 1 – usage of several tools is planned, but no central repository exists
   - 2 – usage of repository that contains partial information of tools and libraries
   - 3 – usage of repository that contains partial information of tools, libraries and usage examples
   - 4 – usage of repository that contains complete information of tools, libraries and usage examples
5. **(IA5) Usage of development indicator** – 1 point is given for each of the listed development indicator
   - Qualitative measurements indicators
   - Performance measurements indicators like number of services/models etc...
   - Existence of NEA reevaluating administrative body
   - Other indicators

A further description of the used indicators is given in Table 3.

Table 3. NEA indicators.

	Indicator	Description
IA1	**Methodology and models**	Evaluates the published methodology and models for NEA development.
IA2	**Model interfaces**	Evaluates the presence of graphical model development tools for NEA development.
IA3	**Architectural standards and principles**	Evaluates the presence and type of architectural standards and principles for NEA development.
IA4	**IT Repository**	Evaluates the presence of IT repository, containing the common architectural artifacts such as development tools, common information etc....
IA5	**Usage of development indicator**	Evaluates the presence of defined development indicators that measures the progress of NEA.

2.3 Evaluation

In our new methodology we propose normalization and calculation of indicator values' average. The final evaluation represents a sum of the both sets of indicators, described as follows.

The result from the first set of indicators (IS) is calculated with the use of the following formula:

$$IS = \frac{IS1 + IS2 + IS3 + IS4 + IS5 + IS6 + \frac{4 * IS7}{9}}{7}. \quad (1)$$

(1) represents the average of all e-government initiatives indicators and all the indicators that have the same weight. Following the same principle, the indicator IS7 is normalized to have value in the range between 0 and 4 (as the values of the rest of the indicators), since it's value may vary between 0 and 9. The maximum value that can be achieved is 4.

The result from the second set of indicators (IA) is calculated with the use of the following formula:

$$IA = \frac{IA1 + \frac{4 * IA2}{3} + IA3 + IA4 + IA5}{5}. \quad (2)$$

(2) represents the average of all NEA indicators and all indicators that have the same weight. By using the equality principle, the indicator IA2 is normalized to have values within the range of 0 and 4 (as the rest of the indicators), since it's value may vary between 0 and 3. The maximum value that can be achieved is 4.

The final result of the evaluation (IV) is calculated as a sum of the results of the both sets of indicators (with the following formula):

$$IV = IS + IA. \quad (3)$$

The maximum value that can be obtained with the use of the formula is 8.

3 Comparison to existing evaluating methodologies

As mentioned before, creation of a NEA methodology is a challenging process. Some reasons that make this process difficult are the following:

- Each country creates a national enterprise architecture according to the needs and the defined goals
- Not all the countries have the same goals and needs
- Different national governance models exists
- The defined national architectures are diverse and not easily comparable
- Different definition levels of NEA exists
- The national strategy documents are not always fully available in public or there not available at all
- Some countries publish their national strategies only in the national language
- The NEA state and progress in a lot of countries is constantly changing

A detailed and sophisticated NEA evaluation methodology has not been published yet. In this paper we introduce a new methodology that presents the most advanced indicators.

One of the most detailed studies on this subject is conducted by Christiansen and Gotze [4], but the results in their report are summarized for all the countries.

The study of Leist and Zellner [5] evaluates part of the well-known and often used frameworks, including few national architecture frameworks. This methodology includes only five constitutive elements and each of the elements is evaluated with only three levels of sophistication: fully accomplished, partly accomplished or not accomplished.

Another detailed study is conducted by the FEAR-project [2] to support the development of the Finnish NEA. This methodology includes evaluation framework that consists of seven viewpoints, and for each of them description of that part of the NEA is filled. That lack of this framework is that the results are only descriptions, so they are not easily comparable.

4 Results

Using the proposed framework we have conducted an evaluation of the values of few countries: Austria, Belgium, Canada, Denmark, Estonia, Finland, Germany, Macedonia, Netherland, New Zealand, Norway, Sweden, Switzerland, USA and United Kingdom. The results for each of the indicator sets are shown in Figure 2. The overall results are shown in Figure 3.

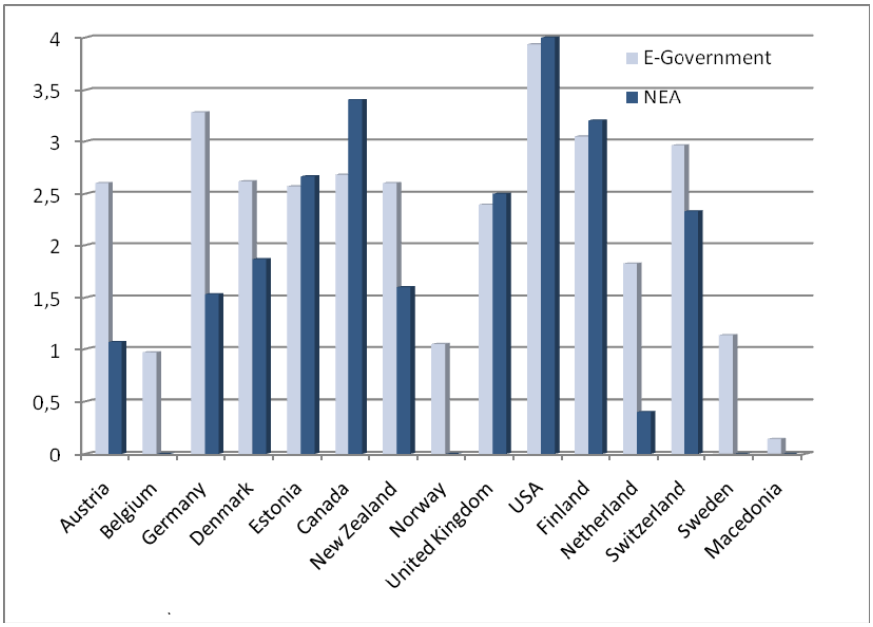


Fig. 2. Results for the E-government initiatives indicators and for the NEA indicators

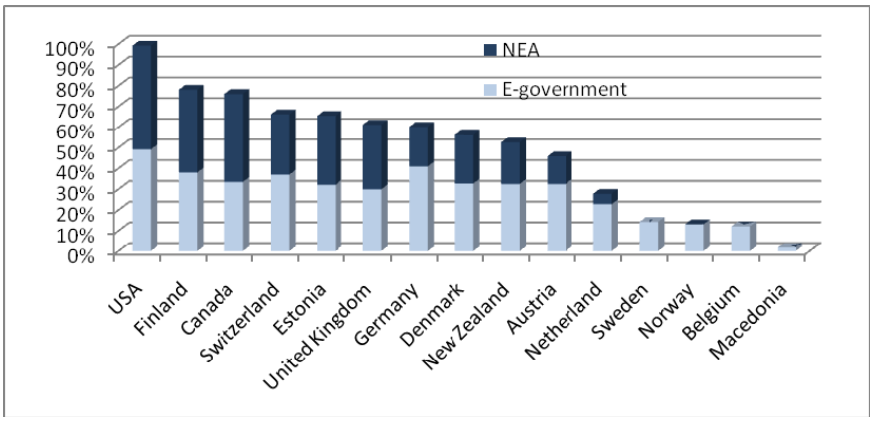


Fig. 3. Overall results

We can conclude that according to the proposed evaluating framework United States of America is the leading country in this area. The lowest values are for the Republic of Macedonia at the moment, although there is slight moving forward. In [12] we show that even if Republic of Macedonia accepts to apply the proposed national architectural framework, than the value of the country will rise for at least

50% of the current value, but will not reach highest level due to fulfillment of other indicators.

5 Conclusion

The proposed evaluating framework is unique and represents a hybrid and detailed evaluation model. As a starting point we used Information Society indicators [7]. Some of the indicators are defined according to the experience of the previous mentioned studies and for each of them we have defined several levels of sophistication. New indicators are additionally defined and also for each of them we have defined several levels of sophistication. The results gained by the proposed framework are comparable.

We believe that this new methodology will help all countries willing to introduce and update their NEA initiative to catch up with most advanced trends expressed by evaluation methodology and indicators introduced in this paper.

References

1. OECD (2005a): OECD Peer Review of E-Government in Denmark, Organization for Economic Cooperation and Development. (September 2005)
2. Liimatainen, K., Hoffmann, M., Heikkilä, J.: Overview of Enterprise Architecture work in 15 countries.Finnish Enterprise Architecture Research Project. Ministry of Finance (October 2007)
3. Accenture: Leadership in Customer Service – Building the Trust (2006)
4. Christiansen, P. E., Gotze, J.: International Enterprise Architecture survey - Trends in governmental Enterprise Architecture on a national level. (2006)
5. Leist, S., Zellner, G.: Evaluation of current architecture frameworks, Proceedings of the 2006 ACM symposium on Applied computing (2006)
6. Janssen, M., Hjort-Madsen, K.: Analyzing Enterprise Architecture in National Governments: The cases of Denmark and the Netherlands. In Proceedings of the 40th Hawaii International Conference on System Sciences, HICSS 40, Jan 3-6, Waikoloa, Big Island, Hawaii. (2007)
7. United Nations: Information Society Indicators. United Nations, New York, (2005)
8. Lallana, E. C.: e-Government Interoperability, United Nations Development Programme e-Primers for the Information Economy, Society and Polity. (2008)
9. Christiansen, P. E., Gotze, J.: Trends in Governmental Erterprise Architecture: Reviewing National EA Programs – Part 1. Journal of Enterprise Architecture, vol. 3, no. 1, 8-18. (2007)
10. Paszkowski, S., Mortensen, R.: Enterprise Architecture and Interoperability Survey, Survey Results - Gathered Evaluation Results at Governmental Level (July 2008)
11. United Nations: World Public Sector Report 2003, E-Government at the Crossroads. (2003)
12. Kostoska, M., Gusev, M.: Macedonian National Enterprise Architecture Framework. CIIT 2010, Bitola, Macedonia (2010)

Implementing Strong Authentication with OTP: Integrated System

Pance Ribarski, Ljupcho Antovski

University Ss. Cyril and Methodius,
Arhimedova bb, 1000 Skopje, Macedonia
{pance,anto}@ii.edu.mk

Abstract. Due to the arising problems with using static passwords, there is a strong need of implementing more secure protocols for authentication. The One-Time Password protocol is implementation for two-factor authentication; the two factors being something you own (a token) and something you know (PIN). This paper presents an open algorithm for OTP and implements a client-server system for secure OTP authentication. The implemented system is full-blown authentication system for multi-site secure authentication ready to be implemented in a real-case scenario.

Keywords : security, password, authentication, multi-factor, two-factor, one-time password, internet

1 Introduction

Problems with authentication online appear more frequently. The security level of nowadays static passwords does not overcome the threats rising in stealing those passwords [15]. There are numerous examples of stolen passwords, phishing scams and many other techniques that allow bad guys to take our passwords [14]. This disrupts the authentication process – process of confirming the identity. This paper introduces a platform for secure authentication employing two-factor algorithm. In chapter 2 we identify the problem. In Chapter 3 we discuss the changes organizations need to embrace in order to bring security to a satisfactory level. These changes have to be enforced in almost every place where classic password authentication is used. In Chapter 4 we review the One-Time Password paradigm and the general ways of implementation. In Chapter 5 we explain the developed working OTP algorithm – algorithm that is under the auspices by the open community of OATH [9]. Chapter 6 underlines our contribution to this OTP area – providing a whole system for secure authentication. Our solution consists of server side – that the authentication organizations should implement, and client side – that consists of software tokens installable on mobile phones.

2 Problem Identification

In computer security, authentication is usually an automated process of verifying the identity of someone or something, such as a computer or application. [19, 1]. Simply, the authentication process is the basis of every multiparty communication in which the parties need to introduce themselves. The introducing part is very difficult to conduct in electronic manner – the other side can never be sure if the introducing identity is the true one.

The classic approach is using a pre-shared secret, that acts as a challenge when the introducing party needs to prove its identity. If the introducing party answers the correct secret, then the identity is proven. This method is copied from times far before the computer age, when a person would have to recite a word or a phrase in order to confirm the right identity [20]. In the computer world this is called a static password.

Static passwords have are flaw in secure manner by design – they don't change over time. Even if they change, the new password must be distributed to the user for further use. Decreasing the life-time of a password has led to users writing their passwords down because they simply can't remember such short-timed passwords. The static nature of the passwords also leads to man-in-the-middle attacks where someone intercepts the password and then uses it to authenticate as the person which stole the password from. The man-in-the-middle scenario and some other password stealing scenarios can be performed variously [21]:

- network administrators or sniffers can sniff for password roaming around [22];
- trojan horses and keyboard listeners can steal passwords as they are typed [23];
- phishing attacks can lead the user to giving away their password [24];
- using social methods to extort a password from a person [25];
- brute-force lookup for a passwords – modern technology spread this way even on non trivial passwords [26].

3 Multi-Factors Authentication

The answer to these problems obviously is the authentication method where the user does not disclose static information. We want to authenticate answering a challenge that always changes. This way, even if someone is looking over our authentication process, it cannot use that information to authenticate as us later on. Even more, we must consider an authentication that requires multi-factors. These required factors ensure that if someone is going to authenticate as us, he must collect all of the factors. The factors that we can choose from are loosely categorized as [16]:

- biometric data – something you are; this factor can be fingerprint, retina scan, voice, or some other bio-information
- static password – something you know; discussed in Chapter 2

- hardware or software – something you have; this is supposed to be unique piece of data that cannot be copied

This research is trying to cover the level of security in the area where we already authenticate with static passwords. It is common sense that if we add more factors we get better security. This is wrong with the authentication process we try to employ – we try to better the plain password authentication and minimize the complexity of the process. This area will be completely satisfied with a two-factor authentication – the method we are proposing. From the list of categories from which we can choose a factor, we choose a static password and a hardware or software. This concludes to a two-factor system in which if we want to authenticate we need to know some password (further noted as PIN) and a device that will generate a random code every next authentication. This solves the problem with static passwords, and generally increases the level of security with increasing the number of factors to two. This kind of authentication is named as One-Time Password, or shortly OTP [27].

4 OTP implementations

There are various implementations of OTP systems [17]. All of them are with the same cause – generate a token that we pass to the authentication server. The authentication server generates one on their own – and checks if the two tokens are equal. If they are equal, than the authentication process is successful. The creation of these tokens is done in various ways:

- create a token using one-way function [18] taking a parameter from the previous token creation; this initial token is something that both the client and server know and is random by nature; randomness is provided from Java Crypto Class KeyGenerator
- create a token using one-way function taking a parameter from an accurate clock; the clock in the client and server must be synchronized
- create a token using one-way function taking a parameter from a secret key and a counter; the secret key is pre-shared with the client and server in an secured manner

There are implementations of OTP that are open – allowing users to see the algorithms and allowing the community to verify that it is secure, or if it is not secure to improve it [8, 10, 11, 12]. But most of them are commercial, meaning that the implementation is hidden and the usage of the system requires registration fees [2, 3, 4, 5, 6, 7].

The implemented systems are meant to be used as a way of two-factor authentication in various systems. They are implemented in different platforms, like C, Java, .NET. The user can generate the token in different ways: using smart cards [4, 6], USB tokens [5], or use the mobile phone as token generator [2, 3, 7]. Some systems are implemented to be used in the UNIX environment [10, 11, 12], thus securing the classical username-password authentication system in UNIX.

The Initiative for Open Authentication [9] is an organization founded by the industry companies trying to unify the authentication process. Their aim is to "provide reference architecture for universal strong authentication across all users and all devices over all networks" [29]. Their joint effort has produced many documents and protocols regarding strong authentication. Among them is the HOTP [8] algorithm which is counter-based one-time password scheme. We are using this scheme in the implementation of our OTP authentication system.

5 Implementation Algorithm

The algorithm proposed from the Initiative for Open Authentication generates numeric OTP values with length from 1 to 8 [8]. The greater the numbers, the stronger the authentication process. This algorithm requires a *symmetric key* (denoted *K*), known only to the authentication server and the client. The transportation of the key to the client should be done over a secure channel. The key transportation is done only one time – at the client token creation. This will be repeated only if the client token is lost, or malfunctioning. In this case it's better to create a new key, as if it is a new client token. Another key part of this algorithm is the parameter that changes over time – exactly every time a new password is generated. This parameter is called a *counter* (denoted *C*). At first the counter has a value of zero at both the client and server. When the client token generates password, the counter is increased by one. If the server is given a generated password for authentication, it generates a password on its own with the current counter for a given client, a check for the equality. There is a possibility that a client will generate a password and not use it. This way, the client counter will increase, but the server counter will not. If the client sends a password to a server, it will not match with the server's. In this scenario, the server has to compensate with the counter, increase the counter couple of times trying to reach the client's counter. The number of times the server increases the counter trying to compensate is denoted as *S – resynchronization parameter*. This value should be kept at minimum to avoid brute force attacks. Another parameter that is handy in avoiding brute force attacks or denial-of-service attacks is the number of wrong *authentication attempts - V*. This parameter will increase every time the server fails to authenticate a client password using the resynchronization parameter S. Like S, the value of V should be kept at the possible minimum to avoid guessing the password, and yet to give the user a chance to wrongly use the system.

The password generation is divided in 3 steps. The first step is to generate Hash-based Message Authentication Code – HMAC [13]. This operation creates a message digest of some text using a secret key (the mentioned key K). Any iterative cryptographic hash function can be used. Our system is currently using SHA1 as cryptographic function, but other functions are tested as a work in progress. The text that is HMAC'ed is the counter C. The result from this step is 20-byte string.

The second step is extracting a 4-byte string from the HMAC value. This is optionally done in two ways – using static offset, or using dynamic offset. The static offset is always extracting 4 consecutive bytes from the 20 bytes HMAC value. Thus the static offset parameter can be 0-16. The dynamic offset is calculated from the last

4 bits from the HMAC value. The 4-byte extracted value the base for generating a password.

The third step is converting the 4-byte string into a decimal number. This decimal number is then shortened to a desired length using modulus with 10^D where *D* is the *password length*. The result from this step is the generated OTP password that we send to the server for authentication.

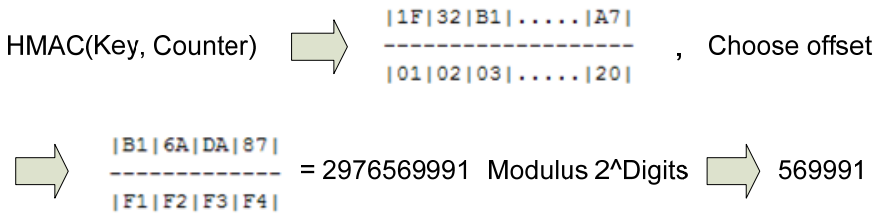


Fig. 1. Algorithm for creating OTP token

The security of this algorithm has been examined thoroughly [8]. The analysis shows that the best possible attack against the algorithm is a brute force attack. Even if attacker has access to consequential successful password values, building of a function based on these values does not have any more advantage over random guessing the values. This concludes that the brute force attack is the only option. Using brute force, the probability of guessing a password is:

$$p = \frac{S * V}{10^D}$$

In a system where resynchronization parameter S is 4, the number of wrong authentication attempts V is 5, and the length of OTP password is 7, the probability of guessing a password is only 0,0000035. This probability is in the acceptable boundaries for a secure system because a brute force attack with this probability rate is technically impossible.

6 The System

The OTP system itself that we are developing is the main contribution in the OTP area. This system is a unified platform consisting of server plugins and client plugins. Server plugins are the tools used at the authentication server. They receive requests for authentication and respond accordingly. The client plugins are tools to generate OTP passwords at client side, they are meant to be user-friendly, not expensive and generally usable. Systems like this already exist, but mainly as commercial products [2, 3, 4, 5, 6, 7]. These commercial solutions offer OTP authentication, but are closed in algorithm manner and users don't have insight in the inner working of the system.

This system allows registration of sites that will use OTP authentication. These sites have the abilities to manage users on their own. The main idea is that this system will work on its own, while all the administrative work will be done from the registered sites that are going to use it.

The server side consists of:

- database to keep all the information
- web site to manage all the administrative tasks
- web service for administrative tasks done by the registered sites and for delegating keys to clients

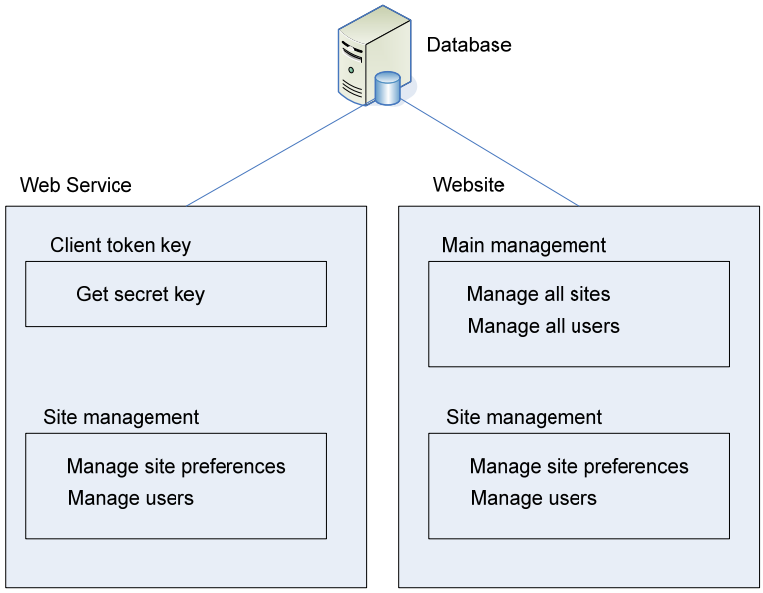


Fig. 2. Architecture overview of OTP system

The database keeps information about the registered sites and about the registered users. The main information for each user are its secret key and counter. These data are needed for the server to generate OTP password for requested authentication from a user. The web site is used for web management and administrative work with the whole system. It is mainly divided into system administrative part and administrative part for particular site. The system administrative part has permission to manage all the sites and all the users. Further, it can create new sites, and add new users to those sites. The administrative part for particular site can manage users only for their site. They can add new users, delete users, recreate a key for a lost device and more. The web service is part of the server tools. The main role is to distribute the secret key to the client token. This can happen only once, the first time a token in initialized. Further reading of the client key is forbidden. Exclusion to this rule is the case when a client token is lost and it needs to be recreated. But even then, a new key is created and this new key is only once distributed. The web service has also an administrative

role. Through the web service sites can manage their user databases. This is an acceptable way of interconnecting legacy systems at the sites' side with the OTP authentication system – programmers can use these web services and connect their own software infrastructure with the authentication system.

The client side is meant to be easy for the client, not expensive but still secure. The token hardware is the "something you have" factor from the two-factor authentication protocol. The only logical conclusion is to use the client's mobile device to act as OTP token generator [30]. Today's mobile devices are capable of running applications with limited resources. There are many platforms on which the mobile phones work: windows mobile classic, windows mobile 7, apple iOS, blackberry, symbian, android, java mobile and others. There is a great probability that any user has mobile phone built on some of these platforms. The system explained here is set to create software tokens – applications for every major platform. In that way the OTP will become easy and very cheap to implement in every organization where secure authentication is required. The token application is designed to operate with more sites – meaning that it holds keys from all the sites and the client is able to generate OTP passwords for many places with only one application. This contributes with the OATH community in a way that these mobile clients can be used with any authentication servers because of the open algorithms used. The clients can generate passwords for sites administered by the system explained here, and can generate passwords for any place that complies with the algorithm explained before.

The process of authentication is the following:

1. client enters authentication site
2. server sends OTP password
3. client checks if the password is correct; sends OTP password
4. server check if the password is correct; authenticates user

This double checking is a double OTP protocol. It is not usual with the OTP authentication protocols but strengthens the security in the process. This way we can avoid the biggest flaw in web authentication – phishing [24]. If we are actually authorizing at a site that pretends to be the site we want to authenticate on, this fake site will not generate a correct OTP password – it will not be equal with the password generated with our token generated password.

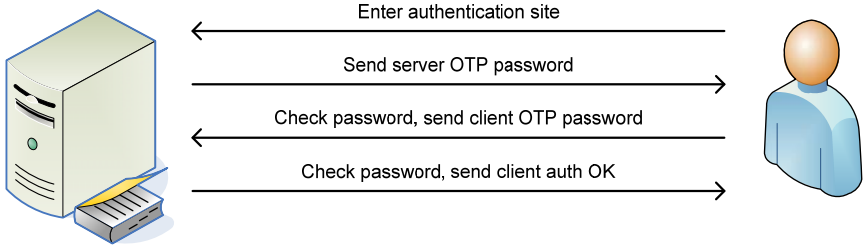


Fig. 3. Authentication protocol between client and server

7 Conclusion

The presented research is a work in progress towards an affordable an easy-to-use system for secure and strong authentication. The current classic authentication method – static password is prone to password stealing. The multi-factor authentication protocols are stepping in, One-Time Passwords are leading the way as currently popular protocol for strong authentication. The presented algorithm is easy to implement and its openness is giving the opportunity to have many developers embracing it in authentication systems. Our system is a solution both for the server side and client side. The server side is consisted of website and web service to serve to administrators and site managers. The client side is consisted of applications that can be installed on mobile devices which will act as hardware token generators. The whole system satisfies and prevents stolen-password, man-in-the-middle, denial-of-service and general phishing attacks. This system is a proof of concept that the transition to a secure authentication can be easily done, and that it can be user-friendly. The technological advantages of mobile devices give the opportunity of slicing costs for extra hardware tokens.

8 References

1. Ribarski, P., Antovski, Lj.: Introducing Strong Authentication for E-Government Services in Macedonia, 2009
2. FiveBarGate, http://www.fivebargate.net
3. FireID, http://www.fireid.com
4. Solid Pass, http://www.solidpass.com
5. Yubico, http://www.yubico.com
6. NagraID Security, http://www.nidsecurity.com
7. OTPXS, http://otpxs.com
8. HOTP, http://www.ietf.org/rfc/rfc4226.txt
9. Initiative for open authentication, http://www.openauthentication.org
10. Haller, N. M., The S/KEY One-Time Password System
11. OPIE – One-Time Password In Everything, http://wiki.linuxquestions.org/wiki/Opie
12. OTPW, http://www.cl.cam.ac.uk/~mgk25/otpw.html
13. Hash-based Message Authentication Code, http://tools.ietf.org/html/rfc2104
14. Dhamija, R., Tygar, J.D., Hearst, M.: Why Phishing Works
15. Fragkos, G., Xynos, K., Blyth, A.: The use of computers idle-time and parallel processing over a network to perform password threat assessment
16. Bhargav-Spantzel, A., Squicciarini, A., Bertino, E.: Privacy Preserving Multi-Factor Authentication with Biometrics
17. Takasuke, T., Akihiro, S.: One-Time Password Authentication Methods
18. Volpano, D.: Secure Introduction of One-Way Functions
19. VeriSign Identity and Authentication Glossary, http://www.verisign.com/authentication/information-center/authentication-glossary/index.html
20. The Roman Military System, http://ancienthistory.about.com/library/bl/bl_text_polybius6.htm
21. Sans Institute, SSL Man-In-The-Middle Attacks
22. Del Armstrog, Password Sniffing

23.Sans Institute, First Step Data Capture – Key Stroke Loggers
24.WordSpy Phishing, http://www.wordspy.com/words/phishing.asp
25. What is social engineering, http://www.microsoft.com/protect/terms/socialengineering.aspx
26.WASC Threat Classification – Brute Force, http://projects.webappsec.org/Brute-Force
27.One-Time Passwords Specifications, http://www.rsa.com/rsalabs/node.asp?id=2816
28. Java Crypto KeyGenerator, http://download.oracle.com/docs/cd/E17476_01/javase/1.4.2/docs/api/javax/crypto/KeyGenerator.html
29.OATH Features and Benefits, http://www.openauthentication.org/features
30.BearingPoint, Multifactor Authentication: A Brief Selection Guide

SOA Approach in Prototype of Intelligence Information System

Jugoslav Ackoski[1], Vladimir Trajkovic[2] and Metodija Dojcinovski[3]

[1] GS of Army/MOD, str. Orce Nikolov bb, 1000 Skopje, Macedonia
[2] Faculty of Electrical Engineering and Information Technologies, str. Rugjer Boshkovik bb, PO Box 574 1000 Skopje, Macedonia
[3] Military Acedemy „General Mihailo Apostolski“, str. Vasko Karangelevski bb, 1000 Skopje, Macedonia

Abstract: Intelligence, as a service has a great significance for the country. An information system for support of intelligence activities is very often in everyday use and from that use comes great influence in the decision making process. Usage of the modern information technology in big way contributes for improvement of the process (activities) which are supporting intelligence cycles (planning, collecting, analyzing and dissemination). Although there is constant improvement as a result of the progress in the area of information technology, significant difference in the quality of work in the field of intelligence has not taken place in the last ten years.

Implementation of Service Oriented Architecture – SOA, i.e. the usage of SOA, is providing possibilities for making new opportunities in the form of expanded solutions for designing intelligence information systems, regarding the more efficient management of information, as well as their use by the end users for whom hum they are intended. In order to keep up with the pace with modern development, short, medium and long term planning is needed for development of information systems for supporting intelligence, in relation to the of IT development.

This paper presents an idea for SOA approach in prototype of Intelligence Information System. Prototype of IIS is a solution which should offer better coordination and Intelligence effectiveness. It is a foundation for establishing integrated system for Intelligence.

SOA approach in information systems is a logical solution, not only for a temporary and short term usage but it is a perspective solution for general strategy in companies and governmental institutions.

Keywords: Service Oriented Architecture, Intelligence IT solutions, process optimization.

1 Introduction

Every modern intelligence system is based on some type of computer information system. Usage of high technology, especially information – telecommunication

technology (ICT), is giving more efficient execution of the intelligence function in terms of collecting, planning, analyzing, and data dissemination process.

Traditional intelligence cycle is a process with precisely defined steps in which separate departments are focused on their part of the process and the remaining aspects of the intelligence cycle are left to the other departments. This approach is treated as an inadequate against netwar (conflict with complex, dynamic, nonlinear nets, which up till now are the biggest challenges for the intelligence community in the area of information age). As a response, presently, different intelligent agencies have established teams for monitoring and response to events on different locations, by using modern IT solutions for better coordination.

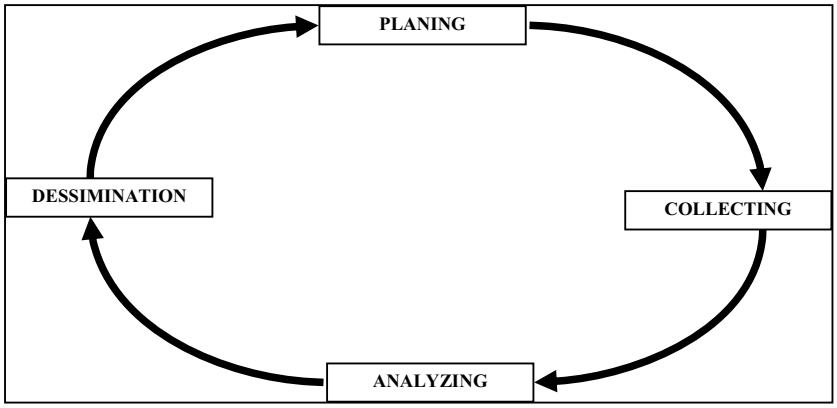


Fig. 1. Intelligence cycle

On the other hand, target-centric model of intelligence, especially the analysis process is not new, but it is not formally accepted by the intelligence population. Individuals and small teams are making ad-hoc temporary information database for improving the process information analysis. However, the information databases are designed as temporary inadequate solution for the direct needs of the information analysts. These information databases are mainly made up of piles of paper and are not understandable for outsiders.

After the changing of the traditional hierarchical model in the target-centric model of intelligence, intelligence analysts as a part of the intelligence community are facing the need of new IT solutions, with one goal – achieving better intelligence products.

SOA, one of the possible solutions, as a particular component of a wider IIS, represents a platform for creating a service for exchanging information in IIS.

2 SOA in Information Systems

SOA is a technological approach towards designing information systems, where the primary aim is using IT advances in business process in a way that would produce bigger efficiency, in order to create synergic and symbiotic relations. Furthermore, there are a lot of doubts whether the SOA approach is going to solve many of the technical problems and NEC advantages. That is only partly correct, because the SOA solutions are IT products themselves and are to a great extent dependent on many other factors which are part of the operating process, for instance human or organization factors as components of the Defense system of one state. SOA approach is abstract concept "service", where the service is technology – independent structure which is simplifying the process „loose coupling“ and provides 'platform for making architectural(modular, opened) components.

On business level, business components exchange large-grained business services (e.g. Common Operating Picture (COP)). On technological level fine-grained technological services are exchanged (e.g. data store). SOA approach is a bridge over which mapping of large-grained business services is made in fine-grained technical services.

Up to this moment, MO are not in a position to achieve flexibility for combining, positioning and configuration of software components in an appropriate way or to create some particular new components. SOA approach is offering architectural structure that is making level of integration of the services as a sure sign of satisfying software and business demands, with only one goal - getting products of higher quality. The abovementioned facts justify investment in the use of the SOA approach, although the real validation is the improvement of the working process, synergy changes and information exploitation, which is not the case with the previously used information systems.

In the early phase of the SAO approach implementation it is often very hard to determine the characteristics of the integrated systems development which are to be used in military purposes. Conclusion can be made that, although we come to a point where further integration and interoperating using the existing systems provides additional advantage. Even in the relatively high level military services where the implementation of the SOA approach or Federated Enterprise Service Bus (ESB) is big and where the value of the group applications is again shown as a group of well defined services (business components), all have achieved optimization. While this is clearly desired and generally very useful achievement, the real challenge, or opportunity, the MoD now is pointing in a direction of using much bigger quantity of useful information together with flexibility and degree of fast re-engineering or creating components for re-using which have characteristics for work in service oriented environment.

3 Use of NEC in the decision making process

The purpose of Networked Enabled Capability (NEC) is to support decision making over “on time found and use of information in intelligence” [1]. Influence of NEC is in area of Defense with aim of getting higher military effectiveness.

The concept of the (NEC) in support of information trade is different compared to the solution for Network - Centric environment, where the net is the base on which military capability is made. The use of the concept for NEC is emphasizing the importance of the decision making process by the authorities in the military area with information that is needed for the right decision making.

From the IT perspective, the implementation of NEC environment presents a challenge. The following may be given as an example, where there is need of distribution of the information received from a big number of sensors (UAV, aircraft and other elements from a fighting schedule) and they must be exchanged between different HQ in joint combined operations.

As an addition to the challenge for information distribution over the NEC is the need of fast response to threat.

3.1 IIS and NEC

In conditions of change of the threats IIS(Intelligence Information System) it must be adapting, with one goal of rational usage of the opportunities which NEC is offering. IIS is a information systems for intelligence, in which are contain reports, overviews, estimations, analyses and other intelligence products in function of decision making process.

That means that NEC is very important solution for IIS, because is giving possibility of connecting between the elements of intelligence which are achiving interaction between them self in the operating environment [4].

3.2 Supporting IIS

In an operating aspect, effective exploitation of information for support decision making process in a military environment is a key to the military success.

As a military threats are changing, and new capabilities are developing (e.g. sensors) it means that is need change in the practices, system IIS is need to develop in higher stadium.

As a response to the new operative requirements from the clients is use Service Oriented Architecture (SOA) for making IIS.

4 IIS prototype

4.1 Type of users

Type of users that exists in IIS can be analyzed in two aspects:

- Aspect of IIS system functionalities;
- Aspect of IIS users functionalities.

Aspect of IIS systems functionalities defines standard users from system aspect. These are following types of users:

- Administrator of communication infrastructure;
- Administrator of applicative solution (definition of web services, integration in informational systems of the institutions);
- Secure administrator (define of polices).

Those types of users are standard types of users which are existing in any more complex information system. From functional aspect IIS have additional four types of users:

- Requestor of service
- Approval of the services
- Companies Компании
- Intelligences (IMINT, SIGINT, OSINT, MASINT)

Requestor of service are any institution (Center for managing with crises, MOI, Agency for intelligence and other institution) how for needs of their work has need to get information, or to give information (as a notification).

Supplier of services is a user who should supply information (in which can ask for proper authorization according to own secure procedures or according with some external norms (where it should be achieved as a request a favor from IIS)

Companies are external users how can be directly referring to some institution which will make their request (e.g. department or section of Government), but they can ask also some public information which should be approved and secure by IIS.

Intelligence is based on four intelligence disciplines: IMINT, SIGINT, MASINT and OSINT.

Also disciplines are dividing on subdisciplines with different specification in part of the work and can organizational part of many institutions or unites in Defenses forces of the state (Agency for intelligence (AI), MOI, MOD, Ministry of Health, MFA and other subjects), but there role and task is to put in information (information, assessments, analyses, reports and etc.), there verification (which can be subject of support of IIS) and to get notification or (e.g. For political-secure situation in one state in relation with secure of the investment).

On picture 2 are shown those users types and the phases of their general users functions.

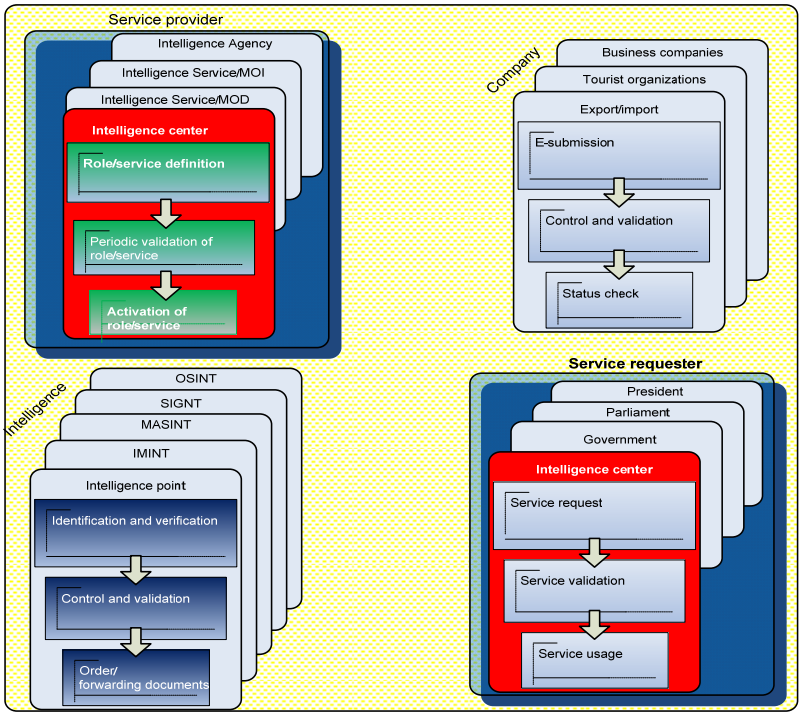


Fig. 2. User of the system for support the intelligence

4.2 Users functionalities

Users functionalities of the system of IIS can be treated from few aspects according to divided users. This chapter will contain definition of the functionalities of final users of IIS. As it was already explained, there are four types of final users of IIS. Those are:

- Requestor of service
- Approval of the services
- Companies
- Intelligences (IMINT, SIGINT, OSINT, MASINT)

Analyzing the process which should exist in the institution included in IIS, it may came to the solution that the final users can use same target for users scenario, meaning always to work in same general procedure (under procedures). That process is containing three phases. Those phases are:

- Phase of evidential,
- Phase of verification,
- Phase of notification.

Those phases exist in every process connected with IIS, but not always are with support from IIS. Sometimes are supported from intern informational systems of the institution included in process of IIS. That is additional reason for need of integration of system for support of workflow on mediation level in every institution included in the process of intelligence.

4.3 Services oriented architecture of IIS

Services oriented architecture of IIS should have central system for support intelligence from logic aspect, while his physical position is distributed.

Services oriented architecture of IIS will be described over explaining logical architecture of IIS and layer model architecture of IIS.

4.3.1 Logical architect of IIS

On picture 5n is shown proposed logical architect of the system for support of IIS. She is using model of levels and contains following levels:

- Level of users;
- Access level;
- Level of process;
- Level of services;
- Level of services providers.

The level of service users is giving approach over using appropriate presentational logic from IIS for all types of functional and system users explained in chapter **Type of users**. From organizational aspect those types user are part of: center of intelligence, Crises Management Centre, embassies of MFA, and other government institutes and agencies, but it can be also other users represented by their companies.

Those users according to secure polices have different way of approach to IIS. The way of the approach is defining in the level of the way of approach. It can be over public internet net (e-mails, approach to web page of IIS), private intranet net (over web pages,e-mail and integration with informational systems that exists in the institutions), or over some private net (over widing the system of MOD, MOI and other with possibility of using the services of IIS, and integration in parts of the information system who have special importance for the work of the institutions). Level of the approach, in accordance with defined polices of types of users is determinating the way of approach of every user.

From the way of approach depend unit of all process from the domain of the intelligence to hum the user will have approach. Possibly process are: planning, collecting, analyses and dissemination. Those processes are uniting all hierarchy of possibly scenario of usage which depend from legal/law and inter procedures of the institutions included in the process of the intelligences.

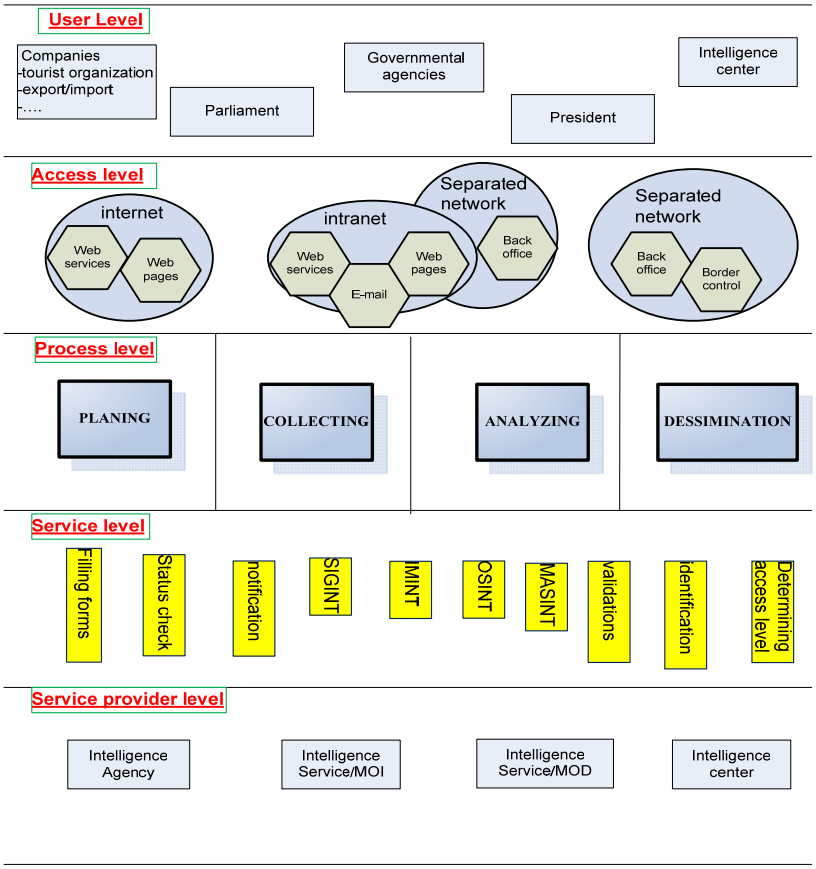


Fig. 3. Level of logical architect of system for support of IIS

Services which are using are defining in the level of service. They are available for integration in the process which are used by the users in accordance with secure polices and availably points they have right to be used. Those services can be: services of control (if the company or document exists as a point in some base of information), services for collecting information (IMINT, SIGNT, OSINT и MASINT), determination (and change) of right of approach according to some regulation (which understands automatic or manual process of improvement), identification (if the given person or document is appropriate with what is shown on base of the rule (comparison of information, statistical analysis)), following (of goods with help of GPS device), notification (for attendance in mutual control, for statistical evidention, for updating geographical informational on base of signal from GPS device), fulfilling forms (for getting information, evaluation,

reports and similar), checking the status of some requests (if the request is recorded, is evaluating, or it is approved).

All services are getting from appropriate service providers through Information systems of government institutions, agencies which are included in Intelligence cycle. It is also possible another Information systems to be service providers for interinstitutional governance. Service providers among system support for workflow processes define web services which are exploited from the users with appropriate security level to service registers.

On that manner is provided flexible architecture which affords to the users of IIS continuous updating and spreading services. In addition, information access is in relation with security polices, but most important is that IIS provides unite approach to the information through set of access points, which is characteristic for centralized systems. Because of that suggested architecture for IIS is centralized systems

4.3.2 IIS architecture

On picture 6 is shown general architecture of IIS prototype. As result of system complexity solution is realized as a layer model of architecture.

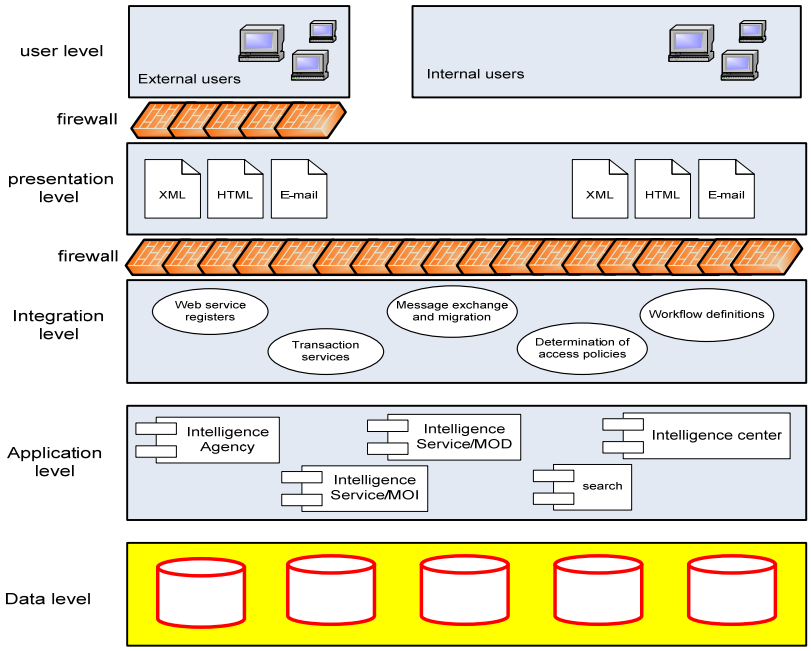


Fig. 4. Prototype IIS Architecture

On the lowest level, IIS prototype has distributed system which consists of heterogeneous database. In this case, most important database for IIS is database which holds data for users who use it. Intelligence center has responsibility for this database.

Access to a separate database will be made with application logic of module which are part of internal information systems on government institution. This application should provide interfaces to the integration logic level. Integration level is key for IIS prototype. That level should provide services through workflow which will be connected with modules of internal information systems and their transformation into web services. As a result of provided web services, integration level should publicized them into appropriate web services registers depending of security level. Also, this level govern security polices and polices for exchanging and adopting messages from different sources, in case of usage in comparable format. Finally this level is taking care for governance of the services offered by IIS in a way of transactions if it is needed. With one sentences, this level is providing the functionality of the services in IIS.

That service should be available in different categories of users. For the purposes of protecting IIS, behind this level should be install firewall after which follows level of presentational logic. This level can be made in a form of portal which can offer: list of web services over approach to service registries, integration of web services with e-mails or directly as far procedure call of the applications (RPC) in a standard format (XML) , but also as a ordinary HTML text for separated union of services – users.

External cooperatives to the system are determent with another additional secure wall-firewall, this way is accomplish maximum protection from unwonted breaking in the system.

5 Conclusion

In the aim of increasing usage of SOA and according offered opportunities, key challenges should be shown. Generally speaking secure challenge is always a problem. Beside that offered standards for flexibility and extensively are making risk of unsecure design much higher.

Flexible usage and accepting of the services and the opportunity for changing the numerous systems which are not in accordance with the levels of the accredit secure is showing the fact of needed careful approach from authorities side for secure accreditation. This type of architecture should be develop in the aim of increasing ability on SOA approach.

Benefits of operational efficiency which might be offer from SOA in a future they will be challenge about solution of security issue.

SOA approach is modern way for increasing IIS capabilities on operational level. Architecture model can help understanding business process and indentifying services which are important for establishing appropriate level of granularity.

Definition of solution according SOA Governance represents a need in the aim of using advantages which offers SOA approach. SOA Governance should set up stakeholders and their interactions, and define lifecycle of services.

Future development will be in a direction of implementing SOA approach in all IIS domain. This approach will be use to afford integration and interoperability with other systems.

References

1. "JSP777 Network Enabled Capability", Edition 1, Ministry of Defence.
2. "Exploring New Command and Control Concepts and Capabilities Final Report", NATO SAS-050, January 2006.
3. "Understanding Command and Control", Alberts D.S, and Hayes R.E., CCRP Publication Series, 2006.
4. "Power to the Edge, Command and Control in the Information Age", Alberts D.S., and Hayes R.E., CCRP Publication Series, 2005.
5. Draft UK Defence SOA Policy, JSP 602-1001, Draft Issue 2.
6. Allied Rapid Reaction Corps, http://www.arrc.nato.int/, accessed on 17 April 2008.
7. IEEE Recommended Practice for Architectural Description of Software-Intensive Systems, IEEE Std 1471-2000, September 2000.
8. Ministry of Defence Architecture Framework (MODAF), version 1.2, http://www.modaf.org accessed on 15 August 2008.
9. NATO Architecture Framework, version 3, AC/322- D(2007)0048.
10. Modeling SOA, IBM developerWorks, http://www.ibm.com/developerworks/rational/library/07/1002_amsden/index.html?S_TACT= 105AGX15&S_CMP=LP, accessed on 21 April 2008.
11. OWL-S v1.2 Pre-release, http://www.ai.sri.com/daml/services/owl-s/1.2/, accessed on 19 August 2008.
12. Protégé ontology editor tool, http://protege.stanford.edu/, accessed on 19 August 2008.
13. SPARQL Query Language for RDF, W3C recommendation dated 15 January 2008.
14. "Get Serious About SOA Governance", BEA, September 2007.
15. Multilateral Interoperability Programme, http://www.mipsite.org/010_Public_Home_News.htm, accessed on 21 April 2008.
16. OASIS Web Services Business Process Execution Language, http://www.oasisopen.org committees/tc_home.php?wg_abbrev=wsbpel, accessed on 21 April 2008.
17. Mule, http://mule.mulesource.org/display/MULE/Home, accessed on 19 August 2008.
18. World Wide Web Consortium (W3C), http://www.w3.org/ accessed on 21 April 2008.
19. soapUI web services testing tool, http://www.soapui.org/, accessed on 19 August 2008.

An Approach for Extraction and Visualization of Scientific Metadata

Stevica Cvetković[1], Miloš Stojanović[2], and Milena Stanković[1]

[1]Faculty of Electronic Engineering, Aleksandra Medvedeva 14, 18000 Niš, Serbia
{stevica.cvetkovic,milena.stankovic}@elfak.ni.ac.rs
[2] High Technical School, Aleksandra Medvedeva 20, 18000 Niš, Serbia
milos.stojanovic@vtsnis.edu.rs

Abstract. In this paper we described an approach for automatic, template-based citation metadata extraction from scientific literature, as well as their visualization. The extraction approach assumes PDF format of file and IEEE reference writing standard. It is based on formally defined templates in form of regular expressions which are utilized to implement finite state machine for metadata extraction. After relations between references are extracted and stored in adequate data structures, their visualization using graphs and treemaps is performed. Graphs and treemaps proved to be very efficient and compact techniques for visualization of citation information, particularly effective to emphasize the most cited papers and authors in specific field of science. Finally, we demonstrated satisfied test results and discussed future plans for improvement of the approach.

Keywords: bibliographies, finite state machines, information retrieval, information visualization

1 Introduction

Information exchange between academic world developed quickly with appearance of scientific paper search engines on the Web. Most popular between them are *CiteSee*r [1], *Cora* [2] and nowadays *GoogleSchoolar* [3]. They enable search and analysis of scientific literature and calculation of citation indexes. Beside traditional citation indexes, which contain very small amount of useful data (e.g. title, author, institution, etc.) assigned for evaluation of scientific subjects, scientific paper search engines enable analysis of relationship between scientific papers, authors, research groups and institutions. For example, Cite Seer downloads PDF or PostScript files from the Web, converts them into text, and then parses files to extract the citation information.

In describing information extraction within previously mentioned services we will use the term *metadata* to refer to the structured information obtained from scientific papers that includes title, authors and the list of references – each including title, authors, year of publication and publishing source (journal, conference, etc). Numerous standards which define the structure of scientific paper (IEEE, ACM,

Springer, Elsevier, etc) are available. In the present work we limited our investigation on metadata extraction to IEEE standard.

Literature analysis in metadata extraction from scientific papers show that generally there are two approaches. One is based on templates (rules) and the other applies machine learning techniques.

First approach searches for previously defined patterns in the scintific paper and then processes extracted information. This technique is known as "template mining" and it is already aplied for reference extraction [4, 5, 6]. Patterns consists of rules which are related to sequence of assignment of reference elements, identifying common punctuation signs and also identificators of specified parts of reference (e.g. "Proc." stands for publication collection, etc.). However, often these patterns are not able to cover wide diversity of reference formats. Also, there is a problem in case of input errors, where every significant deviation from citation standard (e.g. missing quotes at the beginning or end of paper title) leads to disaibility of correct reference extraction.

Second approach applies machine learning techniques in order to find relations between elements contained in a reference. It uses training set of previously marked references to "learn rules" of reference assignment, which are then applied to input documents. Frequently used method for this purpose is Markov model (HMM) applied in [7, 8]. Support Vector Machines [9] as well as Neural Networks [10] also have shown good results in this field. Published results of most methods based on machine learning techniques have shown very good accuracy of over 90%.

The approach proposed in this paper is based on formally defined patterns. Rules for reference writing are defined as regular expressions and later extracted using finite state machine. Finite state machine implemented for reference extraction currently supports only IEEE standard, but there is a flexible mechanism for extension to other mostly used reference standards. In order to present extracted metadata more clearly and to emphasize the most cited documents and authors, graph and treemap visualization is applied. In the rest of the paper is a detailed description of each step of our approach, following by test results and planed activites for its improvement.

2 Metadata Extraction Approach

The proposed approach performs template-based metadata extraction from a document in PDF format. It consists of several major steps presented in Fig 1. The most challenging part of the approach is reference string parsing which is implemented using finite state machine concept. As an output of the algorithm, graph and treemap representations are generated. Bellow is a detailed explanation of each step.

2.1 PDF Document Parsing

First phase in the process of automatic data extraction from paper document is transformation of input document into format which is more suitable for analysis. Scientific papers can be saved in different formats: PDF, PostScript, HTML, Word, etc. Considering PDF is the most frequent file format in which publicly available scientific papers can be found, represented approach assumes this as input format. It should be mentioned that all input formats are not equally complex for metadata extraction. Particulary, HTML and Word documents contain extra information about style (font size, bold letters, etc.) which significantly makes easier recognition of specific segments of scientific paper. For example font size which is used for title writing is always larger then font size of the rest of the paper, which, with additional information, can be applied for title recognition. In contrast to HTML and Word documents parsing, most of PDF parsers returns only text without any additional informations about styles, which makes process of reference extraction even harder.

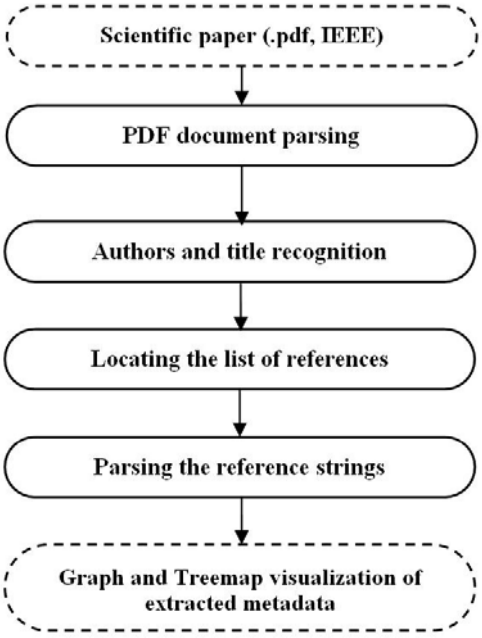


Fig. 1. The approach for metadata extraction

Although there is a number of APIs for PDF documents creation, only few of them enables PDF parsing. In our investigation, PDFNet [11] library is used for parsing of PDF documents. Its features include ability to extract additional text style information (font name, font size, etc).

2.2 Title and Authors Recognition

Usually, font size characteristic is used to determine what the title is. It occurs at the beginning of a paper in a larger font. Our approach starts searching paper from the begining until decreasing of font size is found. Extracted title words are saved in string which represent publication title. Possible problems occure in cases when the title has same font size as the rest of the paper, which is extremly rare in practise.

Once the title has been located, the authors come next. Beginning of the author names section is detected by the font size decreasing after the title. IEEE standard assumes that author names are defined in following order: first name at the begining, then middle name initial followed by fullstop, and surname at the end. Author names (if there are more than one) are terminated using commas, where the last author is preceded by word “and”. Extraction of author full names, that will be saved in authors list, is based on previous rules.

2.3 Locating the List of References

The most of the existing algorithms for reference extraction use the same technique for locating a list of references: Find a line of text that contains the heading “References”. Using the same principle, our approach searches the text starting from the end of paper and stops when it finds a line with a single word “References”. This is very time effective technique.

2.4 Parsing the Reference Strings

Parsing of the reference strings includes extraction of reference parts which are characterised by standard structure, fixed position inside reference and basic literal identificators. This is the most complex step in metadata extraction which offers wide range of abilities for implementation. As we mentioned in introduction, at this moment IEEE is used as reference standard for “template mining”. For the reference extraction, finite state machine is implemented, whose details are given bellow.

In order to define transition graph of finite state machine, it is necessary to determine proper rules for reference writing. As is well known, IEEE standard contains ordinal number at the begining of reference following with the list of authors, paper title, and finally publication information about the magazine or conference. Beginning of the reference is recognized by ordinal number in square brackets, which must be positioned in front of every reference. Two very important rules are that author names must be separated by commas and the last author before paper title is also separated by comma from paper title. Paper title begins with quotes, continues the text of the title and ends with comma and closed quotes.

The approach starts with extractions of lexems from previosly located list of references. Lexems represent arrays of characters between two “blank” signs. For each extracted lexem, its token is determined, where token represents type of lexem and its useful information. E.g., lexem which represents ordinal number is „[11]“, token for that lexeme is *OrdinalNumber* and useful information is 11. These tokens

represent input alphabet for finite state machine (Table 1.). Pseudocode of reference parsing algorithm is given in Fig 2. Procedure *TokenInitialization()* creates list of tokens based on string which is forwarded from extracted reference list in the PDF document. Transitions into the next state is defined by transition graph. Procedure *StateChange()* sets new current state, while *PerformOutputOperations()* pops necessary informations from stack and concatenates it to extracted reference list.

```
FSMReferenceExtraction()
{
  currentState = "Start";
  references[] = null;
  listTokens[] = TokenInitialization();

  foreach (token in listTokens[])
  {
    StateChange(currentState, token);
    PushOnStack(token);
    PerformOutputOperations(currentState, token,
      references[]);
  }

  return references[];
}
```

Fig. 2. Pseudocode of reference parsing algorithm

Token recognition in procedure *TokenInitialization()* is implemented by matching with previously defined regular expressions. Regular expressions which are used in the approach are presented in Table 1.

Table 1. Regular expressions used for reference string parsing (IEEE format)

Token	Regular expression	Example
Ordinal numeral of reference	\[[1-9]\d{0,2}\]	[27]
Author name	[A-Z]\.	D.
Author middle name	([A-Z]\.) |von|van|der|de|la	C.
Author surname	[A-Z]\w+	Chang
Beginning of title	“[A-Z]\w*	“Image
Title end (1)	\w+,”	transformation,”
Title end (2)	\w+”,	transformation”,
Publication	[A-Z]\w*	Trans.

(word)		
Year	(19|20)\d{2}	1998

However, represented regular expressions can not be strictly implemented because of common mistakes made by authors. Those errors could be a reason for incorrect recognition. Therefore, for practical implementation it is necessary to introduce additional rules to regular expressions, which are not formally correct, but very common in practice. One of those errors, which we would like to draw attention, is incorrect use of comma at the end of paper title in the reference section. Actually, very often there are errors in writing of allegation marks and commas at the end of reference (Example 1.), and a correction is also presented (Example 2.):

1. ~~"..... for Internet Presentation", Journal of Web...~~
2. "..... for Internet Presentation**,"** Journal of Web...

3 Metadata Visualization

As a prerequest for metadata visualization, appropriate format of extracted information has to be prepared. After metadata are extracted and saved in XML format, citation index is calculated inside test set. For each paper, a set of referenced papers is determined and stored as hash table. Hash table is then used to establish relations between papers.

For visualization of extracted metadata, we applied two techniques – graphs and treemaps. Graphs are standard approach for visualization of citation data. Using powerful graph layout algorithms we are capable to show large number of nodes (papers) and edges (references). Treemaps, are another approach that becomes very popular for information visualization. However, they are still not applied to the problem of scientific metadata visualization. In our research, we adapted treemaps for this specific purpose. These two approaches are shown as highly effective for concrete application, enabling easy visual identification of most cited authors and papers in specific field of science.

3.1 Graph Layout Algorithms for Scientific Metadata Visualization

Most common graph layout algorithms are those that use an analogy from force physics [17]. In general, these algorithms define an objective function which maps each graph layout into the energy of the layout. This function is defined in such a way that low energies correspond to layouts in which adjacent nodes are near some pre-specified distance from each other, and in which non-adjacent nodes are well-spaced. A layout for a graph is then calculated by finding a local minimum of the energy.

Among force-directed algorithms, one of the most popular is Fruchterman and Reingold algorithm [18], which is characterized with high speed and robustness. In each iteration it computes attractive forces between adjacent vertices and repulsive forces between all pairs of vertices. Since there is no scalar penalty for edge

crossings, the result is a vector, which points out not just how much out of place the vertex is (vector size) but also into which way it should move. The size of actual move is confined to current temperature, which is linearly decreased in each iteration.

Sample result of the graph visualization is presented in Fig 3. Each paper is rendered as one vertex. Referenced papers are connected with directed edge. Radius of the vertex corresponds to the citation index inside test set. In order to emphasize them, the most cited papers are colored in green.

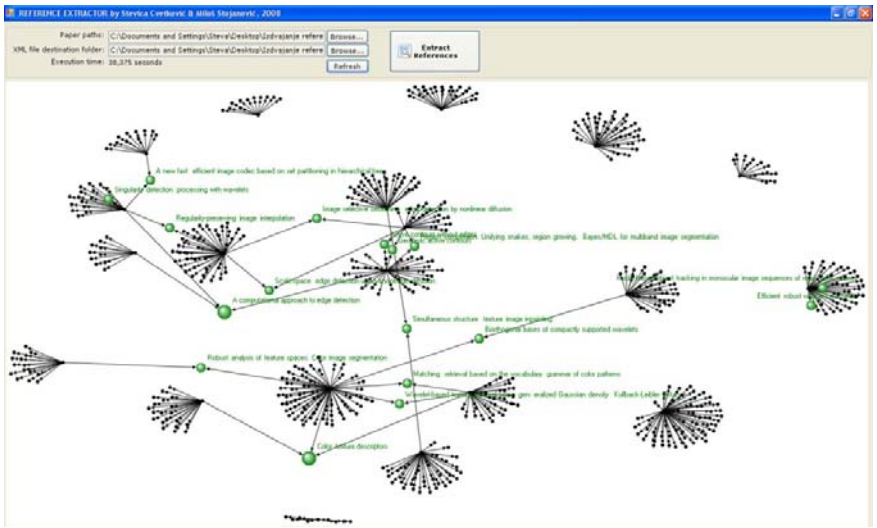


Fig. 3. Example of metadata visualization using graphs

3.2 Treemaps Application to Scientific Metadata Visualization

Treemaps were first designed by Shneiderman and Johnson during the 1990s in response to the common problem of a filled hard disk. Basic idea was to produce a compact visualization of directory tree structures. Later, treemaps were applied to the visualization of: hierarchical information [12, 13], search-tree [14], newsgroups [15], enterprise networks [16], etc.

Treemaps display data as groups of rectangles that can be arranged, sized and colored to graphically reveal underlying data patterns. This visualization technique allows efficient and compact display, which is particularly effective to show the size of the final elements in the structure. Besides the size of each rectangle, it is possible to represent other information by selecting the color used to fill the rectangle.

General concept of treemap representation could be explained on example of hierarchical data structures – trees. The whole tree could be represented as a rectangle. Each sub-tree is represented as a sub-rectangle of its parent rectangle. At the first level of the hierarchy, the whole rectangle is split vertically. Then, sub-rectangles are split horizontally. Sub-sub-rectangles are split vertically, and so on.

Each splitting is done so that the area covered by a rectangle is proportional to the number of nodes it contains.

Treemap which visualize extracted metadata from our test set of papers is shown in Fig 4. Most cited papers cover the largest areas on treemap and the most intense red color is assigned to them. Number at the top of each rectangle shows how many times papers inside that rectangle are cited. If there are more papers which are cited equal number of times, they will be shown in one rectangle, hierarchically as its "children". By clicking on a specific paper (e.g. rectangle) information about the paper is shown (authors, title, publication and year) as well as number and source of citations. Also, information about referenced papers and authors for specific paper can be obtained.

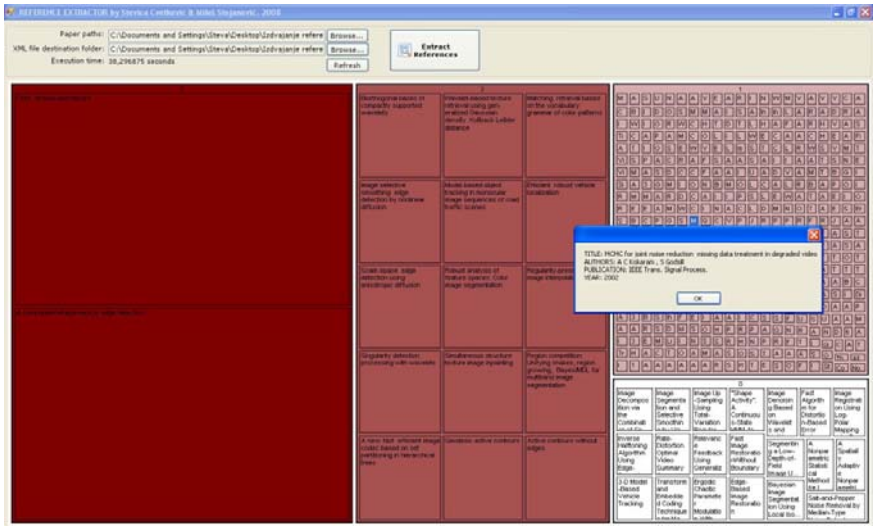


Fig. 4. Example of extracted metadata visualization using treemaps.

4 Results

In order to test the approach, desktop application was developed using Microsoft .NET technologies. For treemap rendering we adapted Data Visualization Components developed by the Microsoft Research Community Technologies Group. Graph visualization is implemented using open source NetMap control. Currently, user interface enables input of path to the folder which contains scientific papers saved as PDF files. We assume that one scientific paper is saved in one PDF file (see Fig. 4 and Fig. 5).

Tests were performed on a set of papers published in IEEE Transactions as well as few conferences where official paper format was also IEEE. Total number of scientific papers was 22, with 645 references in total. As a measure of success we used ratio between number of correctly extracted references and total number of

references (Table 2).

Table 2. Test results

Number of scientific papers	22
Total number of references	645
Number of successfully extracted references	597
Percent of successfully extracted references	92,6 %

5 Conclusion

Presented approach successfuly performs automatic metadata extraction from scientific papers, with limitation that references are in IEEE format. Extracted metadata are then analyzed, relations between them are established and visualized using graph and treemaps. Planned activities include improvement of reference extraction step by intorducing machine learning techiques which should provide robust method able to extract references without previous knowledge about applied format. Additionaly, it should enable automatic correction of common errors that could appear during input of reference data.

References

1. Lawrence, S., Giles, C. L., Bollacker, K.: Digital Libraries and Autonomous Citation Indexing. IEEE Computer Magazine, 32 (6), 67-71 (1999)
2. McCallum, A., Nigam, K., Rennie, J., Seymore, K.: Automating the Construction of Internet Portals with Machine Learning. Information Retrieval Journal, Kluwer, (3), 127-163 (2000)
3. Google Scholar, http://scholar.google.com
4. Bergmark, D,: Automatic extraction of reference linking information from online documents. Technical Report TR 2000-1821, Cornell University – Computer Science Department (2000)
5. Davenport, T., Delong, D., Beers, M.: Successful knowledge management projects. Sloan Management Review, vol. 39, 43-57 (1998)
6. Ding, Y., Chowdhury, G., Foo, S.: Template mining for the extraction of citation from digital documents. The Second Asian Digital Library Conference, pp. 47-62, Taiwan (1999)
7. Seymore, K., McCallum, A., Rosenfeld, R.: Learning hidden Markov model structure for information extraction. AAAI-99 Workshop on Machine Learning for Information Extraction, pp. 37-42 (1999)
8. Takasu, A.: Bibliographic attribute extraction from erroneous references based on a statistical model. The 3rd ACM/IEEE-CS joint conference on Digital libraries, pp. 49-60 (2003)

9. Han, H., Giles, C., Manavoglu, E., Zha, H., Zhang, Z., Fox, E.: Automatic Document Metadata Extraction using Support Vector Machines. Joint Conference on Digital Libraries (2003)
10. Mokriš, I., Skovajskova, L.: Feed-forward and self-organizing neural networks for text document retrieval. Acta Electrotechnica et Informatica, 8(2), 3–10 (2008)
11. PDFNet SDK, http://www.pdftron.com/net/
12. Shneiderman, B., Wattenberg, M.: Ordered treemap layouts. IEEE Symposium on Information Visualization, IEEE Press, Los Alamitos, USA (2001)
13. Bruls, M., Huising, K., Van Wijk, J.: Squarified treemaps. Joint Eurographics and IEEE TCVG Symposium on Visualization (TCVG 2000) IEEE Press, pp. 33-42 (2000)
14. Coulom, R.: Treemaps for Search-Tree Visualization. The Seventh Computer Olympiad Computer-Games Workshop (2002)
15. Fiore, A. T., Smith, M.A.: Interactive Poster: Treemap Visualizations of Newsgroups. Interactive Poster at IEEE Symposium of Information Visualization, Boston, Massachusetts (2002)
16. Goldberg, J.H., Helfman, J.I.: Enterprise Network Monitoring Using Treemaps. 49th Annual Meeting of the Human Factors and Ergonomics Society, Santa Monica, CA, pp. 671-675 (2005)
17. Herman, I., Melancon, G., Marshall, M.: Graph Visualization and Navigation in Information Visualization: a Survey. IEEE Transactions on Visualization and Computer Graphics 6(1), 24-43 (2000)
18. Fruchterman, T. M. J., Reingold, E. M.: Graph drawing by force-directed placement. Software Practice and Experience 21(11), 1129–1164 (1991)

Implementation of Academic Videoconferencing Infrastructure in Macedonia - The ViCES Project

Zdravko Stafilov

Makedonski Telekom, Orce Nikolov bb, 1000 Skopje, Macedonia
stafilov@t-home.mk

Abstract. The ViCES Project in Macedonia is the part for implementation of a videoconferencing infrastructure which covers 7 (seven) sites/universities with Polycom HDX 8000 codecs (platforms), plus a central Videoconferencing Management Centre. The management centre consists of Interconnect bridge server (Polycom RMX 1000), a Streaming and recording server (Polycom RSS 2000), and Converged management application server (Polycom CMA 4000). This paper addresses the successful implementation of this infrastructure, proves the quality, gives suggestions for overcoming some minor issues and network problems, as well as providing additional functionalities of the project overall, such as standardization of the QoS on the MARNET network, for internal usage and for interconnecting the videoconferencing system with the GEANT network. The paper also describes solutions for adding extra functionalities on the current platform with additional reconfiguration and ways to extend its capacity using other Polycom products such as the Video Media Centre server (Polycom VMC 1000).

1 Introduction

The ViCES Project in Macedonia is a project for implementation of a videoconferencing infrastructure which covers 7 (seven) sites/universities with Polycom HDX 8000 platforms, plus a central videoconferencing management centre [1]. The management centre consists of Interconnect bridge server (Polycom RMX1000), a Streaming and recording server (Polycom RSS2000), and Converged management application server (Polycom CMA4000). This infrastructure is displayed on Figure 1.

All of these locations (PMF Skopje, FEIT Skopje, FON University Skopje, Evropski Univerzitet Skopje, UGD Stip, UKLO Bitola and SEEU Tetovo) are interconnected and functional.

Primary location where all of the tests have been initiated for an endpoint testing was PMF Skopje, where the videoconferencing management center is located. The RMX 1000, RSS 2000 and CMA 4000 are also hosted at this location. Interconnectivity of all endpoints to this location is by using the MARNET network infrastructure, and separate sites have differently sized communication links [2].

M. Gusev (Editor): ICT Innovations 2010, Web Proceedings, ISSN 1857-7288

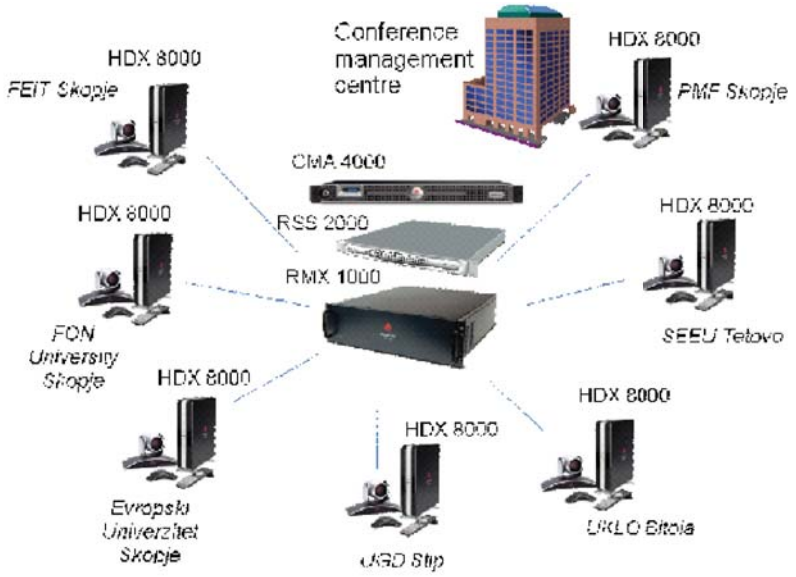


Fig. 1. Videoconferencing infrastructure of the ViCES project in MK

2 Testing the infrastructure and analysis of the results

The results from the tests are shown in Table 1. There were 6 endpoints tested from the project infrastructure, plus one external location (Leuven). Additionally, there are samples from the tests performed on the RMX 1000 with multipoint conferences.

These results show that all of the endpoints are working as expected without any problems with establishing and maintaining a quality videoconferencing call. The overall picture and sound quality was excellent and without any significant video artifacts (pixelization, smearing) or sound issues (distortion, hiccups).

Packet loss should be under 0.05% from end to end [3]. The actual packet loss in this infrastructure is non-existent (0 %).

The latency for a seamless call should be <150 ms [3]. The latency in our case is in the expected and permitted range, varying from 10 ms to 113 ms, with an average of 39 ms.

The jitter within the system should be less than 5 ms peak, for a total of 10 ms peak to peak [3]. The jitter in the tests is varying between 0 ms and 12 ms (one-way), and almost always the jitter is higher on the Rx which indicates that some of the remote endpoints are hosted on a congested networks or congested interconnections with MARNET. These issues should be further investigated so that the jitter is minimized within the expected range.

Table 1. Results from endpoint testing

End Point	Date	Time	Duration	TS Type	Rate (kbps)	Audio Protocol	Video Protocol	Picture Format	Pkt. Loss %	Latency	Jitter (Tx)	Jitter (Rx)
SEEU	4/23/10	9:30:11	0:01:46	h323	512	G.722.1C	H.264	CIF	0	66	0	2
SEEU	4/23/10	9:32:02	0:03:31	h323	512	G.722.1C	H.264	CIF	0	79	0	10
SEEU	4/23/10	9:46:35	0:02:54	h323	512	G.722.1C	H.264	CIF	0	60	0	12
UKLO	4/23/10	13:43:12	0:57:57	h323	384	Siren22	H.264	4CIF	0	113	1	11
FON	4/26/10	13:34:26	0:01:34	h323	512	Siren22	H.264	4CIF	0	10	0	2
UKIM	5/7/10	9:20:32	0:02:11	h323	384	Siren22	H.264	4CIF	0	10	3	3
UKIM	5/7/10	9:23:21	0:01:36	h323	384	Siren22	H.264	4CIF	0	19	0	1
UKIM	5/10/10	11:55:06	0:01:32	h323	1920	Siren22	H.264	720p	0	26	0	0
FEIT	5/31/10	10:56:51	0:33:42	h323	1920	Siren22	H.264	720p	0	37	3	2
FEIT	6/4/10	12:47:27	0:03:48	h323	384	Siren22	H.264	4CIF	0	30	2	2
FEIT	6/7/10	13:56:57	0:10:13	h323	1920	Siren22	H.264	720p	0	38	3	0
FON	6/2/10	9:52:07	0:01:57	h323	512	Siren22	H.264	4CIF	0	12	2	2
FON	6/2/10	15:03:13	0:01:39	h323	512	Siren22	H.264	4CIF	0	12	1	2
Leuven	6/4/10	9:45:15	0:29:59	h323	768	G.722	H.264	CIF	0	70	0	0
RMX	6/4/10	13:08:56	0:11:11	h323	1920	Siren22	H.264	4CIF	0	35	6	1
RMX	6/17/10	13:48:08	0:07:40	h323	1024	Siren22	H.264	CIF	0	30	6	1
RMX	6/17/10	14:04:05	0:01:45	h323	1920	Siren22	H.264	4CIF	0	12	2	1

These results indicate that the videoconferencing infrastructure is well implemented, but some network problems and interconnectivity issues between endpoint sites through MARNET network might introduce lower quality service. It is noticeable that the higher parameters in latency and jitter are present in a specific time period of the day, i.e. in the afternoon hours, when the academic network has higher utilization and thus higher congestion.

To overcome such problems, one of things that can be done is enabling QoS within separate network segments on the MARNET network infrastructure and in all of the endpoint networks in order to guarantee a seamless quality of the live videoconferences and their live and recorded streaming reproductions. (This will be a strong point for defining the SLAs.)

Currently, there are no standards that describe which QoS values are most appropriate for IP video conference. However, since the MARNET network is interconnected with GÉANT network, the recommended values should be the same as the DSCP/ToS values used in GÉANT [4] to classify the traffic of the different QoS classes, shown in Table 2. In addition to the three service classes offered to transiting traffic, there is a DWS (IP commodity service) and a Network Control class, which are traffic classes used internally to the GÉANT network.

Table 2. GÉANT Recommendation for QoS markings

Service	DSCP value	ToS value	Juniper alias	ToS (hex)	DSCP-ToS binary	
Premium IP	46	184	ef	B8	101110 101110xx	-
LBE	8	32	cs1	20	001000 001000xx	-
DWS	32	128	cs4	80	100000 100000xx	-
Network control 1	48	192	cs6	C0	110000 110000xx	-
Network control 2	56	224	cs7	E0	111000 111000xx	-

3 Tests & Analysis of the Videoconferencing Management Centre

The videoconferencing management centre is located at PMF – Skopje, and represents a centralized place for managing end-to-end and multipoint conferences, recording of the conferences, streaming of live & prerecorded streams, scheduling conferences, etc. It has three main components: Interconnect Bridge server (Polycom RMX1000); Streaming and Recording server (Polycom RSS2000); and Converged Management Application server (Polycom CMA4000).

The RMX 1000 interconnect bridge [5] (multipoint conference platform) is fully functional, and has been tested with end-to-end and multipoint conferences. Table 3. shows an example of a successful multipoint conference with 6 participants, with a line rate of 1920 Kbps. Since the RMX 1000 is hosted in the PMF Skopje site, all of the results in Table 1. are also applicable for this test.

Table 3. Testing of multipoint conference

Conference Name	**C20100507 09:55 Test Conference**
Start Time	07.05.2010, 07:55:42
Duration	02:00:00
Line Rate	1920 kbps
Participant 1 Name	PMF
IP Address	194.149.135.79
Participant 2 Name	UKIM-Rectorate
IP Address	194.149.131.115
Participant 3 Name	UKIM-Filoloski
IP Address	194.149.155.252
Participant 4 Name	UNINETTUNO
IP Address	88.35.45.237
Participant 5 Name	UKLO
IP Address	194.149.159.70
Participant 6 Name	UGD
IP Address	194.149.135.75

The RSS 2000 recording and streaming server [6] has been installed and fully functional. Testing with recording of conferences has been successful. Table 4. shows a list of a few successfully made recordings, in different bitrates, different formats, as well as one example of a recorded multipoint session. All of the recordings were then tested for playback through the interface of the RSS 2000 web interface and they played without any problems.

Table 4. List of a successful test recordings

Archive Name	**Start Date & Time**	**Duration**	**Rate (Kbps)**	**Video**	**File Size (Kb)**
admin - Thu Jun 17 20104505_JUN_17_2010_13:22	06/17 2010 13:22:38	00:00:07	1024	H.264	882
C20100617 14:09_JUN_17_2010_13:10	06/17 2010 13:10:46	00:00:04	1920	H.264 SD	913
C20100617 14:00_JUN_17_2010_13:05	06/17 2010 13:05:14	00:00:17	1920	H.264 SD	3827

admin - Thu Jun 17 20106818_JUN_17_2010_12:54	06/17 2010 12:54:05	6	00:01:1	1024	H.264	8938
Proba_snimanje_FEIT JUN_17_2010_12:09	06/17 2010 12:09:09	7	00:13:1	1920	H.264 SD	1723 88

The CMA 4000 Converged Management Application server [7] is successfully installed and put into operation, but from its multiple functionalities, only one is fully implemented: creating a multipoint conference through the CMA Web Scheduler. On Picture 1. it is clearly shown that there were 6 successfully scheduled multipoint conferences and 85 successful adhoc calls. Picture 2. shows the call detail log of these calls.

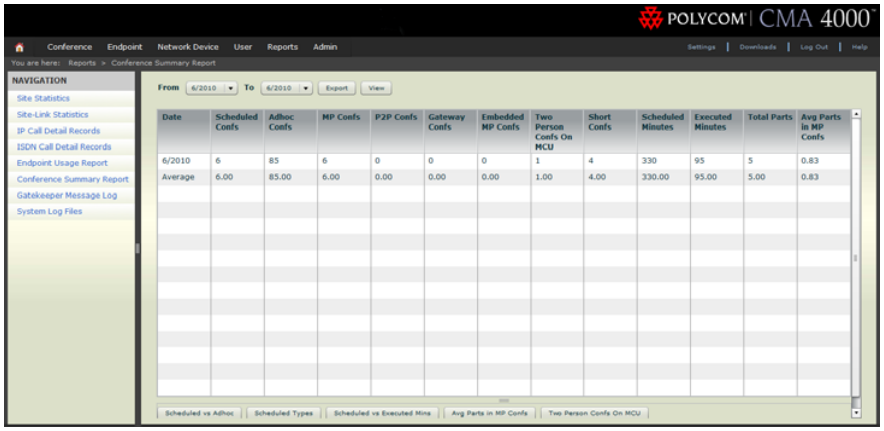


Fig. 2. Screen capture of the CMA 4000 Conference Summary Report

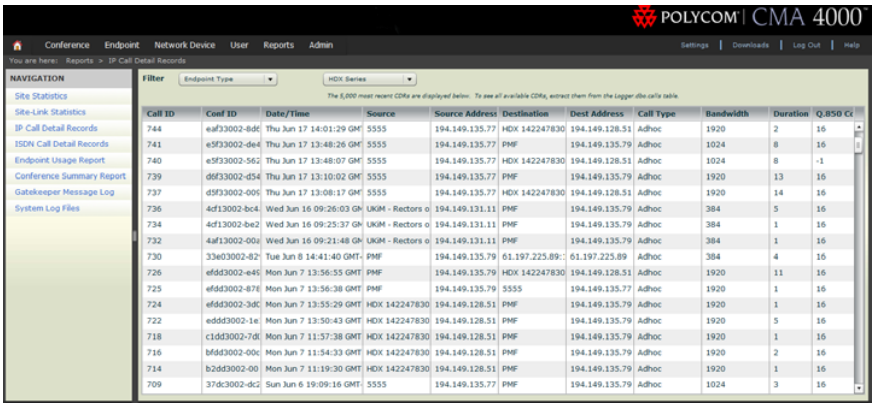


Fig. 3. Screen capture of CMA 4000 Call Detail Records

The other functionalities of CMA 4000 can be implemented, and afterwards tested, only after implementing the automatic and scheduled provisioning of the endpoints.

Since the Polycom CMA 4000 system is a gatekeeper - it manages video and audio devices. However, the system also manages users, because devices are only useful when they provide access to users. Automatic device provisioning, which controls the automatic configuration of devices and the management of video resources, is also tied to users and groups. It is done by applying a provisioning profile to a one or more endpoints at once. The following steps are required in the process of successful implementation the endpoint provisioning [7]:

- Automatic Provisioning - is used to ensure out-of-the-box usability of the endpoints. When a device starts up and at designated intervals thereafter, it automatically polls for new provisioning information from the Polycom CMA system. To ensure this, the following steps need to be done:
  - Define an automatic provisioning profile (works only for HDX systems deployed in dynamic management mode and CMA Desktop Clients);
  - Configure the profile to meet the requirements of the users;
  - Set a profile order and priority.

- Scheduled Provisioning - is a standard/traditional management mode. One can schedule provisioning for one device or a group of devices and they can schedule provisioning to occur immediately or for a date and time in the future:
  - Define a scheduled provisioning profile;
  - Configure the profile to meet the requirements of the users;
  - Apply the profiles to the desired endpoints.

4 Recommendations for enhancing the videoconferencing experience

- The organization support for this project is still into the Project Management organizational structure. After the project is finished, there should be a

separate organizational unit responsible for maintenance and operation of the video conferencing systems which will: set up conferences (end-to-end and multipoint), maintaining the educational activities, defining the lesson structures, moderating the live conferences, administering the recorded sessions, maintaining the equipment, ensuring sustainability and growth, etc.

- Defining a standard for the room requirements. There are a lot of room design types applicable to this project, but the suggestion is to define a standard which will be a mandatory for all the participants in this project, so that the end user experience is not degraded in the different sites.

- Extending the infrastructure with VMC 1000 streaming server in order to provide a wider accessibility and large number of clients for live broadcasting sessions [8], since the RSS 2000 is primary a recording server with a limited streaming capabilities in terms of number of simultaneous streams. The VMC 1000 works together with RSS 2000 and reproduces the live and recorded streams. This is a recommended next step after a video portal is created and its usage starts to grow. Figure 2. displays a diagram of a placement of VMC 1000 at the central site (PMF – Skopje) for supporting up to 1000 clients:

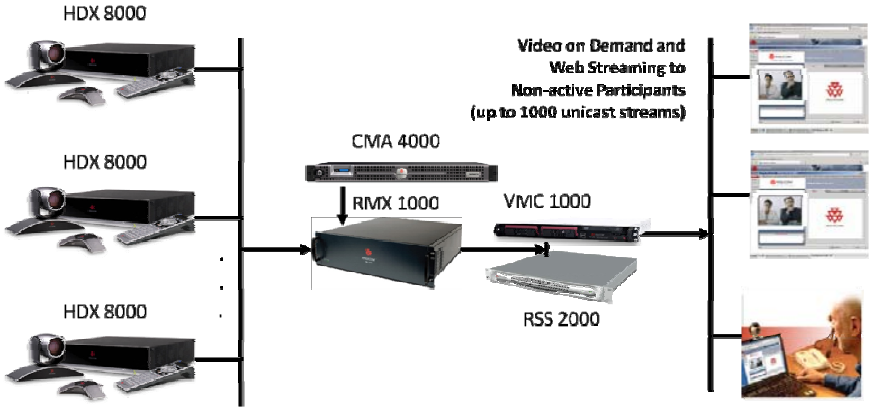


Fig. 4. Extending the videoconferencing infrastructure with VMC 1000

- Enabling the H.323 gatekeeper functionalities on the CMA 4000 [5] when the project interconnects with other videoconferencing nodes on the GÉANT

network. A H.323 gatekeeper is an important element of a H.323 network providing the following services:

- Number based dialing and call routing: calls can be initiated using normal phone numbers. Using of IP addresses or domain names are not needed - apart from that you have the possibility -, calls can be placed by typing usual phone numbers (e.g. 00389xx123456).
- Call admission: after placing a call from a terminal, the local gatekeeper is needed to decide whether the calling party is allowed to originate or receive calls.
- Call authorization: authorization of the terminal to enter the network, to place calls, etc.
- Bandwidth control: controlling call bandwidth.
- Accounting information: generating call detail records (CDR).

In order to use gatekeeper services with a H.323 endpoint (e.g. other videoconference endpoints, Microsoft Netmeeting, etc.), the endpoint must register at the gatekeeper. The gatekeeper then keeps track of registered endpoints and handles calls initiated by them.

5 Conclusion

The ViCES project has managed to implement fully functional videoconferencing infrastructure in R. of Macedonia. The equipment is deployed, codecs are setup and all of the building blocks at the conferencing management centre are up and running.

Tests of the solution overall proved that ViCES established a stable infrastructure for videoconferences, a modular platform capable for future growth, with minor network issues which can be resolved with reconfigurations as recommended in this paper.

Recommendations for enhancing the videoconferencing experience address some important points such as providing operational support (the organizational structure), better user experience (defining the standard for the room requirements), wider accessibility (extending the infrastructure with a streaming server), as well as extra functionalities (enabling the H.323 gatekeeper functionalities) for having more comprehensive infrastructure.

The ViCES project implemented an infrastructure capable of serving the complete academic needs for telecommuting, and presents a successful tool for expanding the possibilities for advanced learning methods and techniques, collaboration, time and cost savings, in addition to fulfilling the most imperative one - distance learning and lecturing.

References

1. Trajkovik V., Enrica C.: Video Conferencing as an Engineering Education System. In: SEFI 2009 Annual Conference, Rotterdam, Netherlands (2009)
2. Trajkovik, V., Caporali, E., Valdiserri, J., Palmisano, E., Cekorov, B.: Establishing Videoconferencing Infrastructure in R. of Macedonia. In: 9th International Conference on Information Technology Based Higher Education and Training - ITHET 2010, Cappadocia, Turkey (2010)
3. Stafilov, Z.: Cisco Telepresence Implementation for Telekom's Corporate Requirements. In: Poster sessions of 1st ICT Innovations Conference 09, Ohrid, Macedonia (2009)
4. GÉANT: GIP Quality of Service in GÉANT. http://archive.geant.net/server/show/conWebDoc.500, (2010)
5. Polycom, Inc.: Polycom RMX 1000 User Guide. http://www.polycom.com/global/documents/support/user/products/network/RMX1000_V1_1_User_Guide_English.pdf, (2010)
6. Polycom, Inc., Polycom RSS 2000 User Guide. http://www.polycom.com/global/documents/support/user/products/network/RSS2000_User_Guide_V4_0.pdf, (2010)
7. Polycom, Inc.: Polycom CMA System Operations Guide. http://www.polycom.com/global/documents/support/user/products/network/CMA_Operations_Guide_v412.pdf, (2010)
8. Polycom, Inc.: Polycom Video Media Center VMC 1000 Data Sheet. http://www.polycom.com/global/documents/products/telepresence_video/datasheets/vmc1000-datasheet.pdf, (2010)

Ensembles of Binary SVM Decision Trees

Gjorgji Madjarov, Dejan Gjorgjevikj and Tomche Delev

Faculty of Electrical Engineering and Information Technologies, Ss. Cyril and Methodius University, Rudjer Boskovic bb, 1000 Skopje, Macedonia
{madzarovg, dejan, tdelev}@feit.ukim.edu.mk

Abstract. Ensemble methods are able to improve the predictive performance of many base classifiers. In this paper, we consider two ensemble learning techniques, bagging and random forests, and apply them to Binary SVM Decision Tree (SVM-BDT). Binary SVM Decision Tree is a tree based architecture that utilizes support vector machines for solving multiclass problems. It takes advantage of both the efficient computation of the decision tree architecture and the high classification accuracy of SVMs. In this paper we empirically investigate the performance of ensembles of SVM-BDTs. Our most important conclusions are: (1) ensembles of SVM-BDTs yield noticeable better predictive performance than the base classifier (SVM-BDT), and (2) the random forests ensemble technique is more suitable than bagging for SVM-BDT.

Keywords: Ensembles, Bagging, Random Forests, Support Vector Machines, Binary decision tree.

1 Introduction

The recent results in pattern recognition have shown that support vector machine (SVM) [1][2][3] classifiers often have superior recognition rates in comparison to other classification methods. However, the SVM was originally developed for binary decision problems, and its extension to multi-class problems is not straightforward. The popular methods for applying SVMs to multiclass classification problems usually decompose the multi-class problems into several two-class problems that can be addressed directly using several SVMs. Similar to these methods, we have developed an architecture of SVM classifiers utilizing binary decision tree (SVM-BDT) for solving multiclass problems [4]. This architecture uses hierarchy clustering algorithm to convert the multi-class problem into a binary tree of SVMs. The binary decisions in the non-leaf nodes of the binary tree are made by the SVMs. The SVM-BDT architecture [4] uses Euclidean distance in the clustering process for measuring the similarity between classes.

The goal of this paper is to investigate whether ensemble methods [5] can be applied to SVM-BDT in order to achieve better performance. Ensemble methods

construct a set of classifiers for a given prediction task and classify new data instances by taking a vote over their predictions. Ensemble methods typically improve the predictive performance of their base classifier [5]. In this paper, we use SVM-BDT as base classifiers. The ensemble methods that we investigate are bagging [5] and random forests [6]. More precisely, the main question we want to answer is: Does building ensembles of SVM-BDTs really improve the predictive performance and which of the two ensemble techniques is more suitable for SVM-BDT. The last comparison is made along running times (training and testing).

The paper is organized as follows. In Section 2, we briefly discuss ensemble methods. Section 3 explains SVM-BDT in more detail. Section 4 presents a detailed experimental evaluation. Conclusions and some ideas for further work are presented in Section 5.

2 Ensemble methods

An ensemble is a set of classifiers constructed with a given algorithm. Each new example is classified by combining the predictions of every classifier from the ensemble. These predictions can be combined by taking the average (for regression tasks) or the majority vote (for classification tasks), as described by Breiman [5], or by taking more complex combinations [7][8][9]. A necessary condition for an ensemble to be more accurate than any of its individual members, is that the classifiers are accurate and diverse [10]. An accurate classifier does better than random guessing on new examples. Two classifiers are diverse if they make different errors on new examples. There are several ways to introduce diversity: by manipulating the training set (by changing the weight of the examples [5][11] or by changing the attribute values of the examples [12]), or by manipulating the learning algorithm itself [13]. In this paper, we consider two ensemble learning techniques that have primarily been used in the context of decision trees: bagging and random forests.

2.1 Bagging

Bagging [5] is an ensemble method that constructs the different classifiers by making bootstrap replicates of the training set and using each of these replicates to construct one classifier. Each bootstrap sample is obtained by randomly sampling training instances, with replacement, from the original training set, until an equal number of instances are obtained. Breiman [5] has shown that bagging can give substantial gains in predictive performance, when applied to an unstable learner (i.e., a learner for which small changes in the training set result in large changes in the predictions).

2.2 Random Forest

A random forest [6] is an ensemble of classifiers, where diversity among the predictors is obtained by using bagging, and additionally by changing the feature set

during learning. More precisely, for each base classifier (SVM-BDT) in the ensemble, a random subset of the input attributes is taken. The number of attributes that are retained is given by a function f of the total number of input attributes x (e.g., $f(x) = 1$, $f(x) = x^{1/2}$, $f(x) = log_2(x) + 1$ . . .). By setting $f(x) = x$, we obtain the bagging procedure.

3 Support Vector Machines Utilizing a Binary Decision Tree

SVM-BDT (Support Vector Machines utilizing Binary Decision Tree) [4] is tree based architecture which contains binary SVM in the non leaf nodes. It takes advantage of both the efficient computation of the tree architecture and the high classification accuracy of SVMs. Utilizing this architecture, N-1 SVMs are needed to be trained for an N class problem, but only log_2N SVMs in average are required to be consulted to classify a sample. This leads to a dramatic improvement in recognition speed when addressing problems with big number of classes.

The hierarchy of binary decision subtasks should be carefully designed before the training of each SVM classifier. There exist many ways to divide N classes into two groups, and it is critical to have proper grouping for the good performance of SVM-BDT. The SVM-BDT method is based on recursively dividing the classes in two disjoint groups in every node of the decision tree and training a SVM that will decide in which of the groups the incoming unknown sample should be assigned. The groups are determined by a clustering algorithm according to their class membership and their interclass distance. SVM-BDT method starts with dividing the classes in two disjoint groups g_1 and g_2. This is performed by calculating N gravity centers for the N different classes and the interclass distance matrix. Then, the two classes that have the biggest Euclidean distance from each other are assigned to each of the two clustering groups. After this, the class with the smallest distance from one of the clustering groups is found and assigned to the corresponding group. The gravity center of this group and the distance matrix are then recalculated to represent the addition of the samples of the new class to the group. The process continues by finding the next unassigned class that is closest to either of the clustering groups, assigning it to the corresponding group and updating the group's gravity center and distance matrix, until all classes are assigned to one of the two possible groups. This defines a grouping of all the classes in two disjoint groups of classes. This grouping is then used to train a SVM classifier in the root node of the decision tree, using the samples of the first group as positive examples and the samples of the second group as negative examples. The classes from the first clustering group are being assigned to the left sub-tree, while the classes of the second clustering group are being assigned to the right sub-tree.

The process continues recursively (dividing each of the groups into two subgroups applying the procedure explained above), until there is only one class per group which defines a leaf in the decision tree. The recognition of each sample starts at the root of the tree. At each node of the binary tree a decision is being made about the assignment of the input pattern into one of the two possible groups represented by

transferring the pattern to the left or to the right sub-tree. This is repeated recursively downward the tree until the sample reaches a leaf node that represents the class it has been assigned to.

An example of SVM-BDT that solves a 7 - class pattern recognition problem utilizing a binary tree, in which each node makes binary decision using a SVM is shown on Fig. 1. a, while Fig. 1. b illustrates grouping of 7 classes.

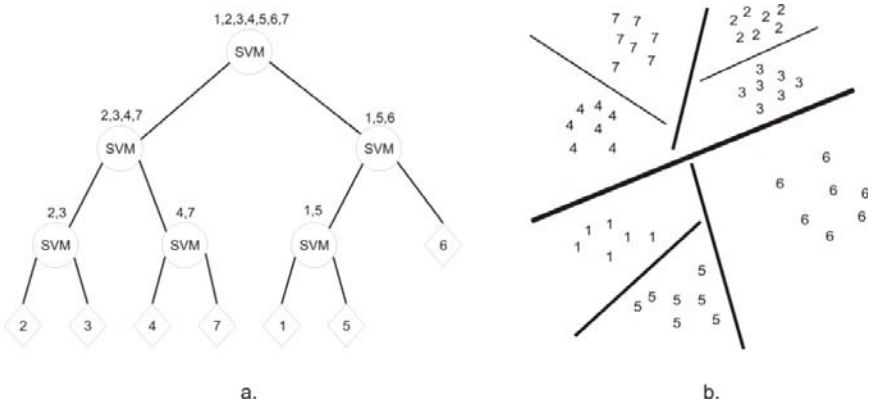


Fig. 1. a. SVM-BDT architecture; b. SVM-BDT divisions of seven classes

4 Experimental Evaluation

In this section, we describe the experimental methodology, the datasets, and the obtained results. The performances were measured on the problem of recognition of digits, letters and medical images.

We empirically evaluate two ensemble learning techniques, bagging and random forests, and apply them to Binary SVM Decision Tree (SVM-BDT). In order to combine the predictions output by the base classifiers, we apply the simple majority vote method. Each ensemble consists of 100 trees. For building random forests, the parameter $f(x)$ was set to $\lfloor log_2 x \rfloor$. The performances of the ensembles were also compared to the performance of the base classifier.

In our experiments, five different multi-class classification problems were addressed by each method (two ensemble methods and the base classifier SVM-BDT). The training and testing time and the recognition performance were recorded for every method.

The first problem was recognition of isolated handwritten digits from the MNIST database [14]. The MNIST database contains grayscale images of isolated handwritten digits. From each digit image, after performing a slant correction, 40 features were extracted. The features are consisted of 10 horizontal, 8 vertical and 22 diagonal projections [15]. The second and the third problem are 10 class problems from the UCI Repository [16] of machine learning databases: Optdigit and Pendigit.

The fourth problem was recognition of isolated handwritten letters, a 26-class problem from the Statlog collection [17]. The fifth problem was recognition of medical images, a 197-class problem from the IRMA2008 collection [18]. The medical images were described with 80 features obtained by the edge histogram descriptor from the MPEG7 standard [19]. The complete description of the datasets (number of classes, number of features, number of training and testing samples) is shown in Table 1.

Table 1. Datasets description

Dataset	# of classes	# of features	# of training samples	# of testing Samples
MNIST	10	40	60000	10000
Pendigit	10	16	7494	3498
Optdigit	10	64	3823	1797
Statlog	26	16	15000	5000
IRMA2008	197	80	12706	1000

In all classification problems the classifiers were trained using all available training samples of the sets and were evaluated by recognizing all the test samples from the corresponding set. All tests were performed on a personal computer with an Intel Core2Duo processor at 1.86GHz and 3GB of RAM under the Windows XP operating system.

The training and testing of the ensembles and base classifier were performed using a custom developed application that uses the Torch3 library [20]. The Torch3 library utilizing the SVMs with Gaussian kernel were used for solving the partial binary classification problems, we have used.

Table 2 through Table 4 shows the results of the experiments using the application of bagging and random forests and the base classifier (SVM-BDT) on each of the 5 data sets. Table 2 gives the prediction error rate of the ensembles and the base classifier applied on each of the datasets. Table 3 and Table 4 show the testing and training time of the ensembles and the base classifier, for the datasets, measured in seconds, respectively.

The results in Table 2 show that for all datasets, the ensemble methods achieved better prediction accuracy comparing to SVM-BDT base classifier. The error rates achieved by both ensemble methods are very similar. However, the bagging method achieved slightly lower error rates for all datasets except for the Optdigit dataset.

Table 2. The prediction error rate %

Dataset	SVM-BDT	Bagging	Random Forests
MNIST	2.45	1.88	1.90
Pendigit	1.94	1.89	1.92
Optdigit	1.61	1.26	1.24
Statlog	4.54	2.88	2.90
IRMA2008	55.80	48.00	48.00

Table 3. Testing time in seconds

Dataset	SVM-BDT	Bagging	Random Forests
MNIST	25.33	2312.76	419.67
Pendigit	0.54	48.54	32,17
Optdigit	0.70	72.46	31.12
Statlog	13.10	1145.82	603.12
IRMA2008	6.50	548.66	195.33

Table 4. Training time in seconds

Dataset	SVM-BDT	Bagging	Random Forests
MNIST	304.25	28455.96	16345.3
Pendigit	1.60	148.34	126,22
Optdigit	1.59	182.45	131.43
Statlog	63.30	6302.67	4252.13
IRMA2008	75.10	6934.59	2884.35

The testing and the training times are shown in Table 3 and Table 4. As expected, the training and testing times for the bagging method are about 100 times slower than the times of a single base classifier for all datasets (the number of the base classifiers in the ensembles was 100). The random forest ensemble is slightly faster than Bagging when training and considerably faster while testing. We believe this is to the smaller size of the feature vector used in the recognition process of the random forests method. The obtained results show that the dependency of the training and testing times of the SVMs in the binary decision trees are not proportional to the size of the reduced feature vector used by the random tree ensemble.

In overall, although the bagging ensemble method has shown slightly better recognition accuracy, the random forests is more suitable ensemble method for SVM-BDT because of its considerably lower time complexity.

5 Conclusion

In this paper, an empirical study on applying ensemble methods to SVM-BDT is presented. The results can be summarized as follows. First, the performance of a SVM-BDT is improved by learning an ensemble (using bagging or random forests) of SVM-BDTs. Second, the random forests ensemble technique is more suitable ensemble technique than bagging for SVM-BDT because of its considerably lower time complexity.

Although the time complexity of random forests ensemble method is lower, it is important to consider that the dependency of the training and testing times of the SVMs in the binary decision trees are not proportional to the size of the reduced feature vector used by this ensemble method.

As future work, we plan to extend the empirical evaluation along two dimensions: (a) how different feature selection strategies can influence on random forests ensemble method; and (b) how boosting ensemble methods can be applied to SVM-BDT.

References

1. Vapnik, V.: The Nature of Statistical Learning Theory, Springer, New York, (1999)
2. Burges, C., J., C.: A tutorial on support vector machine for pattern recognition. Data Min. Knowl. Disc. 2 (1998) 121
3. Joachims, T.: Making large scale SVM learning practical. in B. Scholkopf, C. Bruges and A. Smola (eds). Advances in kernel methods-support vector learning, MIT Press, Cambridge, MA, (1998)
4. Madzarov, G., Gjorgjevikj, D., Chorbev, I.: A multi-class SVM classifier utilizing binry decision tree, An International Journal of Computing and Informatics, Informatica, Volume 33 Number 2, ISSN 0350-5596, Slovenia, (2009) 233-241
5. Breiman, L.: Bagging predictors. Machine Learning 24(2) (1996) 123–140
6. Breiman, L.: Random forests. Machine Learning 45(1) (2001) 5–32
7. Ho, T., Hull, J., Srihari, S.: Decision combination in multiple classifier systems. IEEE Trans. on Pattern Anal. and Mach. Intell. 16(1) (1994) 66–75
8. Kittler, J., Hatef, M., Duin, R., Matas, J.: On combining classifiers. IEEE Trans. on Pattern Anal. and Mach. Intell. 20(3) (1998) 226–239
9. Gorgevik, D., Cakmakov D.: Combining SVM Classifiers for Handwritten Digit Recognition. Proceedings of 16th Int. Conference on Pattern Recognition, ICPR2002, Vol. 3, SII.8p, IEEE Computer Society, Quebec City, Canada, (2002)
10. Hansen, L., Salamon, P.: Neural network ensembles. IEEE Trans. on Pattern Anal. and Mach. Intell. 12 (1990) 993–1001
11. Freund, Y., Schapire, R.E.: Experiments with a new boosting algorithm. In: Proc. of the 13th ICML, Morgan Kaufmann (1996) 148–156
12. Breiman, L.: Using adaptive bagging to debias regressions. Technical report, Statistics Department, University of California, Berkeley (1999)
13. Dietterich, T.: Ensemble methods in machine learning. In: Proc. of the 1th Int'l Workshop on Multiple Classifier Systems. Volume 1857 of LNCS. (2000) 1–15
14. MNIST, MiniNIST, USA http://yann.lecun.com/exdb/mnist
15. Gorgevik, D., Cakmakov, D., An Efficient Three-Stage Classifier for Handwritten Digit Recognition, Proceedings of 17th ICPR2004. Vol. 4, IEEE Computer Society, Cambridge, UK, (2004) 507-510
16. Blake, C., Keogh, E., Merz, C. UCI Repository of Machine Learning Databases, (1998), http://archive.ics.uci.edu/ml/datasets.html [Online]
17. Statlog, http://archive.ics.uci.edu/ml/datasets/Letter+Recognition [Online]
18. http://www.imageclef.org/2008/medaat
19. Martinez, J. M. ed., MPEG Requirements Group, ISO/MPEG N4674, Overview of the MPEG-7 Standard, v 6.0, Jeju, Mar. 2002
20. Collobert, R., Bengio, S., Mariethoz., J.: Torch: a modular machine learning software library. Technical Report IDIAP-RR 02-46, IDIAP, 2002.

Integrated Medical Systems for Improvement of Consultations Between Physicians

Draško Nakik, Vladimir Trajkovik, Suzana Loškovska

Ss. Cyril and Methodius University
Faculty of Electrical Engineering and Information Technologies,
Karpos 2 bb. Skopje, Macedonia
drasko_21@yahoo.com, {trvlado,suze}@feit.ukim.edu.mk

Abstract. In this paper we want to represent the integration of two different and independent remote monitoring medical systems with a consultation system for physicians - Internet Medical Consultant (IMC). The first system is Critical Homecare System (CHS) and the other one is Heartbeat Tracker (HT). The CHS systems monitors the vital signals for chronically ill, convalescents or elderly people with high risk of health disorder, while the HT monitors the heart function for people that go for outdoor physical activity. If integrated with the IMC, these two systems can provide valuable data to enrich the quality of consultations and decision made by its users. By this example the paper should emphasize the importance of medical systems integration towards creating information rich consultations between physicians for improving the decision making process.

Keywords: decision making, medical systems, consultation system, integration, collaboration

1 Introduction

Physicians engage solving diverse problems in their everyday practice. Those problems range from very trivial ones to very complex problems. In occasions when a physician encounters an unfamiliar problem, it is very important to have reliable data and sufficient information to draw the right conclusion. In the modern ICT era harvesting data from patients should be integrated and those data should always be available to the physicians. Therefore, the data that is obtained from various monitoring system and hospital DBSs' should be pushed to integration and processed integrally to provide complete and consistent information to the physicians. Additionally, monitoring systems of body signals that can contribute to the overall image of the patient condition should be designed and considered to be a part of the puzzle. Only that form of information can improve the quality of decision making and decrease the time needed for it.

To face this demand it is always useful to view medical systems as a part of a whole, rather than independent systems. This especially stands for systems that can be utilized as a support for medical consultations between physicians. In this paper we want to show how three systems, which were initially designed as separate sys-

tems, can be integrated to serve a function of consultation enrichment. The development of the systems is a work in progress and it reached a point where it is important to consider their integration and carefully design functionalities for that purpose.

The first system is called Critical Homecare System - CHS. This system is designed to monitor the vital signals of chronically ill, convalescents and elderly people with a risk of health disorders. The CHS continuously monitors the vital signals of a patient while at home and signals critical conditions upon which starts real-time vital signals data streaming to the physicians at the urgent centre and the ambulance paramedic team [4, 5]. This on-time provision with important data contributes to more efficient intervention at the urgent centre and on the point of emergency [6, 7, 8]. The second system is Heartbeat Tracker – HT [3]. The purpose of this system is to monitor the heart function with several groups of people that go for activities that take place in outdoor conditions – mainly outside urban areas. These are people involved in sports like cycling, hiking, walking and cross-country running; extreme sportsmen; people that have some calculated risk of arrhythmia when exposing their body to superoptimal effort like: people who have hypertonia, people that suffered heart attack, brain attack or brain bleed, diabetes patients and people who suffer from problems with heart electrical conductivity; and people in their 50's and above which are getting more involved in sport activity in the past years, and are more prone to arrhythmia under continuous physical effort. The HT recognizes arrhythmia and alerts the user for the developed situation, and suggests ways to temporarily overcome that condition while medical help arrives. The third system that participates in the integration is the Internet Medical Consultant – IMC [1, 2]. This system is to be used by physicians and enables integration of the medical knowledge with providing asynchronous and synchronous consultations between physicians utilizing the modern ICT technologies. The main goal of this system is to come closer to the ideal of integrated healthcare: "*All doctors and all patients in the world are in the same room, with all the knowledge resources they need at hand"*. Having in mind that patient data is valuable knowledge resource, the IMC acts as the integration point of CHS and HT. Namely, both CHS and HT provide information to the IMC that is used to support the consultations between its users.

Fig. 1 shows the integration of the three systems. CHS and HT provide the data and the IMC user makes a decision after performing a consultation. The decision is then sent back to a user of the data providing systems whose data was analyzed in the IMC system. This way the feedback that the users of the data providing systems receive is more reliable because the physician who analyzes the results is networked with other physicians to ask for a second opinion and discuss what he observes.

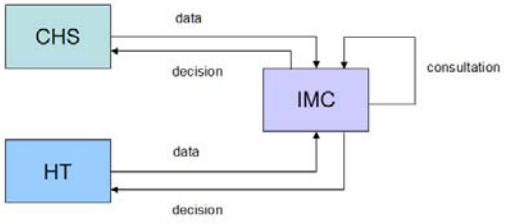


Fig. 1. The integration of CHS, HT and IMC

The paper is organized as follows: in Section 2 and Section 3 we describe the general ideas and functions of the data providing systems, CHS and HT respectively. In Section 4 we introduce the IMC system as a point of integration of the systems described in the previous two sections and elaborate the benefits of the integration. In Section 5 we give a conclusion for the ideas represented in this paper and set the framework for the future work in medical system integration.

2 The Critical Homecare System

The complete architecture of the system is represented in Fig. 2. The VFMD sends RT VF signals to patient's PDA (all PDAs in the system are phone enabled) via Bluetooth. The PDA forwards the data wirelessly, trough Home LAN to the HS, which acts as a VF data hub: it collects VF data from patients and distributes them to all the actors in the system. The HS is general processor of the data which is responsible for storing data and rising alert signals. Both Internet and MTN are utilized for communication between the actors: textual, voice and video. Doctors analyze the data and react upon them giving advice and instructions to the actuator - the ambulance, which on the other hand cooperates remotely with the UC the patient is supposed to be transported to. The patient is included in the communication if able to communicate. Audio-video signal is sent from the ambulance to the UC for more realistic picture of patient's condition [6].

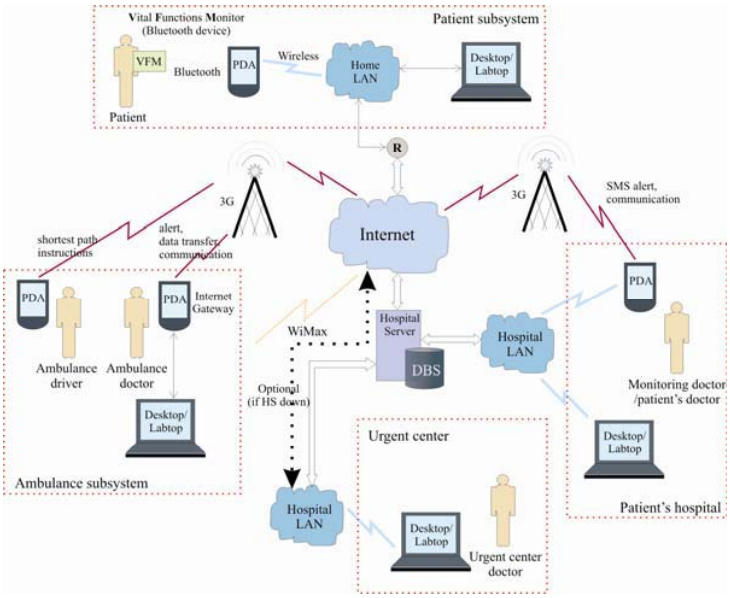


Fig. 2. Architecture of the CHS system

The backbone process of this system is the automatic emergency signal flow. It is depicted in Fig. 3, with the stages marked with grey circles. This signal is raised when the Emergency Condition Evaluator (ECE) component in the HS evaluates some abnormal values. The monitoring doctor receives the alert signal, one for potential problem and one for emergency situation, and has patient's VF displayed immediately on the computer screen. He decides weather to forward the signal to the patient's doctor: GP or specialist who was responsible for the patient while treated in the hospital. If the alert is of first emergency level (marked red) the monitoring doctor sends alert signal to the Ambulance Dispatching Centre. An ambulance vehicle is activated according to the address of the patient. Another signal is sent to the UC according to the sectorization. When emergency situation happens the patient's doctor calls the patient and after gaining some information (if patients is able to answer and communicate) he calls the UC to describe the patient's current condition and his clinical history. If the alert is of second emergency level, the alert signal is sent solely to patient's doctor. The patient can call the monitoring center in the same time when automatic alert signal is raised. In that scenario, the monitoring doctor tries to divert the call to the patient doctor. If patient's doctor is not available, then the monitoring doctor takes his role in the system.

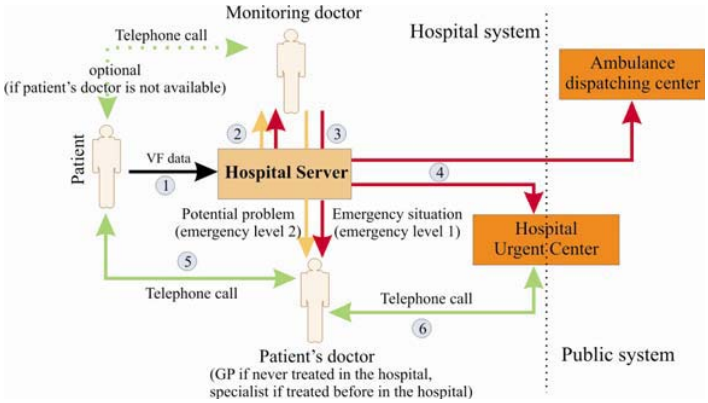


Fig. 3. Alert signal propagation diagram

3 The Heartbeat Tracker

The system (shown in Fig. 4 on the left) is comprised of ECG transmitter, vital signals sensors and Mobile ECG Application. The mobile application has two main modules: the ECG Processing Module - ECGPM and the Mobile Physician Module - MPM. The ECGPM consists of two parts: Pulse Extractor and Arrhythmia Recognition [3]. The MPM is a mobile application and is intended to be installed on smartphones, but not on smartwatches. Smartwatches only raise alert signal and display a short text message, because of their very limited resources.

The signal comes from the ECG transmitter and is then passed trough ECGPM, where pulses are extracted and tested against trained arrhythmia recognition machine. If an arrhythmia is recognized, an alert signal is raised and MPM is launched. Along with the data from the vital signals sensors and trough interaction with MPM the user gets valuable information of how to engage the heartbeat disorder.

The Mobile Physician – MP (Fig. 4 on the right) is intended to be the system's interface to the user and to suggest steps that user should follow until medical help arrives. Isolated ECG signal does not provide enough symptoms to precisely detect the syndrome, so there are additional input vital signals to the MP: blood pressure and oxygen saturation.

Trough a series of questions and answers, the user is guided to perform some actions that will most probably stabilize his/her condition or will just give the user instructions how to remain calm while help arrives in case of more severe disorder. For instance, alarm can be raised for a person that suffers arrhythmia from two reasons: entering an arrhythmia cycle or having a heart attack. The MP can tell the user he/she is entering an arrhythmia cycle and suggest to take a pill and call the ambulance, or say that he/she is having a heart attack and should take nitroglycerin if available and call the ambulance. This is especially important when the user finds him/-herself outside urban area. In some cases the MP can provide steps for first aid to a second person that is accompanying the user or is just passing by. It is important to note that MP does not make any diagnosis on the spot – it only acts as a guide to provide first aid or steps to follow until medical help arrives or while the user is transported to a medical facility. Another important thing is that there is an emergency button that appears in the MP. By pressing it the user calls the ambulance.

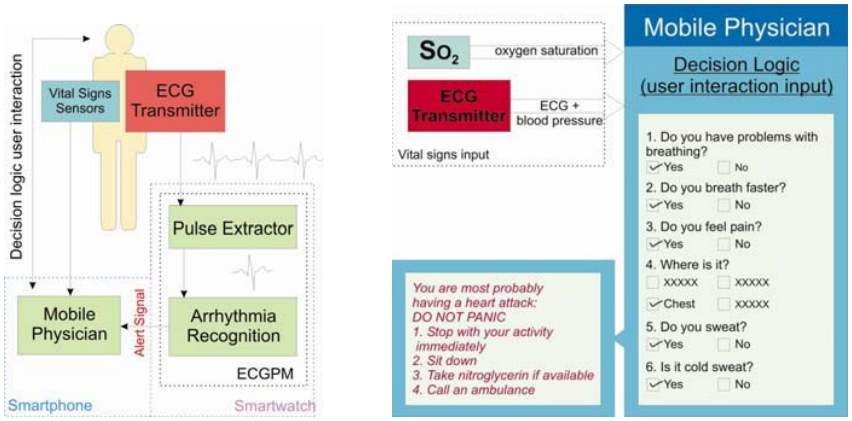


Fig. 4. General architecture of the system (left) and MPM workflow with user interaction (right)

4 Internet Medical Consultant as the point of integration

4.1 The Internet Medical Consultant

The Internet Medical Consultant system is presented in Fig. 5. The main process is demand for consultation. A user can perform that action asynchronously, through message communication (very similar to e-mailing), or synchronously by chatting, VoIP, remote collaboration, teleconferencing, triangular consultation, etc. Other features in the system are designed to support and enhance the quality of knowledge sharing [1].

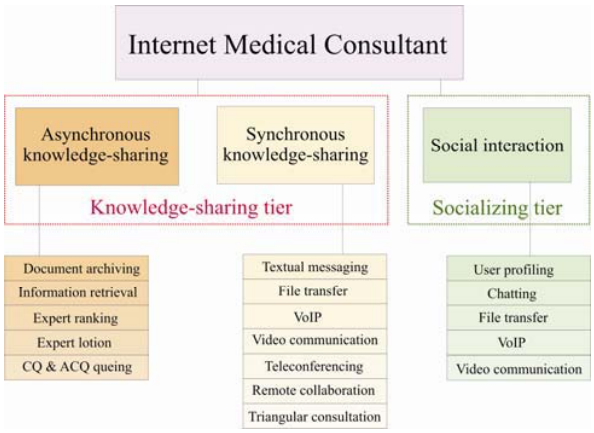


Fig. 5. Logical and functional diagram of the IMC system

The basic mode of knowledge sharing is via concept of messaging. The doctor who requires help creates a CQ and sends it either to the system, or to a list of specified doctors. When a CQ is sent to the system, the location of an expert is determined by an algorithm in the system. Otherwise messages are simply delivered to the doctors listed in the recipient list. The CQ is bound to its answers and formulates the basic knowledge structure in the system. The whole knowledge traffic is archived as such, so that later, users can search through it.

The IMC architecture is depicted in Fig 6. The system is represented by two applications, one for personal computers and another for mobile devices. The personal computer application is designed to be used when more thorough consumption of the systems features is required. The functions like creating a CQ answer, creating files to be attached to the CQ, searching for a certain document, organizing new knowledge gained from answers to a CQ, performing remote collaboration, triangular consultation or teleconferencing are designed to allow more detailed and analytical approach when used on personal computer. The Mobile Application Layer (MAL) is comprised of two modules, one that is a web-based, the Mobile Web Application (MWA), and another that resides on the device itself, the Mobile Device Application (MDA) [3]. The web application carries the real-time remote-collaboration and the

main load of end user's asynchronous communication while outside office, whereas the device application is mainly intended for offline usage and acts like an offline support for the web-application. Regarding the device application capability of modifying contents, there is a possibility for synchronizing the device data with the desktop data, other mobile device data, server data, and interfacing to the hospital DBS.

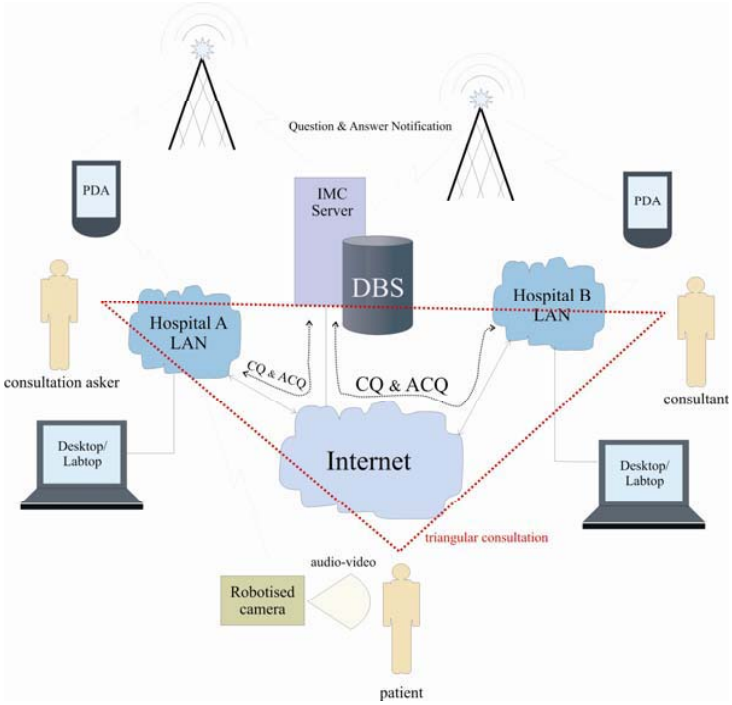


Fig. 6. Architecture of the IMC system

The MWA is designed for viewing and posting CQs, answers to the posted CQs and establishing dialogs for any additional discussions or ambiguities that may occur in the communication. Users can store a consultation (with its answers, dialogs and attached files) in the desired format by selecting which portion of information they find useful in the future. For the purpose of more reliable and time-efficient knowledge sharing there is a sub-module of the MWA that delivers remote collaboration upon an image that can be modified in real-time and viewed by both peers. The MDA is designed to support the web-application. Because the knowledge that is conserved in the consultations has to be available to the physicians at any time, we can't rely solely on the Internet connection or the reliability of the server. Additionally, the utilization of the application mostly relies on browsing through the answers of consultation questions which is not a connectivity inherent problem. Therefore we can afford an offline mode of the application.

4.2 The point of integration

The IMC system looks in the future of modern consultations between physicians. The consultation should be rich with information and consist not only of text and images [1], but be complemented with data obtained from monitoring devices and systems, hospital files (test results, images, segments of EMRs or full EMRs) and specially designed body networks. Teleconference should also be always considered in the notion of modern consultations. In addition to this problem customized remote collaboration tools should be designed to eliminate ambiguities and improve communication. The IMC has all of this in mind and has a remote collaboration module. The remote collaboration consists of three main services: VoIP, Chat, Real-time Image Modification – RIM (two peers) [3]. While the first two services are very popular and spread trough most social networks, the third service brings a new line in the process of mobile remote collaboration. It allows both peers involved in the Remote Collaboration Session to modify an image and see the modifications in-real time. In the prototype we can draw only a rectangle and text to the image, but this can be analogously be broaden to a variety of shapes like lines, curves, circles, etc.

The point of integration is to provide IMC with data from CHS and HT. The systems will be integrated through SOA oriented bonding middleware of web-service network [9, 10] shown in Fig. 7. The systems should support data transfer, streaming and additional functionalities introduced by the integration of the systems. That way the physician that uses IMC can send a CQ along with recorded stream of data from CHS and HT, if performing an asynchronous consultation, or redirect the consultant to the real-time streaming of the data from each of the systems in the case of synchronized consultation.

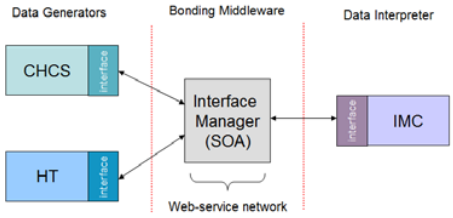


Fig. 7. Introduction of bonding middleware for integrating three initially independent medical systems: CHS, HT and IMC

Regarding the HT system, the Mobile Physician could be replaced with a real physician, which would have insight in the condition of the HT user, so the direction one receives will be more reliable. The Mobile Physician will remain as an offline solution in areas not covered by Internet. The improvement in the CHS due to the integration with IMC is towards enabling the urgent centre physicians to consult colleagues from other institution in real-time about the complicated and unfamiliar condition of the incoming patient. Considering the remote collaboration process (Fig. 8), it could be performed on the data stream obtained from CHS and HT during the real-time consultation, and not only on static data like images, test results or data stream

records. These are valuable improvements that can lead to better lifesaving rate in critical conditions and on-time diagnosis of potential problems.

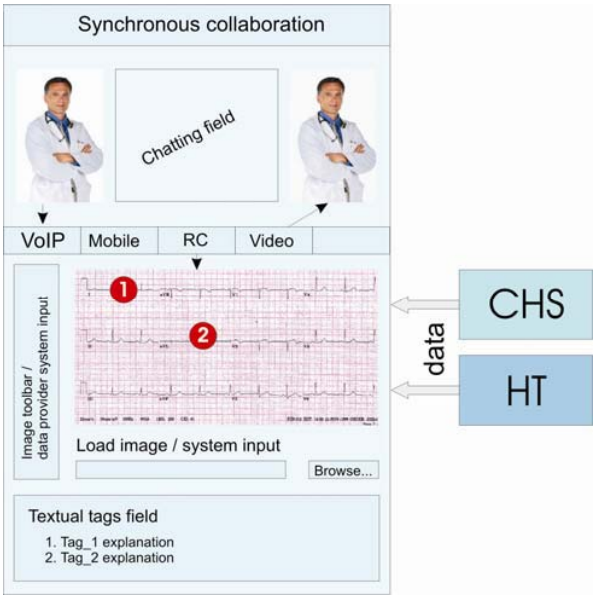


Fig. 8. The remote collaboration module with its functionalities: chat, VoIP and mobile calls, video streaming, modifying images in real-time, textual tagging of images, choosing system input for data streaming, and discussing real-time and recorded data streams.

5 Conclusion and future work

In this paper we showed an overview of how modern medical systems should be observed. We suggest that when designing novel medical systems, their integration is what attention must be paid to. Considering existing systems, an effort should be made to integrate them in a functional concept to improve some properties of each or some of the systems. We showed how three initially independent systems could be synergized to improve the overall performance of each of them. By providing valuable data from Critical Homecare System and the Heartbeat Tracking system to the Internet Medical Consultant, the quality of consultation and remote collaboration improves and consequently do the functionalities introduced by the integration of the systems or previously existed ones. Data rich consultations contribute for more thorough analysis and thus more confident and reliable decisions which are crucial in situations of emergency.

Integration of medical systems is very complex problem and in future work we will investigate how to do that with efficient network of web-services and how to plan the integration oriented design in advance. We will focus on redesigning the

functionalities of the remote collaboration module, so that IMC benefits the most from the integration of the systems. Careful estimation of feasible functionality modifications of separate systems should also be considered.

References

1. Nakic, D., Loskovska, S.: Internet medical consultant — A knowledge-sharing system: Information Technology Interfaces, 2009. Proceedings of the ITI 2009 31st International Conference (2009) 79 - 86
2. Nakic, D., Loskovska, S.: Knowledge Sharing Mobile Application Layer for the Internet Medical Consultant: Information Technology Interfaces, 2010. Proceedings of the ITI 2010 32nd International Conference, (2010) 243-248
3. Nakic, D., Madzarov, G., Djordjevic, D.: Heartbeat Tracking Application for Mobile Devices – Arrhythmia Recognition Module: Information Technology Interfaces, 2010. Proceedings of the ITI 2010 32nd International Conference, (2010) 585-590
4. Al Khaib, I., Poletti, F., Bertozzi, D., Benini, L., Bechara, M., Khalifeh, H., Jantsch, A., Nabiev, R.: A multiprocessor system-on-chip for real-time biomedical monitoring and analysis: ECG prototype architectural design space exploration. ACM Transactions on Design Automation of Electronic Systems (TODAES) (2005) 294-303
5. Koufi, V., Malamateniou, F., Vassilacopoulos, G.: A medical diagnostic and treatment advice system for the provision of home care: 1st international conference on PErvasive Technologies Related to Assistive Environments (2008)
6. Carl, T., Dajani, L.: The future of homecare systems in the context of the ubiquitous web and its related mobile technologies: 1st international conference on PErvasive Technologies Related to Assistive Environments (2008)
7. Larsen, S. B., Bardam, J., E.: Competence articulation: alignment of competences and responsibilities in synchronous telemedical collaboration. Conference on Human Factors in Computing Systems (2008) 553-562
8. Blanchet D. K.: Remote patient monitoring: Medical Conectivity, Telemedicine and e-Health (March, 2008)
9. Beyer M., Kuhn A. Klaus., Meiler C., Jablonski S., Lenz R. : Towards a flexible, process-oriented IT architecture for an integrated healthcare network: Symposium on Applied Computing . Proceedings of the 2004 ACM symposium on Applied computing (2004) 264-271
10. Dalsgaard, E., Kjelstrom, K., Riis, J.: A federation of web services for Danish health care: IDtrust. Proceedings of the 7th symposium on Identity and trust on the Internet (2008) 112-121

Framework for Software-Intensive Ingest System: One Behavioural Description

Aleksandar Spasić[1] and Dragan Janković[2]

[1] College of Professional Studies for Pre-School Teachers, Ćirila i Metodija 29, 18300 Pirot, Serbia
aspasic@string.co.rs
[2] Faculty of Electronic Engineering, University of Niš, Aleksandra Medvedeva 14, 18000 Niš, Serbia
dragan.jankovic@elfak.ni.ac.rs

Abstract. The new model of television production is based upon: digital formats, the centralized management of media content and associated metadata, non-linear assembly of media elements, high-speed networks, format agnostic distribution and automated processes. An effective software-intensive production system should allow incoming material – text, audio, video and associated metadata – to be available to all users on the system as soon as it arrives at the production facility. The framework for ingest system's model in problem space is presented in this paper. During the analyses of new content ingest workflow, behavioural description is modelled by the Use Case diagram.

Keywords: Software Intensive System, Content Management, Television Production, Ingest System, Model of Problem Space (MOPS)

1 Introduction

1.1 Background

The technological changes dramatically altered the way in which television programme is produced and distributed. Small and medium-sized broadcasters had opportunity now to produce the multimedia content in the quality which is comparable with content produced by big players in broadcasting market. Importance of the independent networked local and regional television broadcasters is described in [1] and [2].

Media asset management (MAM) is consisted of the processes and systems responsible for the identification, capture, digitization, storage, cataloguing, retrieval, use and re-use of multimedia materials. Essence is the material itself e.g. video, audio, graphic, still image, text. Metadata is any information related to essence or describing essence, but not the essence itself. Metadata holds the key

to the usefulness of the essence to the user. It can permit media assets to cross over between multimedia applications and to contribute positively in each and every application environment. Content is the combination of essence and metadata. Asset is the combination of content and rights.

Content management (CM) is a system that includes all the applications and tools needed to manage a piece of essence consistently together with its related metadata. Its main goals are maintaining the relations between essence and associated metadata, and maintaining the relations between different versions of the same content.

The pace of change in the television industry, as a result of the convergence of traditional broadcasting and information technology, is quickening dramatically. Searching a solution, programme makers quickly discover that there is no existing model within the broadcast and production industry to which they can turn. Today's "off-the-shelf" digital production solutions rarely do everything needed by the typical media enterprise. Ultimately, what is needed is a complete re-thinking of the way technology can be applied to the art and business of program making.

Business process definition and analyses of production stages are shown in [3]. Behavioural description of ingesting as a most important process in acquisition stage of content production is presented in this paper.

1.2 Traditional vs. Modern Television Production

Traditional production is based on analogue processes and expensive analogue components without computers in the technological system, minimal using of the computers in the other parts of workflow and tape-based technological chain. Media is edited and stored on magnetic tape, production tasks are strictly defined, and workflows are largely linear in nature. Nonetheless, the workflows are still primarily linear and tape-based. The simplified chain of production is shown on Figure 1. [4].

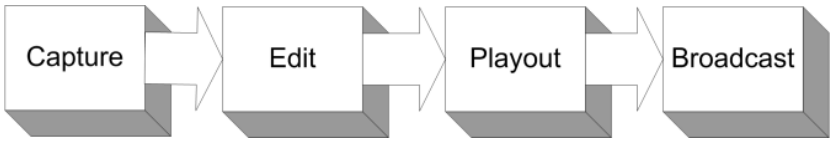


Fig. 1. Simplified chain of traditional production

A production process can be considered in a number of stages. A program's life begins with scheduling, research and planning. Video shots, audio clips and other program items are created during the acquisition stage. In this stage, archived material is also checked for immediate suitability for re-purposing. Next stage is editing, when shots, clips, animations and assembled items are put in the order. After editing, program is sent to delivery point for transmission or play out. Finally, program is archived on tapes.

Traditional tape based production systems have been very successful over the past decades. However they have all suffered from a number of fundamental limitations. Tape based workflows are linear in nature - one task has to be completed before the next can begin. As the essence passes from the idea to the finished program, it passes through the different stages along a chain in which the quantity and complexity of the essence increase and decrease from step to step.

An effective modern digital production system should allow incoming material – text, audio, video and associated metadata – to be available to all users on the system as soon as it arrives at the production facility. This media should be available not only on the work group level but across the entire enterprise, and to users operating remotely.

2 The Method

A model, by its very nature, is an abstraction of the reality. The modeller, depending on his/her needs, keeps parts of the reality that are important to him/her in a particular situation and leaves out others which may be considered less important. Successful modelling needs to consider the areas in which modelling needs to take place. These modelling spaces have been formally considered and discussed by Unhelkar in [5]. The three distinct yet related modelling spaces are defined: problem, solution and background. These divisions provide a much more robust approach to modelling, as they segregate the models based on their purpose, primarily whether the model is created to understand the problem, to provide a solution to the problem, or to influence both of these purposes from the background, based on organizational constraints, and need to reuse components and services.

In UML projects, model of problem space (MOPS) deals with creating an understanding of the problem, primarily the problem that the potential user of the system is facing. While usually it is the business problem that is being described, even a technical problem can be described at the user level in MOPS. In any case, the problem space deals with all the work that takes place in understanding the problem in the context of the software system, before any solution or development is attempted.

Typical activities that take place in MOPS include documenting and understanding the requirements, analyzing requirements, investigating the problem in detail, and perhaps optional prototyping and understanding the flow of the process within the business. Thus the problem space would focus entirely on what is happening with the business or the user [6].

As a description of what is happening with the user or the business, the problem space will need the UML diagrams that help the modeller understand the problem without going into technological detail. The UML diagram that helps express what is expected of the system, rather than how the system will be implemented, is Use Case diagram. Use Case diagrams provide the overall view and scope of

functionality. The use cases within these diagrams contain the behavioural (or functional) description of the system.

3 Basic production stages

Basic production stages are defined here as follows: development, planning, acquisition, processing, control, archiving and publication.

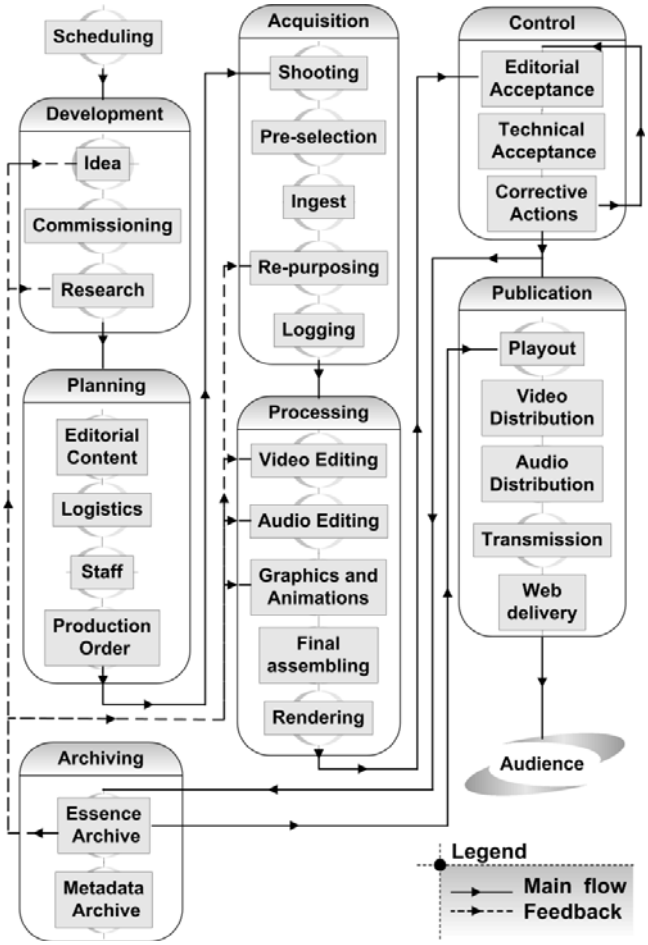


Fig. 2. Program Production Workflow

These stages are shown in Figure 2, as well as what the production processes are consisted of. At each step in the production workflow we can collect, and possibly re-use the metadata.

4 Ingestion Process

4.1 Ingestion Workflow

During the acquisition stage, video shoots, audio clips and other programme items are created, pre-selected, ingested into production system and logged.

The importance of the ingestion process is emphasized by Airola, Boch and Dimino in [7] and noticed that *"crucial problem of Content Management Systems (CMS) is constituted by the ingestion of new content. As we cannot realistically expect that all the aspects of a production/archive environment are under the rules of a CMS, we need to set up gateways through which the content must pass when migrating from a non-managed environment to a CMS. The role of these gateways, that we call Ingestion Systems, is that of collecting and organizing as many relevant information (metadata) on the item as possible... ".*

Ingest is the first stage to efficiently transfer the content to the television production infrastructure. During the ingest we take all the content collected during a shoot, as well as new metadata, and transfer it into the production environment. We assume that the planning and commissioning metadata is already in the system. More metadata can be generated at ingest and this can either be directly entered, for example by an operator marking technically poor sections, or regions for special processing, or it can be extracted automatically.

Simplified ingestion workflow is shown in Figure 4.

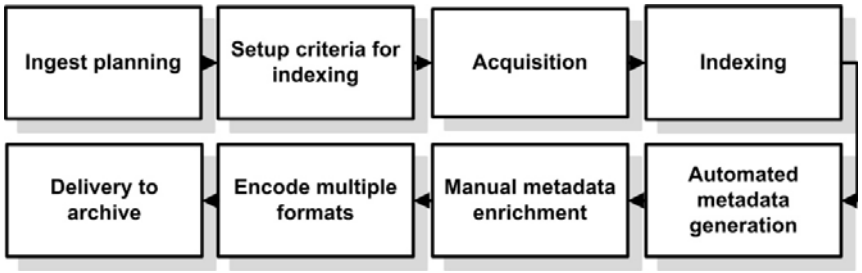


Fig. 2. Ingest Workflow

Ingestion can be considered in terms of two processes, or fundamental tasks:

- Content acquisition and optimization, and
- Content description and referencing.

Content acquisition and optimization assume capturing the audio-video essence and content compression. The obvious capture device is the camera, but equally, sound effects, graphics, stills, captions and music may all be added. Typically, users of Content Management System will want to utilize high resolution master file (MPEG2 encoded) which contains the content in professional broadcasting quality, as well as low resolution proxies of same content for searching and previewing archived material or for web delivery. Ingest system should provide automatic generation of high and low resolution content representations.

During the content optimization key frames should be extracted and recorded. Key frames are valuable for providing asset management solutions with representative images for browsing video, as well as for making edit decisions. Key frames should be extracted and converted to JPEG images based on scene changes or predefined time intervals.

At all points in capture there is an opportunity for metadata collection. Some of the metadata, like producer's comments and annotation, can only be captured by direct entry at the time of shooting. The metadata at this point in the chain should be viewed as 'portable', carried along with the essence. Also, technical metadata, like time code information or light information can be captured from camera. Manual adding of metadata for description and indexing of content, however, should significantly enrich the content and support the asset management applications.

4.2 Modelling Behavioural Description of Ingest Process

The main objective of a behavioural description is to visualize how the user (represented by the actor) will interact with and use the system. This is done by showing the actor associating with one or more use cases and, additionally, by drawing many use case diagrams.

Main actors in problem space of ingest process are producer, ingest operator, essence gathering crew (cameraman, sound recorder) and ingest automated system.

Use cases important for modelling in problem space of ingest process are as follows: Ingest planning, Setup criteria for indexing, Start acquisition, Automated metadata generation, Manual metadata enrichment, Indexing, Encoding multiple formats and Delivery to Archive. Use Case diagram is shown in Figure 5.

4.2.1 Use Case *Ingest planning*

Short Description: The producer plans ingesting activities.

Actors: Producer, often head of production crew or technical manager.

Pre-Conditions: Programme schedule is produced.

Post-Conditions: Plan for ingest activities is done.

Main Flow: (1) Producer suggests an ingest plan in accordance with daily program schedule. (2) Use Case terminates.

4.2.2 Use Case *Setup criteria for indexing*

Short Description: Ingest operator setups criteria for indexing.
Actors: Ingest operator.
Pre-Conditions: Ingest plan is made and producer defined ingesting criteria.
Post-Conditions: Indexing parameters are setup.
Main Flow: (1) Ingest operator setups parameter list for indexing (key frames, time code intervals, external event, spot detection, scene changes, closed captions etc.).(2) Particular parameter is set-up. (3) (2) is repeating until all parameters are set up. (4) Use Case terminates.

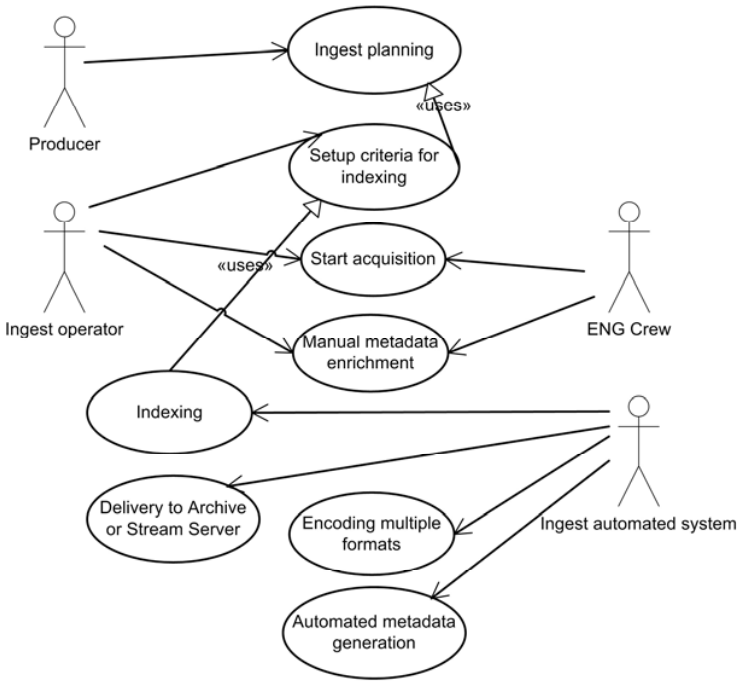


Fig. 3. Use Case diagram of Ingest Process

4.2.3 Use Case *Start Acquisition*

Short Description: Video shoots, audio clips and other essence items, as well as attached metadata are extracted from the continuous acquisition bin, such as cameras, sound recorders, digital tape players, different editing workstations, servers etc.
Actors: Ingest operator and essence gathering crew members.
Pre-Conditions: Video essence (footage) is shoot, audio essence is recorded, other program items are created and pre-selected.

Post-Conditions: All essence materials, as well as related metadata, are ingested into production system.

Main Flow: (1) Ingest operator starts acquiring the raw material or previously produced essence from cameras or other sources. (2) During ingest all the content collected during a shoot, recording and repurposing, as well as new metadata is taken and transferred into the production environment. (3) Ingest operator reviews what he/she has, and marks down its possible use. Use Case terminates.

4.2.4 Use Case *Indexing*

Short Description: Ingest system extracts a number of key attributes from the source essence and converts them to metadata.

Actors: Ingest automated system.

Pre-Conditions: Criteria for indexing are defined and parameters are setup.

Post-Conditions: Key attributes are extracted and related metadata are produced.

Main Flow: (1) System analyzes the acquired essence in accordance with parameters which are previously set-up. (2) Metadata generator module produces related metadata. Use Case terminates.

4.2.5 Use Case *Automated metadata generations*

Short Description: System generates metadata from the acquired essence files.

Actors: Ingest automated system.

Pre-Conditions: Essence material, as well as files with related metadata, are ingested into production system.

Post-Conditions: Metadata are stored in database.

Main Flow: (1) System search for metadata files accompanying acquired essence. (2) System checks the metadata format and if metadata format is in accordance with systems metadata formats, metadata are stored in database. (3) If metadata format does not confirm, corrective actions must be undertaken [A1]. (4) Steps (1) and (2) are repeated until all metadata is generated and Use Case terminates.

Alternate Flow: (A1) No need for corrections. Metadata are approved and stored in data base.

4.2.6 Use Case *Manual metadata enrichment*

Short Description: Automatic extracted metadata can be validated. New metadata (descriptions, business information etc.) can be added.

Actors: Ingest operator, Essence gathering crew member.

Pre-Conditions: Essence material, as well as files with related metadata, including indexed and generated metadata are ingested into production system.

Post-Conditions: Stored essence, as well as related metadata.

Main Flow: (1) Ingest operator validates previously generated indexes and metadata. (2) Ingest operator add new descriptions of essence. Use Case terminates.

4.2.7 Use Case *Encoding multiple formats*

Short Description: Create high-resolution master essence file and low-resolution proxy file.

Actors: Ingest automated system.

Pre-Conditions: Essence acquired, metadata generated and stored in database..

Post-Conditions: Essence encoded in hi-res version and stored in archive. Essence encoded in low-res version and delivered to stream server.

Main Flow: (1) In accordance with production format essence is encoded in high quality broadcast format (usually MPEG2). (2) Essence is encoded in several different versions of low-res files. Use Case terminates.

4.2.8 Use Case *Delivery to Archive and Stream Server*

Short Description: System stores essence in deep archive and send low-res proxies to stream server.

Actors: Ingest automated system.

Pre-Conditions: Essence is encoded in hi-res version. Essence is encoded in low-res version.

Post-Conditions: Essence is stored in archive. Essence is delivered to stream server.

Main Flow: (1) System stores essence in deep archive. (2) System deliver low-res proxies to stream server which serves low-res proxies for searching, previewing, non-linear editing etc. Use Case terminates.

5 Conclusions

Video and audio compression methods, server technology and digital networking are all making a big impact on television production, post-production and distribution.

Management concepts can be applied to a wide variety of applications related to the generation and to the use of audiovisual materials, also outside the broadcast world. Therefore, systems must be conceived with the necessary degree of flexibility, required to tailor software and hardware infrastructure and functionalities, to the performance demanded by specific applications without incurring into bottlenecks of unjustified investments.

As content is one of the most valuable assets for broadcasting companies, ingesting, archiving, accessing, managing, delivering and security of content assets become basic requirements in the everyday life of multimedia producers

and providers; at the same time, it becomes an important way how the company structures its facilities, the processes involved and how it chooses the technologies that best adhere to the purpose related to content handling.

The primary aim of this paper was to describe ingest system as one of the main areas in a production environment and to summarize the essence, metadata and control flow, as well as the main processes involved in a typical ingesting of television content.

The first step in modelling of software intensive ingest system, modelling of problem space, is presented here. Use Case diagram is used as an ultimate tool for behavioural description of the ingest system.

The challenge for the future is to make complete business analysis of problem space using the use case, class, activity, state machine and sequence diagrams. Also, the model of solution space, as well as the model of background space should follow the current work presented in this paper.

References

1. Spasic A., Nesic M.: Informing Citizens in a Highly Restrictive Environment Using Low-Budget Multimedia Communications: A Serbian Case Study, (2005), Informing Science Journal, Volume 8, pages 245-262., Informing Science Institute, Santa Clara, California, USA, ISSN: 1521-4672, http://inform.nu/Articles/Vol8/v8p245-262Spasic.pdf
2. Gill G., Cohen E. (Eds): Foundations of Informing Science: 1998-2008, Part IV: Applied Informing Science, Chapter 17: Aleksandar Spasic, Miloje Nesic: "Informing Citizens in a Highly Restrictive Environment Using Low-Budget Multimedia Communications: A Serbian Case Study", pp. 577-617, 2009, ISBN13: 978-1-932886-15-3, ISBN10: 193288615X, Informing Science Institute, Santa Clara, California, USA, (2009)
3. Spasic A.: Business Analysis of Software-Intensive Television Production: Modelling the Content Production Workflow, Serbian Journal of Management 1 (2) (2006), pp. 17-32, University of Belgrade, Technical Faculty in Bor, ISSN:1452-4864, http://www.sjm06.com/1_2_2006_full text/7 ASpasic OK.pdf
4. Hunter J.R, Lau H., White D.J.: Enhanced television service development, International Broadcasting Convention (IBC 2000), Amsterdam, IEE Conference Publication, (2000) http://www.bbc.co.uk/rd/pubs/papers/pdffiles/ibc00jh.pdf
5. Unhelkar B.: Verification and Validation for Quality of UML 2.0 Models. John Wiley & Sons, Inc. Hoboken, New Jersey (2005)
6. O'Doherty M.: Object-Oriented Analysis and Design: Understanding System Development with UML 2.0. John Wiley & Sons Ltd., Chichester, West Sussex (2005)
7. Airola D., Boch L., Dimino G., Automated Ingestion of Audiovisual Content, (2002) http://www.broadcastpapers.com/asset/IBCRAIAutoIngestAVContent.pdf

One realization of the attribute inheritance mechanism in specific cad/cam applications

Dejan S. Aleksić[1], Dragan S. Janković[2], Leonid V. Stoimenov[2]

[1]Faculty of Sciences and Mathematics, University of Nis, Visegradska 3, 18000 Nis, Serbia
dejan_aleksic@yahoo.com
[2]Faculty of Electronics, University of Nis, Aleksandra Medvedeva 14, 18000 Nis, Serbia,
{dragan.jankovic,leonid.stoimenov}@elfak.ni.ac.rs

Abstract. This paper describes one realization of the attribute inheritance mechanism in specific CAD/CAM software package supporting the design and manufacturing of the facade carpentry. The facade carpentry is made out of a number of profiles and items together with the corresponding parts needed to join and assemble them known as profile system. Each entity in profile system can be described using the arbitrary number of attributes, and each attribute has its type and the momentary value from the list of possible values. Each entity in the profile system is described by arbitrary number of attributes and each attribute has its own type and momentary value from the list of possible values. Frequent repetition of the same attributes has been noticed within the hierarchically organized data which impedes the change of value of typical attributes resulting in extending the necessary space for storing data. The suggested attribute inheritance mechanism should solve these problems.

Keywords: attribute inheritance, CAD/CAM database, object-relational mapping, OODB.

1 Introduction

Specialized software applications in CAD/CAM field usually serve as a solution for only a narrow specter of problems. To justify its existence, they must be maximally adjusted to the user and optimized for the problem they are used for. On the other hand, narrowing its area of usage does not necessarily mean that such applications do not need general enough solutions so that they can be applied to all other/future uses in that area.

Throughout this work, we will illustrate one object-oriented database in a CAD/CAM software package for the project support and the manufacture of the facade carpentry.

During the process of the facade carpentry creation a cluster of profiles and fillings are used, together with the associated parts for their connection and assembly of so called profile system. All the profile systems must follow certain general rules and standards, but, on the other hand, each one of them has some specific characteristics

which make them unique. It is almost impossible to foresee all the new characteristics which some future profile systems can bring.

The choice of storing all the data for the description of profile systems in a relational database has brought certain limitations related to the fixed number of fields inside tables, as well as necessary definition of the data types. On the other hand, it was necessary to have freedom related to the number and types of data to describe the profile system efficiently. The problem arises from the fact that it is impossible to foresee all the new characteristics which some future profile systems can bring.

The solution we have applied in this case is based on the transformation of relational data base into their object model in the memory [1], [2]. Thus, a special group of objects within the same application is created which provides us with a flexibility concerning numbers and types of the data for describing entities but with a realized mechanism for the data type checking. Each entity can be described with a random number of attributes, whereas each attribute has its own type and a current value from the list of potential values. Furthermore, the usage of these classes makes it possible to hierarchically organize the data which describe a profile system in multiple fixed levels.

However, the nature of the problem itself causes a high recurrence of one and the same attributes within the hierarchically organized data. This fact significantly makes the modification of typical attribute values along the whole hierarchy difficult and causes the increasing in space needed for the storage of the data about some profile system. It has been proved that the suggested inheritance mechanism of attribute values, described in the paper, resolves these problems and this has, also, been confirmed in practice.

2 Describing a system in a base using attributes

With the purpose of describing the characteristics and rules governing a profile system more efficiently, a special set of objects has been created within the application, also known as *AttrAPI* [5]. This set of objects provides us with a flexibility concerning numbers and types of the data for describing entities but with a realized mechanism for the data type checking. The basis for this approach is the use of so called attributes i.e. the base class tAttribute along with its inherited class sets. As it was already mentioned, each entity can be described with a random number of attributes and each attribute has its type, current value from the list of possible values. This approach brought a number of benefits regarding an easy execution of not only small, but also large enough differences in profile system description, multilanguage terms support within the base, easy upgrading of the existing database version, automatic data copy from the database into the memory or disk structures and vice versa, quick and easy writing and reading of data about projects on/off the disk, possibility of execution of an automatic written project data upgrade in accordance with data changes in value, as well as in number and type, an accomplishment of all advantages which object-oriented data access brings (data abstraction, succeeding, overload of attributes) over data located inside the base, memory or on disk [6], [6].

All the actions concerning writing, modification and reading attributes inside the base are executed inside *tAttributes* class and *tAttributeList* [8], [9].

3 Realization of ATTr Api

Attr API itself contains greater number of classes, but the complete functionality of this API can be illustrated by its three main classes – *tAttribute*, *tAttributeList* and *tXalNode* with its inherited classes (Fig. 1). The basic functionality in work with the attributes is realized in the basic class *tAttribute*. More precisely, the class *tAttribute*, as the basic class, does not contain much of functionality, it only defines basic variables and methods (virtual and abstract) which will be copied and realized in some of the inherited classes (*tAttrString, tAttrInteger, tAttrReal, tAttrBoolean*). Initially, the backup for the four basic types was realized, but if required, it is possible to realize the backup for some random type of data through the new inherited *tAttribute* class.

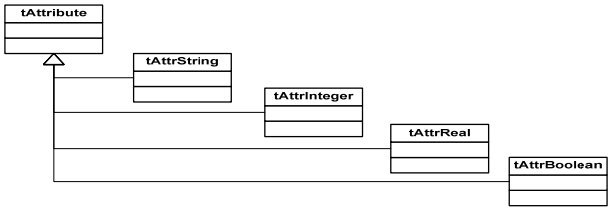
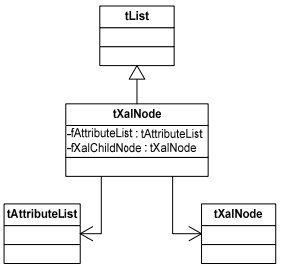


Fig. 1. Inherites tree of base class tAttribute

The second basic class is *tAttributeList* which realizes the complete storage and manipulation of the attribute list, or more precisely, the instance of the basic class *tAttribute, i.e.* its inherited classes. The number of the attributes, their type and value can be completely random.

The third key class of Attr API is *tXalNode* which is *a* inherited class of the class tList. The formation and manipulation with the basic hierarchic structures are its primary task. Each instance of this class represents the node of the hierarchic structure. Typically, it is the structure of the tree where each node of that tree has its own list of attributes which describe its condition (Fig. 2).

The Attr API classes enable us to hierarchically organize and edit data stored in the memory as a set of objects- instances of those classes. In this case, the data access is very fast, but there is still a need for permanent storing of the data. The process ofwriting/reading the hierarchic structure of the Attr API objects in/from the basis is completely realized within basic classes and it is entirely transparent for the user.




```
tXalNode = class (tList)
 fAttributeList:tAttributeList;
 fXalChildNode:tXalNode;
. . .
end;
```

Fig. 2. *tXalNode* - crucial class of *Attr API*

All the data along with their hierarchic structure no matter how complex it is, and no matter the type and attributes in it are being placed inside the three tables within relational data base (Fig. 3).

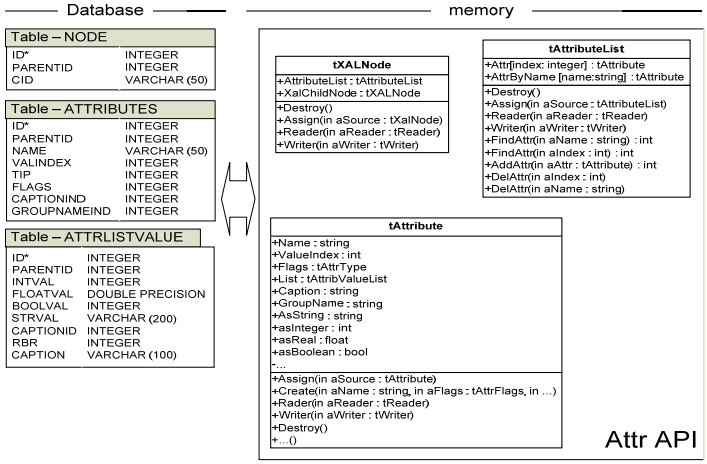


Fig. 3. Relationship between data in relation database and memory (in class of *Attr API*)

Furthermore, if various hierarchic structures are defined within an application (using the tXalNode class) they are all being placed inside the same data base in three above mentioned tables. It is important to stress the possibility of easy creation of complex hierarchic structures using the Attr API classes. The user creates the new classes typically inheriting the tXalNode class. The additional functionality is realized by coping the methods from the parent class and creating the new methods. Data within the newly created classes are being kept in the attribute list (tAttributeList class), which makes the process of data reading/writing from/into the base completely transparent for the user, because it has already been realized in the basic classes. On the other hand, the user can create a random number of the attributes within the newly created class, define their desired type and set the needed value.

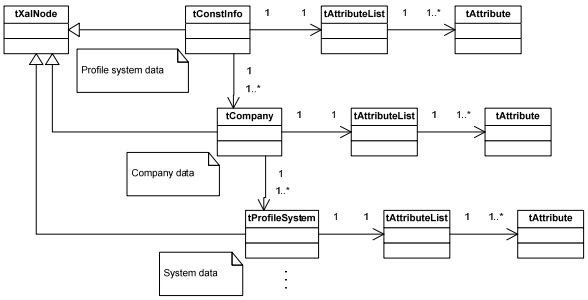


Fig. 4. the hierarchic structure created for the profile system description

Fig. 4 illustrates the hierarchic structure created for the profile system description data handling. It is common that data from this structure are being edited using the special program (ConstInfo Manager) and are being written in the data base. The main application mostly only reads the data from this kind of the structure, thus they are very often being cashed because of the critical speed of data reading.

Another example is the hierarchy structure which handles the project description data (Fig. 5). Depending on the main application configuration, this structure is being placed in the file on the disk (classic application) or inside the database (client/server application type). The complete functionality for both cases has already been realized within the basic Attr API classes, and it is up to user to decide where the data is going to be kept.

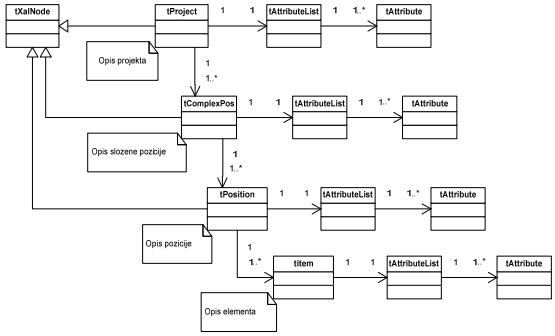


Fig. 5. The hierarchy structure which handles the project description data

In general, hierarchic structures of Attr API objects can be read/written in different forms and formats. The first of those is related to the usage of relational database, where the choice of the base itself does not affect the functionality. It is even possible within the same application to read from the relational database and for example, to write them into the relational database of another manufacturer after the processing. These operations are completely transparent for the Attr API class user. Because of the working speed, it is possible to update the data of only a certain part of the hierarchic structure from the memory into the base, without the need to write all the

structure data. The usage of database for storing hierarchic structure objects is another way according to which we can permanently save the data related to the AttrAPI objects. Dependent on the user, writing multiple data formats in the file is supported. (Fig. 6.).

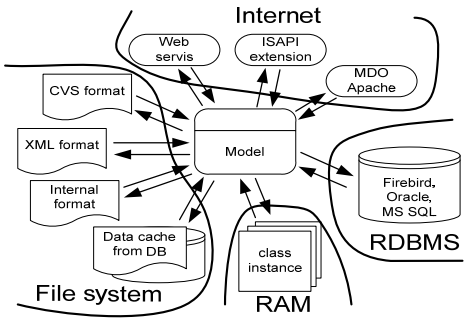


Fig. 6. Illustration write / load the class hierarchical structure of objects in various formats

Each of the formats has its own advantages and disadvantages considering the key parameters such as: the reading/writing speed (internal format) or simple exchange of the data with the other programs of the general usage (CVS format or XML). It is important to stress the possibility of automatic backup of the data in the file which is often used for cashing data on the client's side.

4 Realization of the attribute inheritance mechanism

Several practical examples have shown that it is possible to describe the profile system successfully by using the AttrAPI classes and methods. However, the nature of the problem itself causes a high recurrence of one and the same attributes within the hierarchically organized data. This fact significantly makes the modification of typical attribute values along the whole hierarchy difficult (Fig. 7) and causes the increase in space needed for the storage of the data about some profile system.

In order to solve this problem, the attribute value inheritance mechanism of the hierarchically organized data has been developed. The whole mechanism is realized within AttrAPI classes and it is completely transparent not only for the user but also for the programmer who uses AttrAPI classes for the application development.

The whole mechanism is founded on the fact that if the greater number of elements has the same attribute with the same value (see Fig. 7), than it is much more efficient to assign this attribute to their common parent, i.e. to one nod above the in the hierarchical data tree (see Fig. 8).

Of course, if the already mentioned attribute of some element has different value other than attributes of majority other elements, than this attribute remains in the list of the mentioned element (see Fig. 9 and the element named 390).

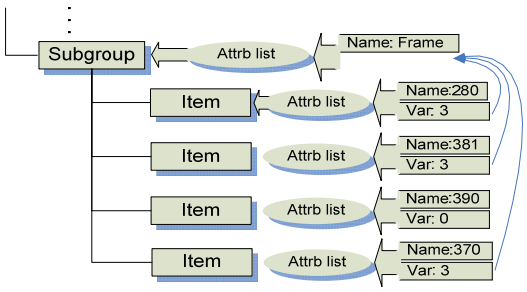


Fig. 7. Repeat var attribute in the hierarchical description of the system profile

This procedure obviously reduces needed space in a memory and makes it easier to change the values of the common attributes. Alternatively, the access to attributes of the elements became more complex, thus it had to be simplified and made transparent for the user/programmer.

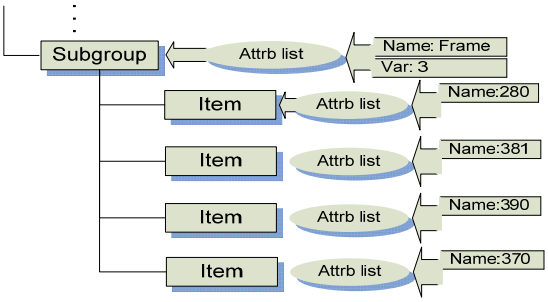


Fig. 8. Var layout attributes after the "transfer" to the list of parents attributes

As an inspiration, we used an inheritance mechanism in objective oriented languages, as expected, adjusted to our specific purpose. The functioning principle is very simple and it is thoroughly illustrated in Fig. 9.

If it is necessary to read the attribute value of some element, firstly its list of attributes has to be found. In case there is a requested attribute in the list, its value is passed on. When there is no requested attribute in the list, the search continues in the parent attribute list. If the desired attribute can be found in that list, its value is passed on to the element and back to the user. The procedure of “searching” for the desired attribute is conducted “along” the hierarchic tree until the attribute is found, and then its value is being passed on along the hierarchy to the element which had requested the attribute value.

In case when the procedure of “searching” ends with the element at the top of the hierarchy, and, if it has no defined requested attribute, the user is notified that such an attribute does not exist.

In this manner, we completely satisfied the demand for transparency, and the user/programmer has an impression that nothing substantially has changed in the writing method of the attribute value.

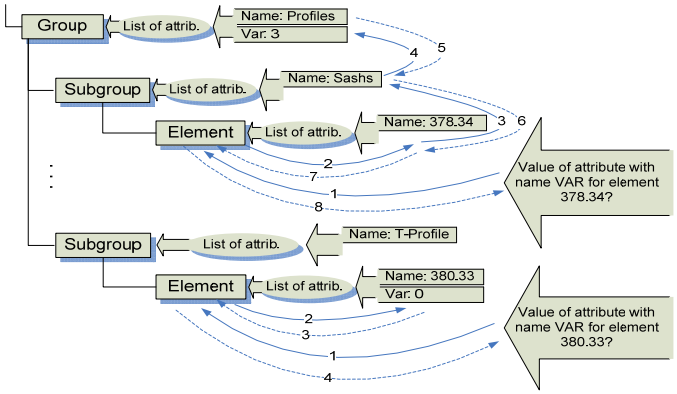


Fig. 9. An example of the functioning of the mechanism of inheritance of attributes of Var in the hierarchical description of the system profile

References

1. Q. Zhang, "Object-oriented database systems in manufacturing: selection and applications", Industrial Management & Data Systems, vol. 101, no. 3, pp. 97-105, 2001.
2. M. Blaha, W. Premerlani, H. Shen, "Converting OO Models Into RDBMS Schema" IEEE Software, vol. 11, no. 3, pp. 28-39, May/June 1994.
3. M.L. Brodie, B. Blanstein, U. Dayal, F. Manola, A. Rosenthal, "CAD/CAM Database Management", IEEE Database Engineering, Vol.7, No.2, pp. 12-20, June 1984.
4. Dejan S. Aleksic, "Softverski paket GenAl Xal za podršku u projektovanju i izradi Al i PVC stolarije", oko 250 instalacija.
5. Dejan S. Aleksic, Dragan S. Jankovic, "The use of scripts in a CAD/CAM database", The X International Conference on Information, Communication and Energy Systems and Technologies (ICEST 2009), June 25 - 27, 2009, Veliko Tarnovo, Bulgaria. http://eedu.tu-sofia.bg/icest/icest-program-01-06/csit-1.pdf
6. D. Maier, "Making database systems fast enough for CAD applications", Object-oriented concepts, databases, and applications, ACM Press, New York, NY, 1989.
7. Ying-Kuei Yang, "An enhanced data model for CAD/CAM database systems", Proceedings of the 25th ACM/IEEE conference on Design automation, pp. 263 - 268, 1988.
8. Dejan S. Aleksic, Dragan S. Jankovic, "The realization of the OO database in the specific CAD/CAM applications", The X International Conference on Information, Communication and Energy Systems and Technologies (ICEST 2009), June 25 - 27, 2009, Veliko Tarnovo, Bulgaria. http://eedu.tu-sofia.bg/icest/icest-program-01-06/csit-1.pdf
9. Dejan S. Aleksic, Dragan S. Jankovic, "One realization of multilingual support in specific CAD/CAM applications", International Conference ICT Innovations 2009, Sept. 28 - 30, 2009, Ohrid, R. Macedonia.

Feature selection in Face Recognition

Ivana Atanasova[1], Ljupco Jordanovski[2]

[1]European University, Faculty of Informatics, bul. Kliment Ohridski 68, Skopje, Republic of Macedonia
[2]Head of the Seismological Observatory, Skopje, Republic of Macedonia
ivana.atanasova@eurm.edu.mk

Abstract. Biometric technology is often used for verification and identification of individuals and objects. Face recognition requires analyzing a particular face and feature processing. Amongst the earliest used methods for face recognition is the feature - based method. The local feature method first includes face region detection from still digital image, then component detection and finally features extraction. This paper develops the idea of a system which identifies a face based on a minimum number of features. Predefined number of features was extracted from the basic face components and analyzed over a training face database with variable expression images. Results from the test dataset confirmed that the minimum 7-feature set is suitable for proper face identification. Research also showed that family-related entities can be recognized and this is something yet to be validated.

Keywords: face recognition, detection, feature-based, face components, family relations

1 Introduction

Facial recognition records the spatial geometry of distinguishing features of the face. Different vendors use different methods of face recognition, however, all focus on measures of key features of the face. The need for passive[1] identification has increased with the Fourth generation of technologies and devices including the so-called "smart environments" [1] that need to be aware of the people in their surroundings. Although there are more reliable and accurate identity authentication methods, such as iris identification and fingerprint, face recognition has a clandestine or covert capability (i.e. the subject does not necessarily know he/she has been observed) and does not require the presence of the participant who is identified. Face recognition for still images generally belongs to two groups of methods: Feature-based and Holistic, however there also exists novel category called Hybrid methods that combines the advantages of both. Most of earlier face recognition methods [2, 3-5] belong to the feature-based category. In these methods, usually a single image per person is used to extract geometrical measures such as the width of the head, the distances between the eyes, etc.

[1] Participant interaction not needed

This paper will present an idea for designing minimum feature set to be used to identify a face from small face workspace. Preprocessing techniques were used such as color conversion and noise reduction to increase the quality of the face images. Next, we determined which features are to be taken into account. This selection of features is next used to define the minimum number of features that is sufficient to identify a face. In future a fully automatic algorithm[2] will be designed that will use this feature set to identify faces.
This paper is organized in six parts. The second part presents the motivation and initiative for the idea. The third part presents early and latest, related work research and articles in our area of interest. The fourth section presents the component and feature selection as well as techniques used for processing them using the feature-based method [6]. The fifth section shows the results and analysis from the implemented research. The sixth section gives the preliminary conclusions made from analyzing the results from the test and future work.

2 Motivation

The face recognition technology has very well advanced and face recognition is one of the most current topics in Computer Vision and Pattern Recognition. Identification systems are no longer limited to verification of identity and surveillance. Numerous systems already use face recognition as an initial step in interpreting human actions, intentions, and behaviors as a central part in future smart environments.
The use of the same existing methods, dating from the early years of research in the field of face recognition based on local feature selection, hides the possibility to improve their disadvantages and performance. Re-election of new set of face features could contribute to more efficient and faster identification of still face images for applications in controlled environments.
The lack of research in terms of face classification and identification of family related faces in the database was an additional incentive to create our own database of images which will undoubtedly include family related entities. The system was intended to make proper identification in addition to successfully notice similarities between family-related entities in the database with the person being identified. In future a public database will be adjoined to the existing one and results will be obtained.

3 Related work in local feature based systems

Perhaps the most famous early example of a face recognition system is that of Teuvo Kohonen of the Helsinki University of Technology[7], who demonstrated that a simple neural net could perform face recognition for aligned and normalized images of faces. The network he employed computed a face description by approximating the

[2] Fully automatic algorithm – defined by the FERET program

eigenvectors of the image's autocorrelation matrix. These eigenvectors are now known as eigenfaces. However, Kohonen's system was not a practical success because it relied on precise alignment and normalization.

Local methods which use local facial features for face recognition are a relatively mature approach in the field with a long history [2, 3-5, 8-9].

In early 1990s, Brunelli and Poggio described a face recognition system, which can automatically extract 35 geometrical features to form a 35-dimensional vector for face representation, and the similarity matching is performed with a Bayes classifier [6]. A good recognition rate of 90% on a database of 47 subjects was reported.

Researcher Gaile G. Gordon in 1992 [8] carried out research in face recognition based on contribution of local features' depth and curvature. The performance of this system shows great promise. Even with a small feature set based on only eye corners, nose description, and head width, and very basic statistical methods for classification, recognition results were better than 70% in all cases. Under the best conditions tested there were only 6 errors in the ranking of the database over all target matches, and for all targets the best match correctly identified the target. These experiments clearly demonstrated the usefulness of the features developed in distinguishing among faces.

M. D. Malkauthekar and S. D. Sapkal [10] presented experimental analysis of classification of facial images. Facial images of different expressions and angles of two classes and three classes are used for classification. For two classes and three classes results are compared by Fisher Discriminant method and Euclidian distance is used as a similarity measure i.e. matching. The experimental results have been demonstrated that performance of Fisher Discriminant Analysis for three classes is same as the performance for two classes.

One of the latest researches in the field of face recognition with feature-based methods belongs to researchers Ramesha K., K. B. Raja, Venugopal K. R. and L. M Patnaik [11] who actually showed the importance of reviewing older feature-based methods. They proposed Feature Extraction based Face Recognition, Gender and Age Classification (FEBFRGAC) algorithm with only small training sets and it yields good results even with one image per person. They obtained the geometric features from a facial image based on the symmetry of human faces and the variation of gray levels; they located the positions of eyes, nose and mouth by applying the Canny edge operator. Finally, the gender was classified based on posteriori class probability and age was classified based on the shape and texture information using Artificial Neural Network.

4 Model design

The goal of our research is two-fold. The first is to define a minimum feature-set that will uniquely identify the faces in the database. The second is to detect whether family-related entities in the database are grouped similar by the system. The system to be designed is necessary to make face detection on portrait images, in a controlled environment[3], as well as extraction of predefined components and discriminatory

[3] One face per image, non-complex background and person's awareness of the photo being

features for the face being identified. The minimum number of features is further used by the system for face recognition.

The questions addressed by this paper are:

1. Which face features are most discriminatory?
2. Which are the minimum features needed to identify a face?
3. Are family relatives associated based on the minimum feature set?

4.1 Limitations

This subsection lists the items that are not covered in the paper.
Methods of image compression and improvement of the performance of the image database was not considered. Two images per face with variable expression are taken in a controlled environment. Lighting issues, pose variation and occlusion are left as future replenishment and improvement of the idea. A small training database of 20 faces is used to record preliminary results and a test dataset of 8 faces is used to test the obtained minimum feature set. We worked with 2D face model, since we need proper understanding of the 2D model before we could use the 3D model and its advantages to explore and improve the process of face identification.

4.2 Face components

All face components do not play an equal role in the process of identification [12]. The selection of features was performed by exploring the basic face components and took into account the mutual distance and position in relation to the whole face. Detected components are shown in Fig.1a).

Components:

- Face shape
- Eyes
- Nose
- Mouth

4.3 Face features

Detected components, distances and positions are used to process the features of a face. In the beginning, the number of features was 14 in order to consider the discriminatory properties of all of them. Next, by analyzing this feature set, we received a reduced feature set that can successfully identify face in the database. Talking with an expert in the field of Art[4] and as stated in [13], we concluded that the following are important face distances that are processed from the basic components, since the time of Ancient Rome [14].

Distances: Positions:

taken

[4] Goran Boev – Faculty of Art and Design at European University

- Left eye-Right eye --Separation of face in three equal parts
- Left eye-Nose(both edges) [5] --Eyes position
- Right eye-Nose(both edges) --Nose position
- Left eye-Mouth(both edges) --Mouth position
- Right eye-Mouth(both edges)
- Nose-Mouth
- Nose length
- Nose width
- Face length
- Face width
- Forehead length

The above positions and distances are shown in Fig.1. Features were presented as ratio of Face length/Distance where Distance represents the distances defined above[6].

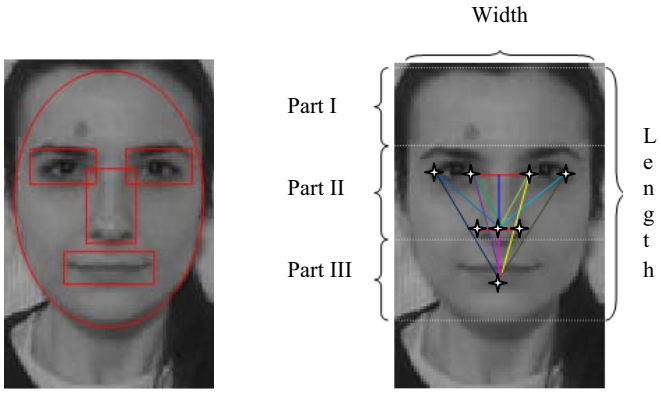
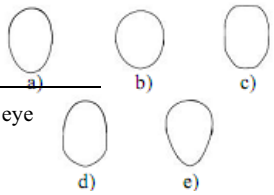


Fig. 1. a) Detected face components b) Face region with detected features

Template matching [15] is a well known method for detecting face features from the early years of face recognition. This method compares the face to a predefined template and extracts the wanted features. We used template matching technique to define the shape of the face being identified. This reduced the search space of the face database. Fig.2 represents the face templates which we used for defining the face shape.



[5] Inner and outer edge of the eye
[6] Excluding the Face length

Fig. 2. Face templates: a) Oval b) Round c) Square d) Pear e) Heart

4.4 Preprocessing

Before detection and processing of the selected features, it was necessary to make appropriate preprocessing and research.

Preprocessing consisted of three stages. In the first stage we built the face training database instead of using an existing one. The purpose of this step was to include family related entities and to make collection of materials in a controlled environment with our own defined parameters[7]. According to the mentioned [16], the second stage included detecting face and non-face region in the image. This was done with custom-made code in Matlab.

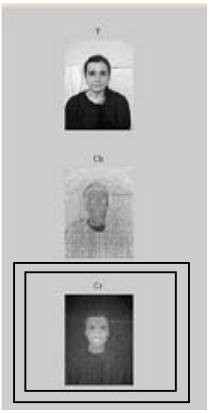


Fig. 3. Portrait image presented in the three components from YCbCr color space

In the second stage, analysis were made [17] on different color spaces, YCbCr (Luminance and Chrominance), RGB (Red, Green and Blue) and HSV (Hue, Saturation and Variance) on the face images. We came to conclusion that the third component from YCbCr space (gray shades) best discriminates the face region unlike other color spaces, because the human visual system perceives color in terms of luminance Y and chrominance CbCr attributes (Fig.3). The final stage covered the detection of the face components which were determined to be most discriminating in the later identification of the faces.

[7] Simple background; lighting, pose and occlusion not included; family related faces included

4.5 Computations

Matlab was used to analyze the discriminative properties of the selected features.

The face identification was implemented using hierarchical cluster trees. A hierarchical cluster tree is created by using the differences or distances among objects in forming clusters. These distances were based on multiple dimensions; each dimension is a rule or condition for grouping objects. In this paper, the distances between faces in multi-dimensional space were determined by calculating the Euclidean distance of the features between each pair of samples (Fig.4). The Euclidean distance is the most commonly used method for geometric distance in a multi-dimensional space (Eq.1).

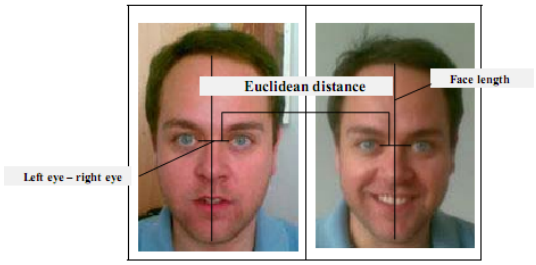


Fig. 4. Sample 1: Euclidean distance on feature Face length/Left eye – right eye in two images with different facial expression

$$d(i)_{1,2} = \sqrt{\sum_{k=1}^{i} |(x_{1i} - x_{2i})\,(y_{1i} - y_{2i})|}$$

Eq.1. Euclidean distance on feature (i) for two images, Image1 and Image2, with coordinates (x_{1i},y_{2i}),

First, the Euclidean distance was calculated with existing function in Matlab, for all the features between all the faces, each face with two images with variable expression. Next, using this information, the faces were grouped into a binary, hierarchical cluster tree. In this step, pairs of faces were linked together that are in close proximity using the *linkage* function in Matlab. The *linkage* function uses the Euclidean distance information generated earlier to determine the proximity of the faces to each other. Later, the *ward*[8] method was used to pair the faces into binary clusters (Eq.2), and the newly formed clusters were grouped into larger clusters until a hierarchical tree is formed [18]. The faces in the hierarchical tree are divided into clusters. The function *dendrogram* generates a dendrogram plot of the hierarchical, binary cluster tree. (Fig.5)

[8]Inner squared distance (min. variance algorithm), appropriate for Euclidean distances; [19]

$$d(r,s) = \sqrt{\frac{2n_r n_s}{n_r + n_s}}\, d(x_r, x_s)$$

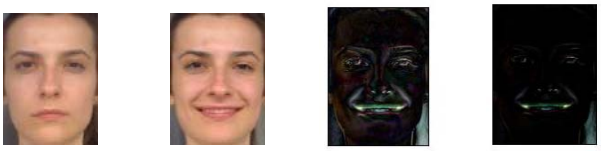
Eq.2. Ward's linkage equation between two clusters, r and s, with number of elements n_r and n_s and centroids x_r and x_s respectively. d (x_r, x_s) represents the Euclidean distance between the centroids x_r and x_s

5 Results and analysis

The hierarchical cluster tree (Fig.5), showed that with all 14 features included, the faces in the training database were successfully identified even on images with variable expression. The circled samples on x-axis show the two images with variable expression belonging to the same person.[9] Each person's two images with variable expression were clustered together in a separate cluster showing that they were properly identified.

Fig. 5. Identified portrait faces of 13 faces with variable expression using hierarhical cluster tree. X-axis (Eucledian distance), Y-axis (Face images, two per face)

Next, the research included testing most of the feature combinations to define a minimum feature set, based on the 14 features already selected. In order to define the minimum feature set, both Matlab and the image tool ImLab were used.



a) b) c) d)

Fig. 6. ImLab results of regions of change of face when smiling
a) Normal b) Smile c) Standard deviation d) Variance

[9] Images 1 and 2 – Face 01, images 3 and 4 – Face 02, etc.

In ImLab standard deviation and variation were calculated over the two images with variable expression for each face in the training database. Fig.6 shows the visual results of the calculations in ImLab for Face 08. The analysis showed that the Face length is a feature without significant change when variable expression images are taken into consideration. Because of this, the ratio Face Length/Distance was taken as a ratio when building the feature vector for each face. Fig.6 c) and d) showed that mouth and eyes' outer corners are changing when face expression changes. The following 6 features were chosen to identify the 20 people in the training database:

- Ratio1: Face lenght/Left eye – right eye
- Ratio2: Face lenght /Right eye – nose (inner edge)
- Ratio3: Face lenght /Left eye – mouth (inner edge)
- Ratio4: Face lenght /Left eye – nose (inner edge)
- Ratio5: Face lenght /Face width
- Ratio6: Face length/Forehead length

The results in Table1 represent the Euclidean distance of the two images of the same face. Table 1 shows the identification results based on all 14 features and the minimum set of 6 selected. Considering a threshold value of 1 the distances show that Face12 and Face13 have high Euclidean distance for their two images which means that they were not found very similar when 14 features were used.

Table 1. Euclidean distance identification results based on 14-set and 6-set feature

Face sample	Euclidean distance	
	14-feature set	6-feature set
Face 01	0,4817	0,3709
Face 02	1,2057	0,8221
Face 03	0,3209	0,2421
Face 04	0,7199	3,2733
Face 05	0,9333	0,2912
Face 06	0,8656	0,3517
Face 07	0,2778	0,1122
Face 08	0,6130	0,2606
Face 09	0,4315	0,1780
Face 10	0,6177	0,2791
Face 11	0,9801	0,3450
Face 12	2,3479	0,2205
Face 13	3,5772	0,3581
Face 14	0,6322	0,2216
Face 15	0,4118	0,3548
Face 16	0,7671	0,2665
Face 17	0,5813	0,3266
Face 18	0,8105	0,3511
Face 19	0,4734	0,2780
Face 20	0,5561	0,1936

However, with the 6 feature set, only Face04's images were not found similar i.e. Euclidean distance was higher than the threshold.

Fig.7 shows the hierarchical cluster tree obtained for the family-related faces in the training database where faces were represented with the minimized 6-feature set. During the identification process, the family related faces were expected to be clustered in lower levels in the hierarchical cluster tree, depending on the inherited local features from the family members. Fig.7 shows that one family relation was not detected i.e. mother and son have similarity above the threshold value and therefore are not clustered as family-related. If a threshold value of 0.8 is taken, it gives better detection of family related faces; however it gives worse face identification. (Table 3 and Fig.7)

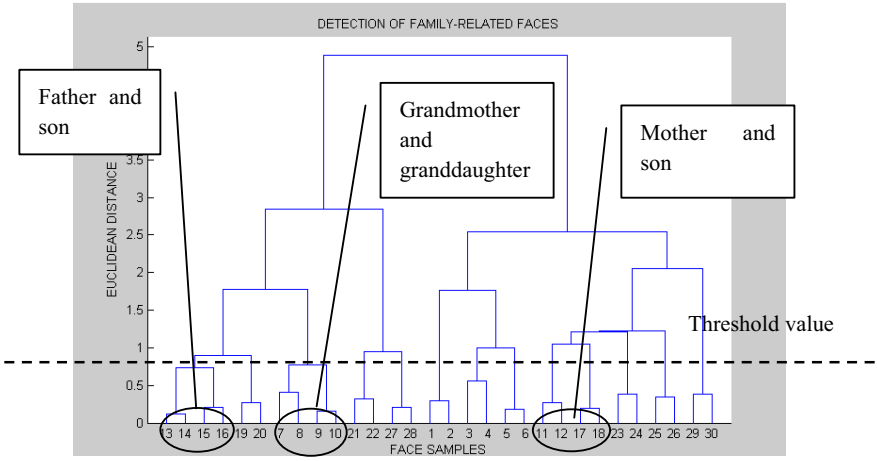


Fig. 7. Family related clustering with 6 features

The face shape was taken to be a 7th feature and was used for initial grouping of the faces, before the hierarhical cluster tree is formed, in order to reduce the face search space. The face templates were shown in Fig.3. Table 2 shows the identification results based on all 14 features and the minimum set of 6 selected where the faces are first grouped by their face shape. This means, identification of the face takes place after the system has determined the shape of the face. Here the search space is reduced and identification process of the face continues only in the selected shape group of faces, instead through the whole database.The results show successful identification on the selected faces with variable expression based on minimum 7 features[10].

The test dataset consisted of 8 face samples, each with 3 images with different facial expressions (Fig.8). Only the neutral expression image for each face in the test dataset is included in the database, while the other two were tested if they are correctly recognized. The identification results of the test dataset are shown in Table 5.

[10] Face shape included

Table 2. Identification results based on 14-set and 6-set feature with face grouping

Face shape – Euclidean distance			
Faces		14-feature set	6-feature set
Shape 1	Face 01	0,4817	0,2942
	Face 02	1,2057	0,5650
	Face 03	0,3209	0,1762
	Face 15	0,7768	0,3812
	Face 17	0,4392	0,3779
Shape 2	Face 09	0,4315	0,1727
	Face 12	0,7172	0,3831
	Face 13	0,6204	0,3326
Shape 3	Face 04	0,7199	0,3972
	Face 05	0,5201	0,1502
	Face 06	0,8584	0,2718
	Face 07	0,2778	0,0936
	Face 08	0,4900	0,1830
Shape 4	Face 10	0,6177	0,2680
	Face 16	0,5491	0,1692
	Face 19	0,8813	0,2969
Shape 5	Face 11	0,7531	0,2058
	Face 14	1,2771	0,5253
	Face 18	0,3480	0,5024
	Face 20	0,7815	0,4482

Table 3. Face identification without face shape grouping

Feature set	No. of Faces	No. of identified (Threshold = 0.8)	Correct Rate (%)	No. of identified (Threshold = 1)	Correct Rate (%)
14 features	20	14	70%	17	80%
6 features	20	18	90%	19	95%

Table 4. Face identification with face shape grouping

Feature set	No. of Faces	No. of identified (Threshold = 0.8)	Correct Rate (%)	No. of identified (Threshold = 1)	Correct Rate (%)
14 features	20	16	80%	18	95%
6 features	20	20	100%	20	100%

Table 5. Face identification without face shape grouping on the test dataset

No. of faces	No. of images	No. of identified (Threshold = 0.8)	Correct Rate (%)	No. of identified (Threshold = 1)	Correct Rate (%)
8	16	14	87.5 %	16	100%

Fig. 8. Test dataset of 8 different faces, each with 3 images.

6 Conclusion

It was concluded that less features can contribute to better identification if selected based on their discriminatory properties. Matlab and ImLab helped discriminate the properties of six important features listed in Section 5. This 6-feature set was extended with another 7^{th} feature and 100% identification was obtained with the training database images. The results were confirmed using small test dataset from a public image database (Fig.8). However, Fig.5 showed clustering of faces that are not family-related if a higher threshold is taken. This could be because of the similar face shape taken as a feature in the 14 feature-set, that significantly influenced the clustering. Fig.7 shows that by taking smaller set of features, family-related members were successfully clustered in lower levels (below 0.8) and proper identification was implemented. In future the influence of the rest of the features in the 14-feature set could be controlled by coefficients which could lead to both, proper identification and detection of family-related faces.

7 References

1. Pentland,A., Choudhury,T. : Face Recognition for Smart environments. In: Biometrics IEEE Computers (Feb. 2000).
2. Kelly, M. D.: Visual identification of people by computer. In: IEEE Trans. Pattern Analysis and Machine Intelligence, 20(3)(1998) 226–239. Tech. rep. AI-130, Stanford AI Project, Stanford, CA.(1970).
3. Kanade T.: Picture processing by computer complex and recognition of human faces, Technical Report, Kyoto University, Department of Information Science, 1973.
4. Goldstein A.J., Harmon L.D., Lesk A.B.: Identification of human faces, Proc. IEEE 59(5) (1971) 748-760.
5. Kaya Y., Kobayashi K.: A basic study on human face recognition, Frontiers of Pattern Recognition, S. 33 Watanabe, (1972) 265-289.

6. Brunelli, R., Poggio, T. : Face Recongition: Features vs Templates. In:IEEE Transactions on pattern analysis and machine intelligence, vol.15, no.10 (1993).
7. Kohonen T.: Self-Organization and Associative Memory, Springer-Verlag, Berlin, 1989.
8. Gaile, G., G.: Face Recongition based on Depth and Curvature Features. U.S. Army Research Office (1992).
9. Yang,M. H., Kriegman, D., Ahuja,N.: Detecting faces in images: A survey. In: IEEE Trans. Pattern Analysis and Machine Intelligence, 24(1) (2002) 34–58.
10. Malkauthekar M. D., Sapkal S. D, : Experimental Analysis of Classification of Facial Images, In: IEEE International Advance Computing Conference, pp.1093-1098, 6-7 March 2009.
11. Ramesha, K., Raja, K., B., Venugopal, K., R., Patnaik, L.,M. : Feature Extraction based Face Recognition, Gender and Age Classification. In: International Journal of Computer Science and Engineering, Vol. 2 No. 01S pp. 14-23 (2010)
12. Pawan S., Balas B., Ostrovsky Y., Russell R.: Face Recognition by Humans: 20 Results all Computer Vision Researchers Should Know About, Department of Brain and Cognitive Sciences Massachusetts Institute of Technology Cambridge, MA 02139
13. Dewi A. R., Suhendra A., Hanum Y.: Facial Feature Distance Extraction as a Face Recognition System Component, The Center for Microelectronics & Image Processing Studies, 2007
14. Bottino A., Laurentini A.: The Analysis of Facial Beauty: An Emerging Area of Research in Pattern Analysis, In: Image Analysis and Recognition, 7th International Conference, ICIAR 2010, Povoa de Varzim, Portugal, June 21-23 2010, Proceeding, Part I
15. Burcu, K.: Face Recognition using Gabor Wavelets. Doctoral thesis. The MiddleEast Technical University, pp.49-60 (Sept. 2001).
16. Skin regions. Stanford University. [Accessed: 20 March 2010.] http://www-csstudents.stanford.edu/~robles/ee368/skinregions.html.
17. Image Processing. Purdue University. [Accessed: 10 April 2010.] www.purdue.edu/VISE/ee438L/lab10/pdf/lab10a.pdf. (Oct. 2008.)
18. Jain K. A., Dubes R. C.: Algorithms for clustering data. Michigan State University, Prentice Hall, 1988.
19. Everitt, B.S., Landau, S. and Leese, M.: Cluster Analysis, Fourth edition, Arnold (2001),.

Improving Content Based Retrieval of Magnetic Resonance Images by Applying Graph Based Segmentation

Katarina Trojacanec, Ivan Kitanovski and Suzana Loskovska

Faculty of Electrical Engineering and Information Technologies, Skopje, Macedonia
{katarina.trojacanec,ikitanovski,suze}@feit.ukim.edu.mk

Abstract. Medical content based retrieval systems are continuously researched and improved. Image segmentation techniques offer a possibility for improvements in the retrieval process. According to this, the aim of the paper is to examine a graph based segmentation technique when it is applied to content based retrieval systems for magnetic resonance images. For this purpose, an evaluation of seven descriptors in both cases: applied to the whole images and on the segmented images is performed. The examination was performed using dataset of hierarchically organized magnetic resonance images. According to the obtained results, we conclude that from the explored descriptors, when Edge histogram descriptor, Region-based shape descriptor and Wavelet transformations are used for feature extraction, the examined segmentation technique leads to improvements in the retrieval effectiveness.

Keywords: Magnetic Resonance Imaging (MRI), Content Based Image Retrieval (CBIR), feature extraction, image segmentation.

1 Introduction

The number of digital images is growing at a rapid rate. This increasing number imposes the need for developing automated systems that will allow efficient storing and precise retrieval. According to this, content based image retrieval systems (CBIR) arise, with the primary goal to enable efficient digital image organization and retrieval from large databases. The concept of CBIR systems revolves around the visual content of images, since they use their content to perform the task of searching images. On the basis of this kind of systems lays the process of extracting information directly from the actual image content, namely, the feature extraction process. Having this information, the CBIR system is able to provide fairly precise retrieval, thus, eliminating the subjective and, even more, error-producing human influence on the process of searching and retrieving.

The medicine is one of the most frequently used terms associated to CBIR systems. Image retrieval for medical applications is becoming more interesting, and is gaining a lot of attention in recent years. The reason for this interest is the vast amount of medical images being produced in digital format by hospitals and medical institutions every day. Medical images are playing an important role in the process of detecting

anatomical and functional information of the body part. Creating CBIR systems for medical purposes improves the process of storing, indexing an analyzing medical image content. Furthermore, tuning the efficiency and precision of the CBIR systems improves the decision making process.

There are many medical CBIR systems developed in recent years. ASSERT (Automatic Search and Selection Engine with Retrieval Tools) [1] is developed by three institutions in USA, Purdue University, Indiana University and University Hospital of Wisconsin. ASSERT works on the image database that consists of High-Resolution Computed Tomography (HRCT) of lung. The system requires users to identify certain specific regions for each image. Then, it extracts 255 features of texture, edges, and shape and gray-scale properties as well. Multi-dimensional hash table is constructed, in the end, to index the HRCT images.

IRMA (Image Retrieval in Medical Applications) [2] is developed by Aachen University in Germany. IRMA is a content-based image retrieval system purely designed for medical applications. The process of image retrieval in this system is split into seven consecutive steps, including categorization, registration, feature extraction, feature selection, indexing, identification and retrieval.

The main constraint of the IRMA system is the absence of retrieval based on local features. To overcome this disadvantage, the new SPIRS-IRMA system is proposed in [3]. The system is a combination of the SPIRS [4] and IRMA system, which are two complementary technologies. SPIRS provides localized vertebral shape-based CBIR methods for pathologically sensitive retrieval of digitized spine x-rays and associated metadata. The idea was to fill IRMA with its missing parts, so that the users may find images that are not only similar in their overall appearance, but also similar in the locally-expressed pathology. The primary use of the system was for research and teaching purposes.

NHANNES II (The Second National Health And Nutrition Examination Survey) is a system developed by National Library of Medicine in USA. The system operates over a database of 17,000 lumbar and cervical spine X-ray images. NHANNES II uses the method of Active Contour Segmentation (ACS) to locate certain parts of the images. This information is later used in the process of optimization of feature extraction. NHANNES II is used for medical purposes and provides good results in the field of X-ray images.

The development of CBIR system is advancing rapidly. But, they are not yet perfected enough. Furthermore, their application in the field of medicine is facing with many problems. Namely, they are not entirely suited for specific medical purposes. Thus, CBIR systems should be adjusted specifically for different medical fields. There are many sensitive aspects related to the concept of medical CBIR systems, such as, the automated segmentation, the lack of standardized criteria for evaluation and the big variability in the feature selection process.

Nevertheless, there are attempts to overcome some of the problems, typical for medical CBIR systems, such as ImageCLEF [5], Insight Toolkit (ITK, for segmentation and registration) [6] etc. Moreover, the IRMA system has many advantages over previous CBIR systems and it is probably one of the most advanced ones. It provides high level of understanding of the image content, but the system lacks semantic labeling and also only global image characteristics can be examined. SPIRS-IRMA, on the other hand, combines both local and global analysis, but its

implementation is far from the theoretical possibilities. There are many CBIR systems in use today, but they are not nearly efficient as they theoretically can be. It is very important for these systems to undertake a clinical evaluation, since their goal is to actively participate in the process of diagnosis, as a reliable medical tool. Thus they need to be continuously researched and improved. Because one possible way for improving the content based image retrieval is including the segmentation process, the graph based segmentation technique is examined in this paper. The main goal is to investigate whether or not this technique is appropriate for magnetic resonance images.

The paper is organized as follows. Section 2 provides a brief overview on CBIR systems for magnetic resonance images signifying their advantages and disadvantages. Section 3 describes the importance of the image segmentation in the field of content based medical image retrieval, while section 4 presents the dataset organization used for this research. The experimental results are shown in section 5. Finally, the concluding remarks are given in section 6.

2 Related Work

One of the most challenging issues of CBIR is to make easier and more efficient the process of organization and searching the increasingly large biomedical image collections. One special case of medical CBIR is the content based retrieval of Magnetic Resonance Images (MRI). There are many researches in this field [5][6][7][8] and [9].

A two-tier, CBIR architecture with special application to brain MRI data was presented in [5]. The first tier is concerned with the overall description of the image and the second tier is concerned with specific regions of interest. The system provides flexibility because it allows users to choose, whether to use only one tier or two tiers. The main purpose of the first tier is to provide general information about the images, while the task of the second is to analyze specific regions of interest. The second tier, in fact, provides a level of semantics about the images. The semantics was defined for a specific application, since it remains hard to define important semantics suited for any application. The system was tested over a dataset of 120 MR images.

Improvement of the two-tier architecture in [5] is presented in [7]. The crucial part of the proposed system is the phase which follows the presentation of the results to the user. This architecture allows the medical expert to add its opinion, as an addition to the systems decision. It is considered that this will improve the precision of the results and make them more relevant.

A system for automated content extraction of images, which was developed as a combination of image registration and natural language processing was proposed in [6]. The choice of specific algorithms, for feature extraction and segmentation, depends on the type of image, the anatomy and the demographic characteristics of the patients. The algorithms and the appropriate parameters can be specified by the user or automatically by the system. They manage to combine the registration and the natural language processing, to indentify relevant images in the brain MR image dataset.

The CasImage [9] system is developed in the University Hospital of Genève, Switzerland. The system allows content-based image retrieval of medical images, including MR images. CasImage is integrated in a PACS environment. Searching and retrieval in system is made via medGIFT searching system. Three types of features are extracted from the images such as the local and global color characteristics and the texture of the images, so that the searching is enabled. The interesting part of the system is that it uses a combination of textual annotation and visual features to perform the search.

These systems have some advantages and disadvantages. Most of the developed MRI CBIR systems are not fully tested on large data sets, because it is hard to find a large, relevant and good set of MR images. According to this, it is hard to make an objective evaluation of these systems. The system proposed in [5] was tested on small number of images, so its full capabilities cannot be assessed properly. The extension, presented in [7], should add significant improvements. Finally, CasImage is a good idea for complex medical CBIR system, since it combines both, the textual information and visual features of the images. But the problem, here, is that textual information is not added automatically, so the entire system is not fully automated. Nevertheless, all these systems have significantly contributed to the CBIR research process.

The concept of MRI CBIR systems is to allow efficient analysis of relevant information gathered from MR images, since MR images hold valuable information about the health of the patients. Having that in mind, these systems should be able to handle the specific characteristics of MR images, efficiently. The development of specialized CBIR systems is intensively going towards efficient and precise clinical decision making support. That is why it is important to continue the development of these systems. Image segmentation is usually one of the most important steps that lead to improvements in this context. According to this, in this paper, a graph based segmentation technique is examined with the aim to improve the content based retrieval for magnetic resonance images.

3 Image Segmentation

Image segmentation is a very important step that leads to better image understanding, analysis and interpretation. It provides richer information than that which can be obtained with feature extraction only methods in the process of information extraction directly from the image content. Automatic segmentation methods offer a mechanism for better image understanding, analysis and interpretation. One of the most important applications of image segmentation in the domain of medical imaging is localization or identifying regions of interest (ROIs) that describe anatomical structures and (pathological) regions of interest. Image segmentation is widely used for medical purposes such as tumor localization, blood cells labeling, surgical planning, image registration, tissue classification [4][10][11][12][13][14] and many others. In fact, the fundamental role of ROIs is to enable quantification and reduction of the searching dataset by focusing the quantitative analysis only on the certain ROIs.

As a result of the low spatial resolution, bad defined edges, noise, objects shape variability, intensity inhomogenity, and many other artifacts coming from the medical image acquisition from different modalities, medical image segmentation is in continuous research and improvement. Moreover, since the segmentation has wide area of application and its ultimate goal depends on the domain of the performed examination, it is impossible to develop one segmentation technique that will provide good results in all medical fields and can be applied to medical images from different modalities.

With the aim to improve the description of the image content, the segmentation process is usually involved in the existing CBIR systems. In this context, image segmentation enables separating ROIs from the whole image that are used for local feature extraction. This approach may induce a more efficient and more precise CBIR system. However, the appropriate image segmentation technique should be chosen on the bases of the specific characteristics of the images and the performed examinations [15].

In this paper, a graph based segmentation technique proposed in [16] was examined. Our goal was to investigate whether this algorithm is appropriate in the case of magnetic resonance images. This technique represents the image segmentation task as a graph where every pixel form the image is represented as a node in the graph. Each edge from the graph connects two neighbor pixels. The graph based image segmentation approach defines edges between different regions from the image on the bases of two concepts: pixel intensity difference from the both sides of the edge and intensity difference between the neighbor pixels within the region. What is specific for the algorithm proposed in [16] is that unlike the classical methods, it adaptively adjusts the segmentation criterion based on the degree of variability in neighboring regions of the image. This is the first reason why we choose this technique for magnetic resonance images segmentation. Namely, this characteristic of the segmentation algorithm deals with the specific MRI characteristics, such as intensity inhomogenity, overlapping tissue intensity distributions, similar intensities of the neighbor tissues. The second reason why we decided to examine this technique is its computational efficiency [16]. Fig.1 depicts the result (the left hand side image) of the applied segmentation technique to the brain MRI (the right hand side image). From this picture it should be noticed that this segmentation performs well in the separation of the pathological regions in the depicted image.

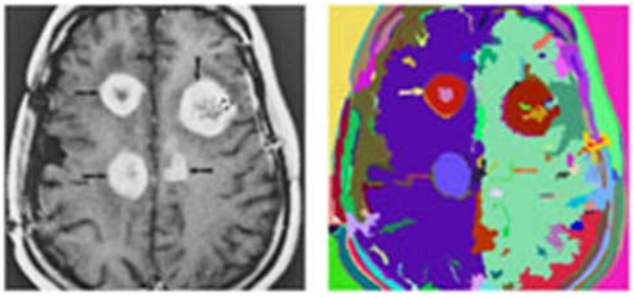


Fig. 1. Graph based segmentation applied to MRI.

4 Dataset description

The dataset used for analysis consists of magnetic resonance images provided by [17] and [18]. The dataset includes brain and abdomen MRIs and MRIs from gynecology domain. A brief textual description is available for each image from the dataset. The provided images did not have any organization. We organized the images, firstly, according to the part of the body they represent, i.e. brain, abdomen, gynecology. Then, we divided each of these classes on the bases of pathology present in the image characteristic for the specified class. The hierarchy that represents this classification is depicted on Fig. 2 [19][20].

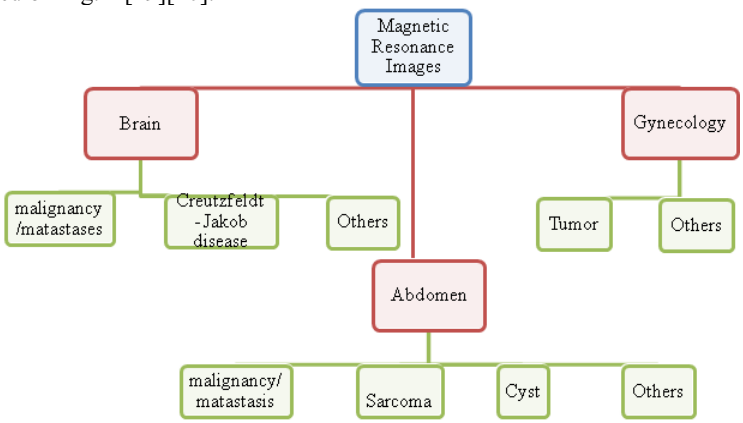


Fig. 2. Hierarchy organization of the dataset.

As we can see from the Fig. 2, the first level of the hierarchy contains three categories: Brain, Abdomen and Gynecology. There are three subclasses contained in the Brain class. The first one contains images taken from patients in whom malignancy, metastases or tumor has been diagnosed. The second subclass represents MRIs where Creutzfeldt-Jakob disease is present. The last subclass, Others, includes images with none of the mentioned pathologies and/or images where no pathological region has been detected. The Abdomen class was divided into four subclasses. The first class contains images with presence of malignancy, metastases or tumor in the abdominal part of the human body, while the second class represents the images with presence of sarcoma. The third subclass includes MRIs that denote presence of cysts in the abdominal part of the examined patients. All other abdominal MRIs are classified in the fourth subclass of the Abdomen class. In the third, Gynecology, class two separated subclasses are obtained, according to the presence or absence of tumor, respectively. Therefore, the examined magnetic resonance images are classified into nine classes, presented by the leaf nodes in the hierarchy from Fig. 2.

There are 1870 magnetic resonance images in the dataset in total. The training set consists of 1247 MRIs, while the test set consists of 623 MRIs. Table 1 depicts the distribution of the number of images through the classes [19][20].

Table 1. Distribution of the number of the images through the classes

Level 1	Level 2	Class No.	Training set	Test set	Total
Abdomen	malignancy /matastases	0	67	34	101
	Sarcoma	1	28	14	42
	Cyst	2	36	18	54
	Others	3	455	228	683
Brain	malignancy /matastases	4	53	27	80
	Creutzfeldt - Jakobdisease	5	13	7	20
	Others	6	343	171	514
Gynecology	Tumor	7	56	27	83
	Others	8	196	97	293
Total			**1247**	**623**	**1870**

5 Experimental Results

The aim of the paper is to examine a graph based segmentation technique when it is applied to magnetic resonance images. To provide that, we compare the results provided by evaluation of seven descriptors after the segmentation of the images with the efficient graph based segmentation technique [16] with the results when no segmentation is used [19][20]. The ultimate goal is to conclude whether the segmentation technique improves the retrieval effectiveness for certain descriptor used for feature extraction. In fact, feature extraction process was performed by using seven algorithms: Edge Histogram Descriptor (EHD) [21], Homogeneous Texture Descriptor (HTD) [21], Region-based Shape Descriptor (RSD) [21], Wavelet transformations [22], Moment Invariants Descriptor (MID) [23], Directional Edge Histogram Descriptor (DEHD) [23], and Directional Edge Histogram Moments Descriptor (DEHMD) [23]. The result of the feature extraction process performed by each of the mentioned algorithms is a separate feature vector for each image from the dataset of MRIs explained in the previous section. The feature vectors are then normalized using min-max normalization technique.

The evaluation of the retrieval effectiveness is provided by calculating the precision and recall [24]. Precision rates are obtained to evaluate the retrieval efficiency based on recall rates from 10% up to 100% with step 10%. Each recall rate shows the fraction of relevant training images that have been retrieved. A training image is considered to be relevant if it has the same class label as the test image.

Precision recall diagrams in both cases, with and without segmentation, for each of the descriptors used for feature extraction are depicted on Fig. 3 and Fig. 4.

Additionally, mean average precision was calculated in both cases. The results in the case when no segmentation is used and in the case when the segmentation process was included are depicted on Fig. 5 and Fig. 6, respectively.

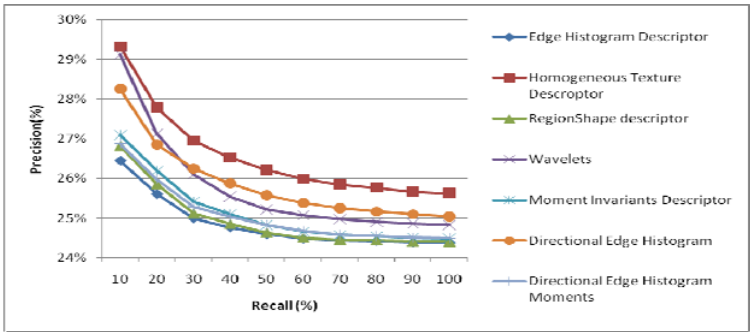


Fig. 3. PR – curves when no segmentation is used.

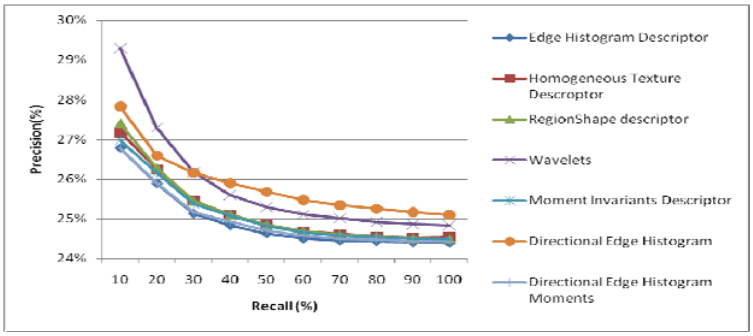


Fig. 4. PR – curves when the segmentation process is included.

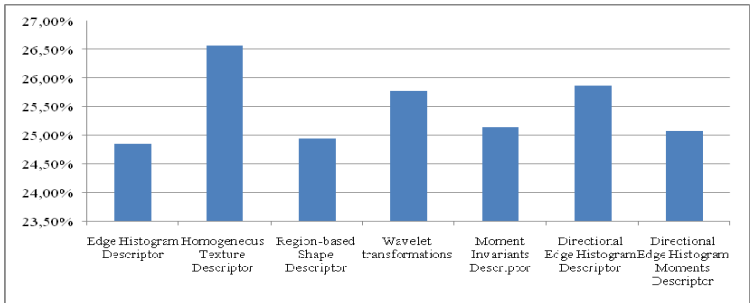


Fig. 5. Mean average precision when no segmentation is used.

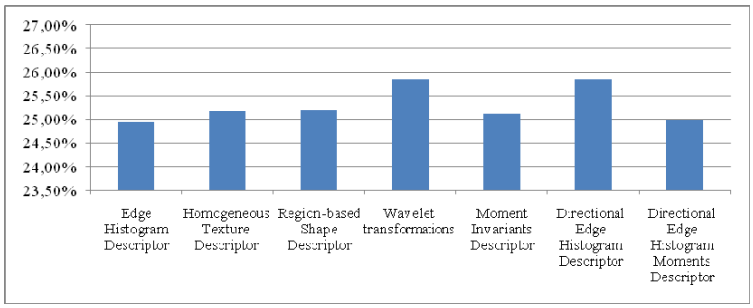


Fig. 6. Mean average precision when the segmentation process is included.

According to the performed examination and the calculated mean average precision, we can conclude that the segmentation process leads to better results in the case of Edge histogram descriptor (without segmentation: 24,85207%, with segmentation: 24,94153%), Region-based shape descriptor (without segmentation: 24,95445%, with segmentation: 25,19852%) and Wavelet transformations (without segmentation: 25,77465%, with segmentation: 25,84249%). In all other cases, the examined graph based segmentation does not improve the retrieval process.

6 Conclusion

In this paper, efficient graph based segmentation was examined when it is applied to magnetic resonance images. For the feature extraction process from the whole MR images and from the segmented images, seven descriptors were evaluated. The aim of the paper was to reveal in which case the segmentation techniques improve the retrieval process. The overall examination was performed on the dataset of magnetic resonance images organized in a specific hierarchy. According to the results provided by the investigation, it can be concluded that the examined segmentation technique improves the retrieval effectiveness when Edge histogram descriptor, Region-based shape descriptor and Wavelet transformations are used in the feature extraction process.

References

1. C. R. Shyu, C. E. Brodley, A. C. Kak, A. Kosaka, A. Aisen, and L. Broderick. Assert: A physicianin-the-loop content-based retrieval system for HRCT image databases. Computer Vision and Image Understanding, 75, 1/2, 111–132 (1999).
2. T. M. Lehmann, H. Schubert, D. Keysers, M. Kohnen, B. B. Wein, Berthold. The IRMA code for unique classification of medical images, Medical Imaging 2003: PACS and Integrated Medical Information Systems: Design and Evaluation, Proceedings of the SPIE, Vol. 5033, pp. 440–451 (2003).

3. Antani, S. K., Deserno, T. M., Long, L. R., Güld, M. O., Neve, L., & Thoma, G. R., Interfacing global and local CBIR systems for medical image retrieval, In Proceedings of the Workshop on Medical Imaging Research, 166-71 (March 2007).
4. Hsu, W., Antani. S. K., & Long L. R. SPIRS: A Framework for Content-based Image Retrieval from Large Biomedical Databases. In Proceedings of the MEDINFO,12(1),188-92 (2007, August).
5. J. Moustakas, K. Marias, S. Dimitrijadis, S.C. Orphandoudakis, A Two-level CBIR Platform with Application to Brain MRI Retrieval.
6. U. Sinha, A. Ton, A. Yaghmai, R. K. Taira, H. Kangarloo. Image Content Extraction: Application to MR Images of the Brain.
7. K. Trojacanec, I. Dimitrovski, S. Loskovska. Content Based Image Retrieval In Medical Applications: An Improvement Of The Two-Level Architecture, IEEE Region 8 Eurocon 2009 Conference, St. Petersburg, Russia, pp. 138–141, 18–23 May 2009.
8. Trojacanec, K.,Dimitrovski, I., Loskovska, S., Texture-based Descriptors Applied to Magnetic Resonance Images, 9-th International Conference Electronics, Telecommunications, Automatics and Informatics (ETAI 2009), Ohrid, Macedonia, (2009).
9. http://www.casimage.com
10. K. O. Lim, and A. Pfefferbaum. Segmentation of MR brain images into cerebrospinal fluid spaces, white, and gray matter, *J. Comput. Assist. Tomogr.*, **13**, 588–593 (1989).
11. D. Brzakovic, X. M. Luo, and P. Brzakovic. An approach to automated detection of tumors inmammograms, *IEEE Trans. Medical Imaging*, **9**, 233–241 (1990).
12. Z. Liang, J. R. MacFall, and D. P. Harrington. Parameter estimation and tissue segmentation from multispectral MR images, *IEEE Trans. Med. Imaging*, **13**, 441–449 (1994).
13. B. A. Ardekani, M. Braun, B. F. Hutton, I. Kanno, and H. Iida. A fully automatic multimodality image registration algorithm, *J. Comput. Assist. Tomogr.*, **19**, 615–623 (1995).
14. I. N. Bankman, T. Nizialek, I. Simon, O. B. Gatewood, I. N. Weinberg, and W. R. Brody. Segmentation algorithms for detecting microcalcifications in mammograms, *IEEE Trans. Inform. Technol. Biomed.*, **1**, 141–149 (1997).
15. R. Albatal, P. Mulhem, Y. Chiaramella, T. Chin. *Comparing image segmentation algorithms for Content Based Image Retrieval Systems*, October 17, 2008.
16. P. F. Felzenszwalb and D. P. Huttenlocher. Efficient graph-based image segmentation. *Int. J. Comput. Vision* (2004).
17. http://www.imageclef.org/ImageCLEF2008, 20.06.2010.
18. http://www.info-radiologie.ch/index-english.php, 20.06.2010.
19. K. Trojacanec, I. Kitanovski, S. Loskovska. Comparison of Classification Techniques Applied to Magnetic Resonance Images. Proceedings of the 7th International Conference for Informatics and Information Technology. 2010.
20. K. Trojacanec, G. Madzarov, D. Gjorgjevikj, and S. Loskovska. Classification of magnetic resonance images. Proceedings of the 32nd International Conference on Information Technology Interfaces (ITI). 2010.
21. Manjunath, B.S., Salembier, P., Sikora, T.: Introduction to MPEG-7: Multimedia Content Description Interface. John Wiley and Sons, pp. 214-229 (2002).
22. Y. M. Latha, B.C.Jinaga, V.S.K.Reddy, Content Based Color Image Retrieval via Wavelet Transforms, IJCSNS, International Journal of Computer Science and Network Security, VOL.7 No.12 (2007).
23. M. Hu. Visual pattern recognition by moment invariants. IEEE Trans. Information Theory, 8(2):179–187 (1992).
24. Vranic, D.V.: 3D Model Retrieval. Ph.D. Thesis. University of Leipzig (2004).

Influence of Segmentation over Magnetic Resonance Image Classification

Ivan Kitanovski, Katarina Trojacanec and Suzana Loskovska

Faculty of Electrical Engineering and Information Technologies, Skopje, Macedonia
{ikitanovski,katarina.trojacanec,suze}@feit.ukim.edu.mk

Abstract. Magnetic resonance imaging is an image based diagnostic technique which is widely used in medical environment. Thus, the efficient automated analysis of this kind of images is of great importance for both, scientific and clinical environment. In this paper, analysis of evaluation results of the classification of magnetic resonance images with different classifiers is conducted. This analysis is provided in both cases, with and without application of graph-based segmentation technique. The aim of the paper is to investigate whether or not this kind of segmentation technique induces improvements in the classification of MRIs. Seven descriptors are used for feature extraction in our paper, and the classification is analyzed in all seven cases. The ultimate goal of the paper is to signify in which combination of classification technique and feature extraction algorithm, the examined segmentation technique is the most appropriate for magnetic resonance images. For the overall investigation in this paper, a specific hierarchical organized dataset of magnetic resonance images is used.

Keywords: Magnetic Resonance Imaging (MRI), image classification, image segmentation, feature extraction, graph-based segmentation.

1 Introduction

The amount of medical images is continuously increasing as a consequence of the constant growth and development of techniques for digital image acquisition. Manual annotation and description of each image is impractical, expensive and time consuming approach. Moreover, it is an imprecise and insufficient way for describing all information stored in medical images. In fact, humans tend to be highly subjective and can produce an amount of inconsistency and insecurity. In other words, manual image analysis is practically impossible taking into account this enormous number of digital medical images.

According to this, an automated approach to annotation is required. Automated image annotation can be achieved with the use of different classification techniques. With the aim to improve the efficiency and precision of the content based image retrieval systems, feature extraction and automatic image annotation techniques are subject of continuous researches and development. Including the classification techniques in the retrieval process enables automatic image annotation in an existing

CBIR system. It contributes to more efficient and easier image organization in the system.

Magnetic resonance is a technique which is widely used in medical environments. Magnetic Resonance Imaging (MRI) has become a useful tool since it provides rich medical information via images, additionally providing high sensitivity with its non-invasive nature. Furthermore, MRI provides high spatial resolution, contrast and superior soft tissue differentiation. It has become a crucial part in the medical diagnosis process. However, MRI characterizes with the intensity ihomogenity and a lot of noise that induce overlapping tissue intensity distributions. These specific characteristics of MRIs make MRI classification a very sensitive problem and a big challenge.

The necessity for efficient and automated analysis of magnetic resonance images rapidly increases as the number of images grows. The classification of MR images is a very difficult task and one that has to take into consideration the specific characteristics of MR images. However, it is of great importance for efficient analysis of this kind of images.

Classification of MR images is an important taskt and it is widely used in research and clinical studies [1],[2],[3],[4],[5],[6]. Researchers have used various classification techniques for MR images, such as Aritifical Neural Networks (ANN) [5], k-Nearest Neighbors (kNN) classifier [7], Bayes classifier [1],[8], Suport Vector Machines (SVMs) [9] and Expectation Maximization (EM) as a statistical classification scheme. SVM classifier is applied on breast multi-spectral MR images in [10]. In [9] a SVM based method for automated segmentation and classification of brain MR images is used.

The paper is organized as follows. Section 2 provides a brief explanation of the classification techniques used in this paper for classification of magnetic resonance images. Section 3 describes the importance of involving segmentation technique in image analysis and gives description of the examined segmentation technique. In section 4 the dataset used in this paper for evaluation is presented, while section 5 provides the experimental results. The concluding remarks are given in section 6.

2 Classification Techniques

Various classification techniques are subject of continuous research such as support vector machines (SVM) [11], the k-nearest neighbor algorithm [12], C4.5 algorithm [13],[14], neural networks [12], CART [15], Bayesian classifier [12] etc.

2.1 Support Vector Machines

Support Vector Machines [16][17][18] are amongst the most efficient classifiers that exist today. The SVMs are designed basically for binary classification which is their main disadvantage.

SVM classifiers, however, can be tuned to solve multiclass classification problems as well. There are many strategies to implement this type of behavior. One possible strategy is the one-against-all (OvA) scheme. It works by dissolving the multiclass

problem, with M classes, where M is larger than two. The idea is to create M number of classifiers where the i-th classifier separates the class i from all other classes. All classifiers should be able to distinguish whether a sample which belong to one class or whether it belongs to all other classes [19].

There is one approach where SMVs [20] are represented in binary tree architecture (SVM-BTA). The core of the algorithm is that, at each node of the tree, the classes are being divided into two groups of classes. Then a SVM classifier, at the node, decides to which group the sample, that needs to classified, belongs to. After that the sample is being moved to the left or right sub-tree, depending on the decision. This is repeated until the sample reaches a leaf node, which represents the class that will be assigned to the sample.

SVMs utilizing Binary Decision Tree (SVM-BDT) involves hierarchical clustering [21]. Using distance matrixes, similar to [20], the classes are divided into two groups, at each node. Only, here the groups are being chosen to be as distant as possible, to achieve better results. When a sample needs to be classified, it goes through the tree in the way previously described, until it reaches a terminal node. An extension of the SVM-BDT method is SVM utilizing Balanced Binary Decision Tree (SVM-BBDA) [21]. The only difference is that here the binary is balanced i.e. the distribution of the classes through the groups is uniform, in order to achieve better results.

The reason for using support vector machines is the fact that the strategy beyond them has many advantages like its simple process of implementation, the precision of recognition and the speed in the training phase and the recognition process.

2.2 C4.5 Algorithm

C4.5 algorithm is a widely used classification method in pattern recognition [22],[23]. It is used for building decision trees for a given training set. This done by use of the concept of information entropy, extracted from the training data.

C4.5 has a few advantages such as the fact that it handles missing values, since they are simply discarded in the entropy calculations. The algorithm also handles with both discrete and continuous values and with nominal attributes. When it comes to continuous values, C4.5 creates a threshold with a purpose to split the values of the features which are above the threshold and those which are equal or below that threshold. In the end of the training process C4.5 goes through the whole decision tree and looks for the branches, which do not influence on the decision making process, and replaces them with leaf nodes.

The decision tree is very easy and fast to train and built, but on the other hand it requires a large amount of data to develop relevant decision making abilities.

2.3 K Nearest Neighbors Classifier

The k-nearest neighbor algorithm (k-nn) [24] is one of the most simple and efficient classification techniques. The core principle of the classification of k-nn lies in the finding the closest training samples. Then the class of the unlabeled sample is decided by a voting procedure i.e. a sample is classified based on the majority vote of the k-

nearest neighbors. The parameter k is manually adjusted integer number, which specifies the number of nearest neighbors that will be considered in the voting. The neighbors are taken from the training set, which consists of pre-classified samples. The main issue at this point is the way the distance is calculated between samples.

K-nn is also used for regression, by assigning the property value of the sample to be the average of the values of its k nearest neighbors.

But, k-nn has a serious flaw in that it tends to promote classes which have an overwhelmingly large number of samples. Classes, which have a number of samples far greater than other classes, tend to dominate the prediction process. So, when we want to classify a certain sample, there is a greater probability that it will be labeled as a member of these classes.

3 Image Segmentation

To provide better image description, the segmentation process is continuously researched [26],[27],[28],[29]. In fact, the main goal of the segmentation techniques is to provide richer information extracted from the images, than it can be obtained with feature extraction process applied on the whole image. Moreover, automatic segmentation methods provide a mechanism for overcoming the disadvantages of the manual segmentation of large datasets, and also promise reproducibility which is difficult with manually defined results [29]. Especially when detailed or quantitative information such as the appearance, size, or shape of the patient anatomy need to be analyzed, then image classification plays a crucial role [26].

Image segmentation for MRI applications is a very sensitive problem, because of the specific nature of MRI characteristics. There are two main reasons behind this: the imaging process itself and the anatomy that is being imaged [26]. In fact, from the view point of the first aspect, magnetic resonance images provides clinically relevant information about the tissue being imaged, but this does not mean that the anatomical feature of interest will be separable from its surroundings. This is because of the noise in the image acquisition process and the intensity nonuniformity of the tissues. Thus simple segmentation techniques, such as threshold based segmentation algorithms, are not precise enough when applied to MRIs. From the point of view of the other aspect, the complexity and variability of the anatomy, involves huge problems in the segmentation process. This may lead to the necessity of involving the detailed anatomical knowledge. According to this, MRI segmentation is not a trivial problem, but it is a wide area of interest where improvements need to be performed.

In this paper, a graph based image segmentation technique proposed in [30] is examined when it is applied to magnetic resonance images. Two fundamental aspects of this technique make it appropriate and interesting for investigation for medical purposes. The first one is that unlike the classical methods, it adaptively adjusts the segmentation criterion. The second important aspect is its computational efficiency. Because this algorithm belongs to the graph based kind of segmentation algorithms, it represents the problem as a graph where each pixel corresponds to a node in a graph. Each pair of neighboring pixels is connected by undirected edges. The dissimilarity between pixels is represented by the weights on each edge. The main difference

between this method and the other classical methods lies exactly on the adaptive adjustment of the segmentation criterion on the bases of the degree of variability in neighboring regions of the image [30]. Fig. 1 depicts the result of the application of this segmentation technique (the right hand side image) to the magnetic resonance image (the left hand side image).

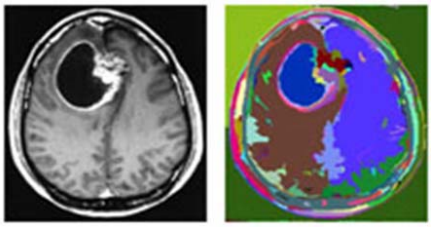


Fig. 1. Graph based segmentation applied to MRI.

4 Dataset Description

For the purposes of our research, the hierarchical organized dataset of magnetic resonance images is used. In fact, the considered dataset contains magnetic resonance images provided by [31] and [32]. The dataset contains brain MRIs, abdominal MRIs and MRIs from the gynecology part. A brief textual description is available for each image from the dataset. Hence, on the bases of textual retrieval we organized the images in a hierarchical way. The first level contains three classes of images classified according to the body part they represent, i.e. brain, abdomen, gynecology. The second level of the hierarchy includes subclasses of the first level classes separated on the bases of pathology present in the patient. The hierarchy that represents this classification is depicted on Fig. 2 [33][34].

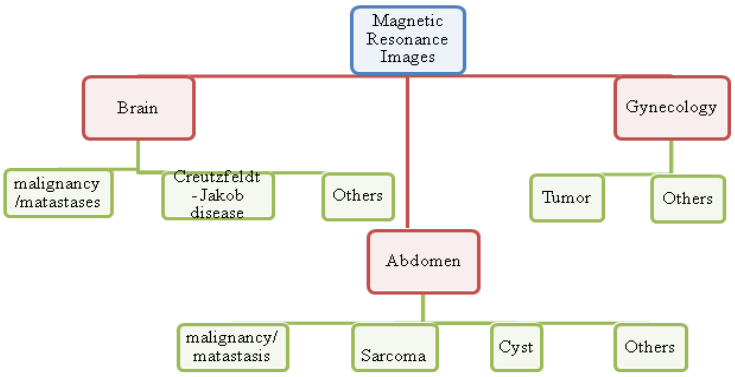


Fig. 2. Hierarchical organization of the MRI dataset.

According to the hierarchical representation shown on Fig. 2, the first level of the hierarchy contains three classes: Brain, Abdomen and Gynecology. The class Brain is separated in three subclasses in the second level of the hierarchy. The first one contains images taken from patients in whom malignancy, metastases or tumor has been diagnosed. The second subclass consists of MRIs where Creutzfeldt-Jakob disease is present. The last subclass, named Others, represents images with none of the mentioned pathologies and/or images where no pathological region has been detected. The Abdominal class was divided into four subclasses. The first class contains images with presence of malignancy, metastases or tumor in the abdominal part of the patients. The second class includes the images with presence of sarcoma, while the third subclass represents MRIs that denote presence of cysts in the abdominal part of the examined patients. All other abdominal MRIs are labeled as Others, which is the fourth subclass of the Abdomen class. Finally, the third, Gynecology class, has two separated subclasses in the second level, in accordance to the presence or absence of tumor, respectively. Therefore, the investigated magnetic resonance images could be classified into nine classes, presented by the leaf nodes in the hierarchy from Fig. 2.

There are 1870 magnetic resonance images in the dataset, from which 1247 MRIs are part of the training set and 623 MRIs belong to the test set. Table 1 depicts the distribution of the number of images through the classes [33][34].

Table 1. Distribution of the number of the images through the classes

Level 1	Level 2	Class No.	Training set	Test set	Total
Abdomen	malignancy /matastases	0	67	34	101
	Sarcoma	1	28	14	42
	Cyst	2	36	18	54
	Others	3	455	228	683
Brain	malignancy /matastases	4	53	27	80
	Creutzfeldt - Jakobdisease	5	13	7	20
	Others	6	343	171	514
Gynecology	Tumor	7	56	27	83
	Others	8	196	97	293
Total			**1247**	**623**	**1870**

5 Evaluation Results

The aim of the paper is to compare different classification techniques in both cases with and without application of segmentation technique to magnetic resonance images. This should provide information whether the examined graph based segmentation algorithm improve the classification results and in which cases. The evaluation of the classifiers we performed on the dataset of magnetic resonance images described in the previous section. The results of this work are an extension of previous research, presented in [33] and [34].

Two main processes are characteristic for our examination, the feature extraction process and the classification process. We used seven descriptors for feature extraction from the visual image content: Edge Histogram Descriptor (EHD) [35], Homogeneous Texture Descriptor (HTD) [35], Region-based Shape Descriptor (RSD) [35], Wavelet transformations [36], Moment Invariants Descriptor (MID) [37], Directional Edge Histogram Descriptor (DEHD) [37], and Directional Edge Histogram Moments Descriptor (DEHMD) [37].

The result of the feature extraction process is separate feature vector obtained for each of the images belongs to both, the train and the test set, for each descriptor. The feature vectors obtained from both segmented and unsegmented images are then normalized using min-max normalization technique and separately passed through the classifiers.

The classification process was performed using seven classification techniques, previously described: Support vector machines based on one-against-all (OvA) scheme, Support vector machines based on one-against-one (OvO) scheme, SVMs in binary tree architecture (SVM-BTA), SVM utilizing Binary Decision Tree (SVM-BDT), SVM utilizing Balanced Binary Decision Tree (SVM-BBDT), K - nearest neighbor classifier and C4.5 algorithm.

The first five algorithms are support vector machines for multiclass classification that that we implemented using the Torch library [38]. For the k nearest neighbor classifier and C4.5 algorithm we used Weka implementation.

The classification error calculated for each classifier when no segmentation was used is depicted in Table 2. Table 3 shows the minimum classification error in the case when the segmentation process was involved.

Table 2. Classification error without segmentation

Classification error (%)	EHD	HTD	RSD	Wavelets	MID	DEHD	DEHMD
SVM OvO	**17,66**	**47,51**	**41,73**	**44,12**	**56,02**	**46,22**	**60,19**
SVM OvA	**18,14**	**47,51**	**44,63**	**42,35**	**55,86**	**48,8**	**68,22**
SVM-BTA	**18,62**	**44,94**	**43,02**	**40,58**	**55,7**	**45,43**	**58,91**
SVM-BDT	**18,78**	**46,71**	**42,54**	**43**	**56,34**	**55,06**	**59,87s**
SVM-BBDT	**18,46**	**45,26**	**43,98**	**42**	**55,54**	**46,73**	**58,59**
k-nn	**18,29**	**50,56**	**43,82**	**44,28**	**51,36**	**49,44**	**61**
C4.5	**32,91**	**51,04**	**47,83**	**51,21**	**50,72**	**59,71**	**58,59**

Table 3. Classification error with segmentation

Classification error (%)	EHD	HTD	RSD	Wavelets	MID	DEHD	DEHMD
SVM OvO	**21.35**	**43.34**	**52.65**	**53.77**	**59.55**	**41.57**	**62.44**
SVM OvA	**22.79**	**44.14**	**54.57**	**52.32**	**62.44**	**44.78**	**73.35**
SVM-BTA	**23.62**	**43.98**	**53.93**	**48.80**	**60.99**	**41.25**	**56.02**
SVM-BDT	**23.78**	**43.18**	**54.90**	**51.69**	**60.67**	**44.78**	**59.71**
SVM-BBDT	**23.46**	**43.98**	**53.77**	**51.69**	**60.67**	**42.38**	**56.02**
k-nn	**21.83**	**44.46**	**56.02**	**49.27**	**57.63**	**43.34**	**56.02**
C4.5	**43.02**	**51.85**	**58.59**	**62.92**	**57.78**	**56.02**	**59.71**

According to the results depicted in Table 2 and Table 3 we can conclude that the segmentation technique raises the classification error in the case of all classifiers, only when Homogeneous texture descriptor and Directional edge histogram descriptor are used for feature extraction. The best classification error in these two cases is obtained when:

- SVM classifier utilizing binary decision tree is used for classification in the case of Homogeneous texture descriptor (classification error of 43,18%)
- SVM classifier utilizing binary tree architecture is used for classification in the case of Directional edge histogram descriptor (classification error of 41,25%)

In all other examined cases, the involved segmentation technique does not improve the classification error.

6 Conclusion

In this paper, investigation of the graph based segmentation technique was conducted to examine whether it improves the classification of magnetic resonance images. For this purpose, seven classifiers were evaluated on the bases of classification error as an evaluation technique in both cases when segmentation technique is included and when it is not included. For the feature extraction purposes, seven descriptors were used. According to the obtained results, we can conclude that classification error is improved only in the case of Homogeneous texture descriptor and Directional edge histogram descriptor in the cases of all classifiers. The best classification error with included segmentation technique when Homogeneous texture descriptor is used as a feature extraction technique was provided by SVM classifier utilizing binary decision tree, while in the case of Directional edge histogram descriptor, SVM classifier in binary tree architecture showed the best results.

Magnetic resonance image analysis is a very sensitive problem. However, every improvement in this field is of great scientific and clinical importance.

References

1. Collins, D., Montagnat, J., Zijdenbos, A.,Evans, A., Arnold, D., Automated estimation of brain volume in multiple sclerosis with BICCR, In: Insana, M. F., Leahy, R. M. (Eds.), Proc. of IPMI 2001. Vol. 2082 of LNCS. Springer-Verlag, pp. 141-147, (2001).
2. MacDonald, D., Kabani, N., Avis, D., Evans, A. C., Automated 3-D extraction of inner and outer surfaces of cerebral cortex from MRI. Neuroimage 12 (3),340-56 (2000).
3. Paus, T., Zijdenbos, A., Worsley, K.,Collins, D. L., Blumenthal, J., Giedd, J. N.,Rapoport, J. L., Evans, A. C., Structural maturation of neural pathways in children and adolescents: in vivo study. Science 283 (5409), 1908-11, (1999).
4. Rapoport, J. L., Giedd, J. N., Blumenthal, J., Hamburger, S., Jeries, N., Fernandez, T., Nicolson, R., Bedwell, J., Lenane, M., Zijdenbos, A., Paus, T., Evans, A., Progressive cortical change during adolescence in childhood-onset schizophrenia. A longitudinal magnetic resonance imaging study. Arch Gen Psychiatry 56 (7), 649-54, (1999).
5. Zijdenbos, A. P., Forghani, R., Evans, A. C., Automatic 'pipeline' analysis of 3D MRI data for clinical trials: Application to multiple sclerosis. IEEE Trans Med Imaging 21 (10), 1280-91, (2002).
6. Trojacanec, K., Madzarov, G., Gjorgjevikj D., Loskovska, D., Classification of Magnetic Resonance Images, Proceedings of the ITI 2010 32nd Int. Conf. on Information Technology Interfaces, Cavtat, Croatia (2010).
7. Kamber, M., Shinghal, R., Collins, D. L., Francis, G. S., Evans, A. C., Model-based 3-D segmentation of multiple sclerosis lesions in magnetic resonance brain images. IEEE Trans Med Imaging 14 (3), 442-53, (1995).
8. Wareld, S. K., Kaus, M., Jolesz, F. A., Kikinis, R., Adaptive, template moderated, spatially varying statistical classication. Med Image Anal 4 (1), 43-55, (2000).
9. H. Selvaraj, S. Thamarai Selvi, D. Selvathi, L. Gewali, Brain MRI Slices Classification Using Least Squares Support Vector Machine, International Journal of Intelligent Computing in Medical Sciences and Image Processing, Vol. 1, No. 1, Issue 1, (2007).
10. Chuin-Mu Wang, Xiao-Xing Mai, Geng-Cheng Lin, Chio-Tan Kuo, Classification for Breast MRI Using Support Vector Machine, Proceedings of IEEE 8th International Conference on Computer and Information Technology Workshops, (2008).
11. J. C. Christopher Burges. A Tutorial on Support Vector Machines for Pattern Recogni-tion, Data Mining and Knowledge Discovery, 2, 121–167 (1998).
12. S. K. Wareld, M. Kaus, F. A. Jolesz, R. Kikinis. Adaptive, template moderated, spatially varying statistical classification. Med Image Anal, 4, 1, 43–55 (2000).
13. B. E. Boser, I. M. Guyon, V. N. Vapnik. A training algorithm for optimal margin classifiers, Fifth Annual Workshop on Computational Learning Theory, pp. 144–152, Pittsburgh, ACM (1992).
14. J. R. Quinlan. C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers (1993).
15. I. H. Witten, E. Frank. Data Mining: Practical Machine Learning Tools and Techni¬ques, Second Edition (2005).
16. V. Vapnik. The Nature of Statistical Learning Theory, 2nd Edition Springer, New York, 1999.
17. C. J. C. Burges. A tutorial on support vector machine for pattern recognition. Data Min. Knowl. Disc. 2 (1998) 121.
18. T. Joachims. Making large-scale SVM learning practical. In B. Sch¨olkopf, C. J. C. Burges, and A. J. Smola, editors, Advances in Kernel Methods—Support Vector Learning, pages 169–184, Cambridge, MA, MIT Press, (1999).
19. Yi Liu Zheng, Y.F., One-against-all multi-class SVM classification using reliability measures, Neural Networks, 2005. IJCNN '05. Proceedings. 2005 IEEE International Joint Conference, Volume 2, 849- 854, 31 July-4 Aug. (2005).

20.Madzarov G., Gjorgjevikj D., Chorbev I., Multi-class Classification using Support Vector Machines in Binary Tree Architecture, International Scientific Conference, Gabrovo, (2008).
21.Madzarov G., Gjorgjevikj D., Chorbev I.,A Multi-class SVM Classifier Utilizing Binary Decision Tree, Informatica 33, pages 233-241, (2009).
22.Quinlan, J. R. C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers, (1993).
23.J. R. Quinlan. Improved use of continuous attributes in c4.5. Journal of Artificial Intelligence Research, 4:77-90, (1996).
24.Xindong Wu, Vipin Kumar, J. Ross Quinlan, Joydeep Ghosh, Qiang Yang, Hiroshi Motoda, Geoffrey McLachlan, Angus Ng, Bing Liu, Philip Yu, Zhi-Hua Zhou, Michael Steinbach, David Hand, Dan Steinberg, Top 10 algorithms in data mining, Knowledge and Information Systems, Vol. 14, No. 1., pp. 1-37 (1 January 2008).
25.Elena Deza, Michel Marie Deza , Encyclopedia of Distances, Springer, page 94 (2009).
26.Lauren O'Donnell, Semi-Automatic Medical Image Segmentation, Master Thesis, Massachusetts Institute of Technology, Department of Electrical Engineering and Computer Science (2001).
27.D.J. Withey and Z.J. Koles, Three Generations of Medical Image Segmentation: Methods and Available Software, International Journal of Bioelectromagnetism, Vol. 9 No. 2 (2007).
28.D.L. Pham, et al., "Current methods in medical image segmentation," Annu. Rev. Biomed. Eng., vol. 2, pp. 315-337 (2000).
29.L.P. Clarke, et al., "MRI segmentation: Methods and applications," Magn. Reson. Imaging, vol. 13, no. 3, pp. 343-368 (1995).
30.P. F. Felzenszwalb and D. P. Huttenlocher. Efficient graph-based image segmentation. Int. J. Comput. Vision (2004).
31. http://www.imageclef.org/ImageCLEF2008, 20.06.2010.
32. http://www.info-radiologie.ch/index-english.php, 20.06.2010.
33.K. Trojacanec, I. Kitanovski, S. Loskovska. Comparison of Classification Techniques Applied to Magnetic Resonance Images. Proceedings of the 7th International Conference for Informatics and Information Technology. 2010.
34.K. Trojacanec, G. Madzarov, D. Gjorgjevikj, and S. Loskovska. Classification of magnetic resonance images. Proceedings of the 32nd International Conference on Information Technology Interfaces (ITI). 2010.
35.Manjunath, B.S., Salembier, P., Sikora, T.: Introduction to MPEG-7: Multimedia Content Description Interface. John Wiley and Sons, pp. 214-229 (2002).
36.Y. M. Latha, B.C.Jinaga, V.S.K.Reddy, Content Based Color Image Retrieval via Wavelet Transforms, IJCSNS, International Journal of Computer Science and Network Security, VOL.7 No.12 (2007).
37.M. Hu. Visual pattern recognition by moment invariants. IEEE Trans. Information Theory, 8(2):179–187 (1992).
38. R. Collobert, S. Bengio, J. Mariethoz, Torch: a modular machine learning software library, Technical Report IDIAP-RR 02-46, IDIAP (2002).

An Application of Ternary Quasigroup String Transformations

Vesna Dimitrova and Hristina Mihajloska

"Ss Cyril and Methodius" University
Faculty of Natural Sciences and Informatics
Institute of Informatics, Skopje, Republic of Macedonia
{vesnap,hristina}@ii.edu.mk

Abstract. One of the most popular current trends in cryptography is the research for new approaches for designing cryptographic primitives. Using the algebraic structures for constructing such primitives is studied by many researches. Our investigation is focused on finding ternary quasigroups with good properties for designing cryptographic primitives. In this paper we define ternary quasigroup transformations for encryption and decryption and show that these transformations are applicable in cryptography for cryptosystems based on quasigroups.

Keywords: ternary quasigroups, $e_{t,f}$-transformation, $d_{t,f}$-transformation

1 Introduction

The applications of quasigroups in cryptography are increasing rapidly. The structures, properties and their large number allow them to be applied in this field. For cryptographic purposes quasigroups have to be of a good quality. This implies that for successful application of quasigroups it is very important to know which quasigroups have good properties for some designs. There are several known cryptographic primitives based on quasigroups [6], [9], but most of them are based on binary quasigroups. Our current research is focused on ternary quasigroups and their application in cryptography. In our previous paper [4] we investigate the structures of ternary quasigroups and give classification based on their structures. In this paper we are focused on finding ternary quasigroups with good properties for designing cryptographic primitives. We define ternary quasigroup string transformations for encryption and decryption and show why these transformations are applicable in cryptography for designing cryptosystems based on quasigroups.

The number of ternary quasigroups of order greater than 4 is too large, so for our research we consider ternary quasigroups of order less than or equal to 4. In section 2 we give a brief introduction to the notions of quasigroups and n-quasigroups. Some mathematical background for ternary quasigroups and their lexicographic ordering are given in section 3. In section 4 we define two ternary

quasigroup string transformations useful in cryptography as encryption and decryption functions. Graphical representations of strings obtained by Ternary Quasigroups String Transformations (TQST) are given in the last section 5.

2 Mathematical background

2.1 Quasigroups and Latin squares

At the beginning of this paper we will give some basic definitions and notations about quasigroups.

A quasigroup (Q, f) is a groupoid (i.e. algebra with one binary operation * on the set Q) satisfying the law:

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

In other words the equations $x * u = v$, $u * y = v$ for each given $u, v \in Q$ have unique solutions x, y.

Latin square is equivalent combinatorial structure to a quasigroup. A Latin square can be associated to any finite binary quasigroup $(Q,*)$ given by its multiplication table. It consists of the matrix formed by the main body of the table, since each row and column of the matrix is a permutation of Q. Conversely, each Latin square L on a set Q gives rise up to $|Q|!^2$ different quasigroups (depending on the bordering of the matrix of L by the main row and the main column of the multiplication table).

For a given quasigroup $(Q,*)$ of order n, $n! - 1$ new operations on the set Q, called *parastrophes* can be derived. For our research we use one of them known as *left parastrophe* (denoted as "\") defined as follows:

$$x_1 * x_2 = x_3 \ \Leftrightarrow \ x_1 \backslash x_3 = x_2 \tag{1}$$

Example 1: Let $Q = \{1, 2, 3, 4\}$ *and* $(Q,*)$ *is binary quasigroup with the Cayley table given on Fig.*1. *a*)*.Using* (1) *we obtain its left parastrophe* $(Q,\backslash)$ *given with the Cayley table on Fig.*1. *b*).

$*$	1	2	3	4
1	4	1	2	3
2	1	4	3	2
3	2	3	1	4
4	3	2	4	1

a)

$\backslash$	1	2	3	4
1	2	3	4	1
2	1	4	3	2
3	3	1	2	4
4	4	2	1	3

b)

Fig. 1. Quasigroup $(Q,*)$ and its parastrophe quasigroup $(Q,\backslash)$

2.2 *n*-Quasigroups and Latin cubes

An n-groupoid $(n \geq 1)$ is algebra (Q, f) on a nonempty set Q as its universe and with one n-ary operation $f: Q^n \to Q$. [1]
An n-groupoid (Q, f) is said to be an n-quasigroup if the equation

$$f(a_1, a_2, \dots, a_{i-1}, x, a_{i+1}, \dots, a_n) = b \tag{2}$$

has single solution x in Q for each $a_1, a_2, \dots, a_n, b \in Q$ and for each $i = 1, 2, 3, \dots, n$. In other words it means that (Q, f) is an n-quasigroup if knowledge of n elements of the $n + 1$ elements in the equation (2), uniquely determine the remaining unknown element.

In the special case when $n = 1$ the quasigroup is called a unary quasigroup, when $n = 2$, we have a binary quasigroup or only quasigroup and when $n = 3$, this kind of quasigroup is called a ternary quasigroup.

Equivalent combinatorial structure to n-quasigroup is an n-Latin square. The main body of the multiplication table of an n-quasigroup (Q, f) is an n-Latin square (see [10]). On the other hand, from an n-Latin square we can obtain A^3 n-quasigroups, where A is the number of all binary quasigroups of order n. (Note that a 1-Latin square is a permutation of Q, a 2-Latin square is a Latin square and a 3-Latin square is a Latin cube).

3 Ternary Quasigroups

In this section we will give some notations and properties of ternary quasigroups. Let $Q = \{a_2, a_2, \dots, a_k\}$ be a given finite set and $f: Q^3 \to Q$ is a ternary operation in Q. Then the quasigroup (Q, f) is called a ternary quasigroup.

Let $x_1, x_2, x_3 \in Q$. If we fix x_1 and define $f_{x_1}(x_2, x_3) = f(x_1, x_2, x_3)$, then (Q, f_{x_1}) is a binary quasigroup that can be represented with its Cayley table. Binary quasigroup (Q, f_{x_1}) we call x_1-plane quasigroup of (Q, f), and the corresponding Latin square we call x_1-plane Latin square.

Lemma 1. *Let $(Q, f_1), (Q, f_2), \dots, (Q, f_k)$ be k binary quasigroups and let $L_1, L_2, \dots, L_k$ be the corresponding Latin squares. The binary quasigroups form ternary quasigroup (Q, f), if in the three-dimensional matrix $L = [a_{ijt}]_{k \times k \times k}$ constructed from the corresponding Latin squares placed one above the other, are satisfied the following implications*

$$\begin{aligned} i \neq i' &\Rightarrow a_{ijt} \neq a_{i'jt}, \\ j \neq j' &\Rightarrow a_{ijt} \neq a_{ij't}, \\ t \neq t' &\Rightarrow a_{ijt} \neq a_{ijt'} \end{aligned}$$

for each $i, j, t \in \{1, 2, \dots, k\}$.

Here we give one example of ternary quasigroup of order 4.

Example 1: Let $Q = \{1,2,3,4\}$ *and* (Q, f_i) *for* $i = 1,2,3,4$ *are binary quasigroups with Cayley tables given on Fig.* 2.

f_1	1	2	3	4
1	1	2	3	4
2	2	1	4	3
3	3	4	2	1
4	4	3	1	2

f_2	1	2	3	4
1	2	3	4	1
2	3	2	1	4
3	4	1	3	2
4	1	4	2	3

f_3	1	2	3	4
1	3	4	1	2
2	4	3	2	1
3	1	2	4	3
4	2	1	3	4

f_4	1	2	3	4
1	4	1	2	3
2	1	4	3	2
3	2	3	1	4
4	3	2	4	1

Fig. 2. Ternary quasigroup of order 4

All these quasigroups satisfy the condition that elements on the (i,j) *position in all 4 quasigroups are different i.e the 4-tuple of all these elements is permutation of* Q. *So using them we can construct ternary quasigroup* (Q, f) *where* $f(x_1, x_2, x_3) = f_{x_1}(x_2, x_3)$ *for fixed* x_1 *from* Q *and* x_1, x_2, x_3 *also from* Q. *Also, using these quasigroups we can make* $4! = 24$ *different ternary quasigroups depending on the order of their arrangement.*

3.1 Lexicographical ordering of ternary quasigroups

For our research we use the lexicographic ordering of the finite ternary quasigroups. We take that the universe set is $Q = \{1, 2, \dots, k\}$ and that the ternary quasigroups are given by their Latin cubes. The linear presentation of ternary quasigroup (Q, f) of order k is given by string which consists of all its k linearly presented binary quasigroups. Now the lexicographic ordering of the linear presentations of all ternary quasigroups of order k gives the ordering of ternary quasigroups.

For our research we consider the ternary quasigroups of order 4 which number is 55296. We take the set $Q = \{1, 2, 3, 4\}$. Using the lexicographical ordering of binary quasigroups given in [2] and the method describe above, we present a ternary quasigroup as a string of 64 characters that is a concatenation of the rows of the corresponding Latin squares. Then we apply the lexicographic ordering of all strings obtained from ternary quasigroups, assuming that the characters are already ordered. Therefore, the linear presentation of the first (lexicographically ordered) ternary quasigroup is the following:

1: 1234|2143|3412|4321||2143|1234|4321|3412||3412|4321|1234|2143||4321|3412|2143|1234

Using the lexicographical numbers of binary quasigroups, we can see that the first ternary quasigroup is built up from the 1-th, 172-th, 405-th and 576-th binary quasigroups.

According to the previous for better vision of ternary quasigroups of order 4 we present them graphically by their Latin cubes, when we put corresponding Latin squares of order 4 one above the other. On Fig.3. is given graphical presentation of the Latin cube with lexicographic number 55296.

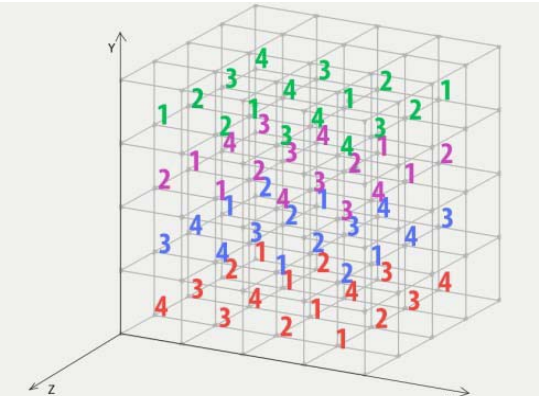


Fig. 3. 55296-th Latin cube

4 Ternary Quasigroup String Transformations (TQST)

Quasigroup string transformations based on binary quasigroups are already defined in the paper [8]. In this section, similarly as in [8], we will define quasigroup string transformations based on ternary quasigroups and will call them ternary quasigroup string transformations (TQST).

Let $Q = \{1, 2, 3, 4\}$. We denote by $Q^+ = \{a_1 a_2 \dots a_k | a_i \in Q, k \geq 2\}$ the set of all finite strings with elements of Q. Let (Q, f) be a given ternary quasigroup consisted from the binary quasigroups $(Q, f_{n_1}), (Q, f_{n_2}), \dots, (Q, f_{n_k})$ from the bottom to the top. For fixed elements l_1, l_2 from Q, called leaders, we define ternary quasigroup string transformations (TQST) $e_{t_{l_1,l_2},f}$, $d_{t_{l_1,l_2},f} : Q^+ \to Q^+$ as follows:

Let $a_i \in Q, \alpha = a_1 a_2 \dots a_n$, for each $i \in \{1, 2, \dots, n\}$. Then

$$e_{t_{l_1,l_2},f}(\alpha) = b_1 b_2 \dots b_n \Leftrightarrow \begin{cases} f(l_1, l_2, a_1) = f_{l_1}(l_2, a_1) = b_1 \\ f(l_2, b_1, a_2) = f_{l_2}(b_1, a_2) = b_2 \\ f(b_1, b_2, a_3) = f_{b_1}(b_2, a_3) = b_3 \\ \dots \\ f(b_{n-2}, b_{n-1}, a_n) = f_{b_{n-2}}(b_{n-1}, a_n) = b_n \end{cases}$$

The ternary function $e_{t_{l_1,l_2}}$ is called $e_{t,f}$-transformation of Q^+ based on the ternary operation f with leaders l_1 and l_2 and its graphical representation is shown on Fig.4.

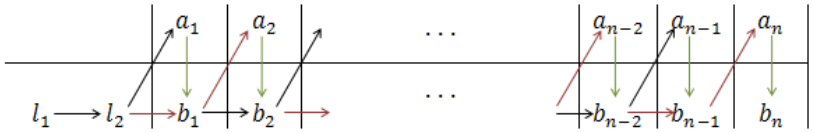


Fig. 4. Graphical representation of $e_{t_{l_1,l_2},f}$-transformation

$$d_{t_{l_1,l_2},f}(\alpha) = c_1c_2...c_n \Leftrightarrow \begin{cases} f(l_1,l_2,a_1) = f_{l_1}(l_2,a_1) = c_1 \\ f(l_2,a_1,a_2) = f_{l_2}(a_1,a_2) = c_2 \\ f(a_1,a_2,a_3) = f_{a_1}(a_2,a_3) = c_3 \\ \dots \\ f(a_{n-2},a_{n-1},a_n) = f_{a_{n-2}}(a_{n-1},a_n) = c_n \end{cases}$$

The ternary function $d_{t_{l_1,l_2}}$ is called $d_{t,f}$-transformation of Q^+ based on the ternary operation f with leaders l_1 and l_2 and its graphical representation is shown on Fig.5.

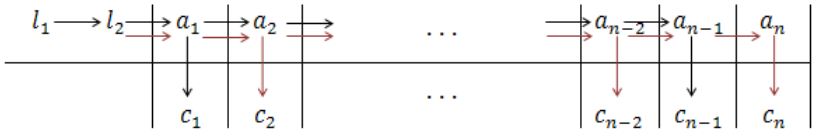


Fig. 5. Graphical representation of $d_{t_{l_1,l_2},f}$-transformation

We can apply consecutive $e_{t,f}$ or $d_{t,f}$ -transformations on a given string, as a composition of $e_{t,f}$ or $d_{t,f}$ -transformations using the same or different leaders l_1 and l_2 for each transformation. For our research we use the same starting leaders l_1 and l_2 for each transformation.

Let (Q,f) be a given ternary quasigroup consisted of four binary quasigroups $(Q,f_1),(Q,f_2),(Q,f_3)$ and (Q,f_4). If we make parastrophe quasigroups $(Q,f_i^{'})$ for $i=1,2,3,4$ where $f_i^{'}$ is the left parastrophic operation of f_i defined as in (1) for each (Q,f_i) and construct ternary quasigroup from these quasigroups then we will obtain ternary quasigroup $(Q,f^{'})$ which is parastrophe on given (Q,f) quasigroup.

For cryptographic purposes we take $e_{t,f}$ -transformation based on operation f and $d_{t,f^{'}}$ -transformation to the operation $f^{'}$.

Similarly as for binary transformations, the following lemmas are true for TQST.

Lemma 2. *The transformation functions* $e_{t,f}$ *and* $d_{t,f'}$ *are permutations on* Q^+ *and* $(e_{t,f})^{-1} = d_{t,f'}$.

Lemma 3. *For each* $\alpha \in Q^+$, $e_{t,f}\left(d_{t,f'}(\alpha)\right) = \alpha = d_{t,f'}(e_{t,f}(\alpha))$

Example 3: Let $Q = \{1,2,3,4\}$, (Q,f) *is ternary quasigroup given on Fig.2,* (Q,f') *is a left parastrophe ternary quasigroup given on Fig.6,* $l_1 = 1$ *and* $l_2 = 2$ *and* $\alpha = 2134122134221$ *is the finite sequence of elements of* Q.

f_1'	1	2	3	4
1	1	2	3	4
2	2	1	4	3
3	4	3	1	2
4	3	4	2	1

f_2'	1	2	3	4
1	4	1	2	3
2	3	2	1	4
3	2	4	3	1
4	1	3	4	2

f_3'	1	2	3	4
1	3	4	1	2
2	4	3	2	1
3	1	2	4	3
4	2	1	3	4

f_4'	1	2	3	4
1	2	3	4	1
2	1	4	3	2
3	3	1	2	4
4	4	2	1	3

Fig. 6. Corresponding parastrophe ternary quasigroup of quasigroup given in Fig.2

Firstly, if $e_{t,f}$ *-transformation with operation* f *is applied on the starting sequence* α *and then* $d_{t,f'}$ *-transformation with operation* f' *on the obtained sequence* β *the starting sequence* α *would be returned (see Fig.7).*

l_1	l_2	2	1	3	4	1	2	2	1	3	4	2	2	1	$= \alpha$
1	2	1	2	4	3	2	3	1	3	2	1	3	4	2	$\beta = e_{t,f}(\alpha)$
1	2	2	1	3	4	1	2	2	1	3	4	2	2	1	$\alpha = d_{t,f'}(\beta)$

Fig. 7. Encryption and decryption functions

For designing cryptographic primitives $e_{t,f}$ -transformation based on operation f is used as encryption function and $d_{t,f'}$ -transformation on the operation f' is used as decryption function.

5 Graphical presentation of TQST

We give a graphical presentation of ternary quasigroup string transformations (TQST) in order to obtain a suitable tool for their better vision. We can use this presentation to discover and investigate some of their properties. The method for obtaining graphical presentation of $e_{t,f}$ or $d_{t,f}$ -transformations given in [3], we adjusted for TQST.

Example 4. The corresponding images for ternary quasigroups with lexicographic numbers 1 and 54594 and the periodical starting sequence $s = 12341234 \ldots 1234$ *with length* $t = 200$*, leaders* $l_1 = 1$ *and* $l_2 = 2$ *and* $k = 100$ *times of* $e_{t,f}$*-transformation, are shown on Fig.8.*

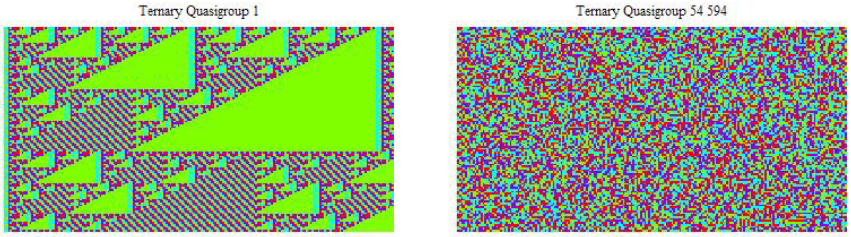


Fig. 8. Graphical presentation of $e_{t,f}$-transformation

Example 5. The corresponding images for ternary quasigroups with lexicographic numbers 1 and 54594 and the periodical starting sequence $s = 12341234 \ldots 1234$ *with length* $t = 200$*, leaders* $l_1 = 1$ *and* $l_2 = 2$ *and* $k = 100$ *times of* $d_{t,f}$*-transformation, are shown on Fig.9.*

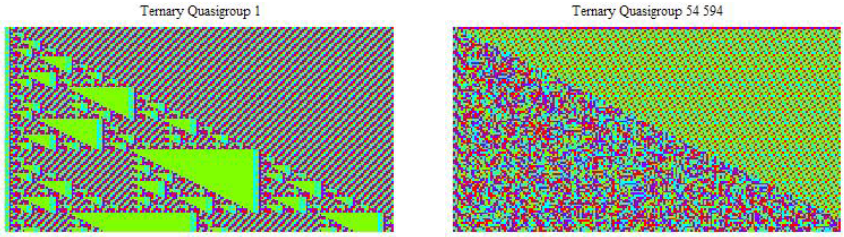


Fig. 9. Graphical presentation of $d_{t,f}$-transformation

The graphical representations of ternary quasigroup transformations are similar to graphical representations of binary quasigroup transformations. Also, using these transformations we obtain two types of images, one with fractal structure and the other with non-fractal structure. For cryptographic usage are interesting the ternary quasigroups that belong to the class of non-fractal and pure non-linear quasigroups denoted as $C_{nf,pnl}$,according the classification given in [4]. Analyses are made on all of the ternary quasigroups that belong to this class. Graphical representations on the ternary transformations made with the same starting sequence and leaders gave the images with non-fractal structure.

6 Conclusion and further directions

In this paper we defined ternary quasigroup string transformations (TSQT) which are useful in cryptography for encryption and decryption and showed that these

transformations are applicable for cryptosystems based on quasigroups. We gave some results of our analysis on sequences obtained by these transformations, but still a lot of researches have to be done. Our future research will be deeper analysis of the properties of TQST and exploitation of these transformations for designing some cryptographic primitives as stream ciphers, block ciphers, hash functions, etc.

References

1. Belousov, V. D: n-arnie Kvazigruppi (n-ary Quasigroups), Stiinca, Kisiniev, (1972)
2. Dimitrova, V: Kvazigrupni transformacii i nivni primeni (Quasigroup Transformations and Their Applications), MSc thesis, Skopje, (2005)
3. Dimitrova, V., Markovski, S: Classfication of quasigroups by image pat terns, Proc. of the Fifth International Conference for Informatics and Information Technology, Bitola, Macedonia, (2007), pp. 152 – 160
4. Dimitrova, V., Mihajloska, H: Classification of ternary quasigroups of order 4 applicable in cryptography, in print, 7 International Conference of Informatics and Information Technology, Bitola, Feb. 2010
5. Gligoroski, D., Dimitrova, V., Markovski, S: Quasigroups as Boolean functions, their equation systems and Groebner bases", Book: "Groebner Bases, Coding, and Cryptography, Springer (2009), pp. 415–420
6. Gligoroski, D., Markovski, S: Quasigroup transformations and their cryptographic potentials, talk at EIDMA Cryptography Working Group, Utrecht, October 10, 2003
7. Mahesar, Q., Sorge, V: Classification of Quasigroup - structures with respect to their Cryptographic Properties, Proc. of the ARW 2009 Bringing the GAP between Theory and Practice, Liverpool, UK (2009) 23–25
8. Markovski, S: Quasigroup string processing and applications in cryptography, First Intern. Conf. Mathematics and Informatics for Industry, Thessaloniki, Greece (2003), 278–289.
9. Mileva, A: Cryptographic primitives and their applications, PhD Thesis, Skopje, Macedonia, 2010
10. Markovski, S., Dimitrova, V., Mileva, A: A new method for computing the number of n-quasygroups, The Journal "Buletinul Academiei de Stiinte a Republicii Moldova. Matematica", No.3 (52), (2006), pp. 57–64
11. Mullen G. L., Weber R. E: Latin cubes of order <= 5, Discrete Math. 32 (1980), no. 3, 291–297

Location Based Systems for retrieval using mobile devices

Aneta Mirceska[1], Vladimir Trajkovik[2], Katerina Ristevska[1]

[1]TTK Bank, Skopje, R. Macedonia
[2]Faculty of Electrical Engineering and Information Technology, University Sts. Cyril and Methodius, Skopje, R. Macedonia
{amirceska,ketiris@}yahoo.com, trvlado@feit.ukim.edu.mk

Abstract. Location-awareness is a key issue for mobile and handheld computing and many people expect a high potential of location-based services as city guides or navigation systems for m-commerce. Mobile information retrieval is crucial to share the vast information and multimedia content. Mobile information retrieval has three challenges; mobile devices have a small display area, so, users are more interested in precision than in recall, also mobile device portability means users frequently change their interest with location and though wireless development is progressing; its response rate and mobile device processing ability are lower than for a wired network and PC. Several approaches are described and to meet these challenges, we propose keyword-based semantic mobile search system using ontology and location information using GPS module and map with mashup service. The objective is to combine all the good things about mobile technology while trying to avoid the shortcomings that come with it.

Keywords: Location-based Service, Mobile Information Retrieval, Semantic Information Retrieval, Global Positioning System

1 Introduction

Determining a mobile user's current location will be one of the most important functions of future mobile computing environments. Location-awareness is a key issue for mobile, ubiquitous and handheld computing and many people expect a high potential of location-based services as city guides or navigation systems for m-commerce scenarios [1]. Mobile systems and services are an important part of the liberty most of us want to rely on. They enable us to stay in contact all the time with other persons or to access data stored on remote systems. Location based services are offered from various sides for users of cell phones. Persons can get guiding support to the next restaurant or ATM, and even meetings of friends can be enhanced by these services. For these applications the cell phones can be localized by use of the cellular phone network or GPS (Global Positioning System). GPS receivers are inexpensive and the corresponding location output is accurate, thus GPS is widely accepted. GPS however only works outdoors since the receiver must have a direct "view" to at least four GPS satellites. If a location-based service is designed for a high coverage, it has to use other (e.g. indoor) positioning systems. Accessing more positioning systems increases the complexity of location-based service. The World-Wide Web has reached a size where it is becoming increasingly challenging to satisfy certain information needs. While search engines are still able to index a reasonable subset of the (surface) web, the pages, and the user is really looking for may be buried under hundreds of thousands of less interesting results. Thus, search engine users are in danger of drowning in information. A natural approach is to add advanced features to search engines that allow users to express constraints or preferences in an intuitive manner, resulting in the desired information to be returned among the first results [2]. We expect that geographic search engines, i.e., search engines that support geographic preferences, will have a major impact on search technology and associated business models. First, geographic search engines provide a very useful tool. They allow a user to express in a single query what might take multiple queries with conventional search engines. Second, geographic search is a fundamental technology for location-based services on mobile devices. Third, geographic search supports locally targeted web advertising, thus attracting advertisement budgets of small businesses. Other opportunities arise from mining geographic properties of the web, e.g., for market research. The paper is structured as follows: Section 2 outlines related work. In Section 3, we describe the entire system, keyword-based semantic retrieval module and location based retrieval module and mash-up service. The last section contains concluding remarks of proposed system and future directions.

2 Related work

Many location-based applications have been developed over the last years. Geographic Information Systems (GIS) and spatial databases provide powerful mechanisms to store and retrieve location data [3]. Such systems primarily concentrate on accessing large amounts of spatial data. General purpose search engines, such as Google, are "live", in the sense that they have the ability to constantly find information without the sources of that information having to report or register it explicitly, or even know that they are being analyzed. Yokoji et al [4] developed a location-based search system for web documents on the Internet; the system can find web documents based on the distance between locations that are described in web documents and a location specified by a user using three modules: a robot that gathers documents from the Internet, a parser that extracts address strings from web documents and associates latitude-longitude information to the original document and a retrieval module.

2.1 Geo coding, geographic search engine and Semantic Web

In this section, we describe related work on geo coding, existing geographic search engines, and the Semantic Web. The process of assigning geographic locations to web pages that provide information relevant to these locations is called geo coding. A document can be associated with one or multiple locations. Geo coding can be dividing into three steps, geo extraction, geo matching, and geo propagation. A good discussion of geographic hints commonly found in web pages is provided by McCurley [5], who introduces the notion of geo coding. He describes various geographic indicators found in pages, such as zip codes or town names, but does not discuss in detail how to extract these, or how to resolve the geo/geo or non-geo/geo ambiguity. Several geographic search engines are already available online. Some are academic prototypes based either on small specialized collections or a meta search approach. In particular, [6] performs automatic geo coding of documents based on the approach in [7]. Most other prototypes, such as [8], require pages either to carry special mark up tags or to be manually registered with the search engine. It seems natural to extend the Semantic Web to a Geographic Semantic Web, such as proposed in [9], where each web page contains some meta data, defining its geographic footprint. Several models are already available [10,11]. Other models from the GIS community, such as GML from the Open GIS Consortium [12], can be adapted.

2.2 Mobile Information Retrieval

Mobile information retrieval is important to share the vast information and multimedia content. Mobile information retrieval differs from existing non-mobile information retrieval. The features of mobile device and ubiquitous computing environment confer many challenges, such as the mobile device has a small display area, therefore, in the mobile search environment, the users are more interested in precision than in recall [13] . Another challenge is the locality where the search for information is focused may continually change due to the portability of mobile devices. Thus, the user's interest is frequently changed as the user moves to a new location [15]. Fast response is important for user satisfaction. Although there is rapid progress in research and development of wireless networking and communication technologies [16], the speed and mobile device processing ability remain lower than for a wired network and PC. P. Coppola [13,14] divided information retrieval into e-relevance and w-relevance. He used the term "relevance" to denote information relevant to the user. E-relevance denotes classical relevance, i.e. the non-mobile information retrieval case. Conversely, w-relevance concerns mobile information retrieval. He proposed four dimensions – information resources, representation of the user problem, time, and components - that can be used in the implementation and evaluation of information retrieval systems. The two relevancies differ, because e-relevance is in the "information world", whereas w-relevance is in the "real/physical world".

2.3 Semantic Search

The search is divided generally in two; navigational search and research search [17]. They differ with the type of the query; navigational search requires the exact word or sentence the user wants to find. For research search, the user inputs the query that explains or describes information related to that the user wishes to collect or research. Then, research search differs from navigational search, because the user does not know about the documentation before searching. Thus, the user tries to find the related document by getting information about the place of the document, as the result of navigational search. Semantic-based search belongs to research search; it is formulated to improve existing keyword-based web search. The query user inputs indicate one (or two) real world concepts. This helps understand the users expected categorization of the search result. The result of semantic search is independent of the result of the keyword-based search, because it is based on semantic web. Therefore, it can increase the quality of the result, as well as the quantity.

3 Proposed Architecture

3.1 System Architecture

The system consists of several parts that communicate with each. All communication between handset and servers, and all three interfaces, use the HTTP protocol [18]. This is the best supported protocol, both by cellular networks and phones. The user employs the mobile device and transmits the created multimedia content to the web server through the wireless TCP/IP socket after commenting on the content and receives the coordinates of the current location through the GPS. This is the first level - creation level. The second level is sharing level. At this level, the user can share the multimedia content created and ontology information using the keyword-based semantic retrieval module. The user creates and sends the input query to find the information or content. The server receives the query and responds it to the keyword-based module and semantic-based module, respectively. The system integrates the search result and creates the structured XML (eXtensible Markup Language) document. XML is a hierarchical format for data representation and exchange. An XML document consists of nested XML elements starting with root element. Each element can have attributes and text values, in addition to nested sub-elements.
The user checks the list resulting from the search result by accessing this XML document. The user selects the content and displays it if s/he finds what they wanted. The system consist three modules: keyword-based retrieval module, semantic-based retrieval module [19] and location-based retrieval module. Fig. 1 show the overall data flow. A mobile device gets raw location data from one or more positioning systems. Framework transforms these data and produce unique location data using the decentralized infrastructure. To achieve an optimal flexibility, the framework provides physical coordinates as well as semantic locations. As these data are globally unique with a well-defined format, they can easily be used as a search key to access database, user registers or web services. Mobile users can switch between satellite navigation systems such as GPS (Global Positioning Systems), positioning systems based on cell-phone infrastructures or indoor positioning systems without affecting the location-aware application.

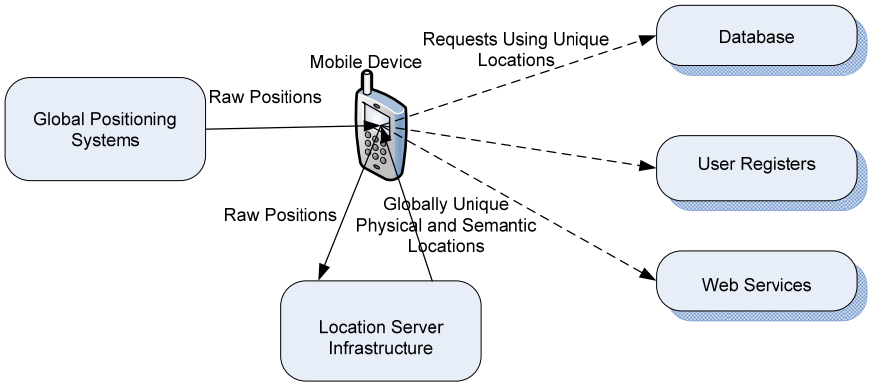


Fig. 1. Architecture of mobile based search system with keyword-based retrieval module, semantic-based retrieval module and location-based retrieval module

Keywords are meaningful text associated with entities; they are more appropriate for humans than for machines. The parser extracts keywords from two different sources: FQNs (Fully Qualified Name) and comments. Keyword extraction from FQNs is based on language specific heuristics that follows the commonly practiced naming conventions in that language. For example, the Java class name "QuickSort" will generate the keywords "Quick" and "Sort". Keywords extraction from comments is done by parsing the text for natural and meaningful keywords. These keywords are then mapped to the entities that have unique IDs and that are close to the comment, for example the immediate entity that follows the comment. These keywords are also associated with the immediate parent entity inside which the comment appears. The keyword-based query the user inputted retrieves using the two interfaces; keyword-based retrieval interface and semantic-based retrieval interface. Keywords are used generally in retrieving items from a vast information source, for instance a catalogue or a search engine. An index term or descriptor in information retrieval is a term that captures the essence of the topic of a document, and can consist of a word, phrase, or alphanumerical term. They are created by analyzing the document either manually with subject indexing or automatically with automatic indexing. Methods that are more sophisticated use keyword extraction. Keywords are stored in a search index. The semantic-based retrieval module reflects real world information connected to location information by a semantic relation. Each class has a semantic relationship to each other as well as the address. We use the semantic relationship between classes in the ontology for semantic-based retrieval.

We propose an architecture where customers can use a mobile device to retrieve information and objects using GPS coordinates of a current position. The user obtains current coordinates using the GPS module. The location mapping module maps the coordinates to the real address we called. The obtained address is used to query. Therefore, the user can select whether or not to use the address as the query. The results can be accurately overlaid to produce a composite map. This solution is for the mobile station with a GPS unit-based mixed-Web map interface using a location-based mobile mash-up with Google maps. Fig. 2 shows the collaboration diagram and interactions between location mapping module, keyword and semantic based retrieval modules.

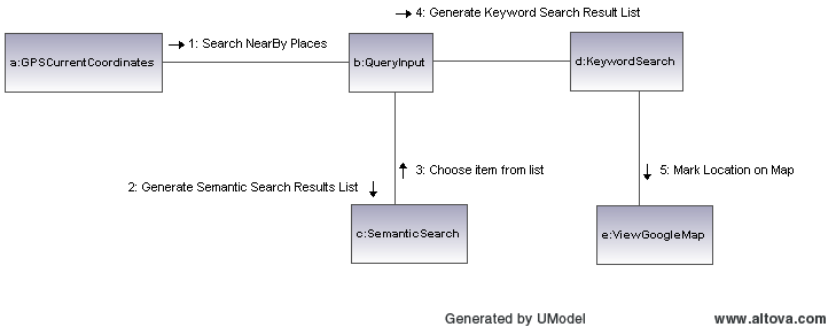


Fig. 2. Collaboration diagram of keyword-based retrieval module, semantic-based retrieval module and location-based retrieval module and their interactions

Fig. 3 shows the activity diagram of three active processes that flows after the user enters the input query. The first process is LBS and the flow begins with entering the input query and after obtaining the current coordinates using GPS module, the query transfers to server where the web service is running. Web server is second process with retrieval activity such as keyword and semantic retrieval and the generated result list is sending by xml. The LBS process accepts the result list and displays on mobile device where the user can pick the result item and can choose one of the options: Check result or Display the Retrieval Result using Google Maps. Searching for Google Maps activate the third process and generation of wanted map is the last step of this activity.

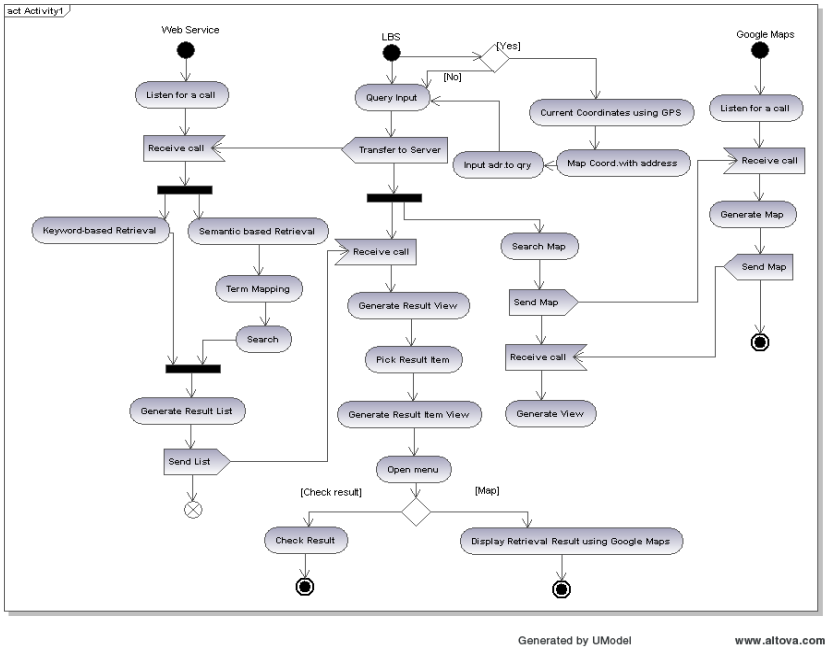


Fig. 3. Activity diagram of location-based system for retrieval using mobile device with three active processes

4 Conclusion

The proposed system introduced the need to offer a mobile location based retrieval solution combined with a location mapping module. Therefore, we made a brief description of the different technology alternatives that exist for geo coding, geographic search engine and semantic web, as well as for mobile information

retrieval and semantic search. Then, we defined what mobile location based systems are and what the issues are today.

Finally, we tried to propose an architecture that would resolve some of the current issues learned from failures in mobile location based systems for keyword and semantic retrieval. The objective is to combine all the good things about mobile technology while trying to avoid the shortcomings that come with it. In conclusion, mobile location based systems for keyword and semantic retrieval is very important for often using the mobile devices, especially when the input query, for example is hospital where the correct address has great significance.

References

1. J. Roth. Flexible positioning for location-based service. IADIS International Journal on WWW/Internet. 1 (2):18-32.
2. A. Markowetz, Y. Chen, T. Suel, X. Long, B. Seeger. Design and Implementation of a Geographic Web Search Engine. Technical Report TR-CIS-2005-03, CIS Department, Polytechnic University. 2005.
3. C. D. Tomlin. Geographic Information Systems and Cartographic Modeling, Prentice Hall. 1990.
4. S. Yokoji, K. Akahashi, N. Miura. Kokono Search: A Location Based Search Engine. 10th International World Wide Web Conference (WWW10). Hong Kong. 2001.
5. K. McCurley. Geospatial mapping and navigation of the web. In Proc. of the 10th World Wide Web Conference. 2001; 221–229.
6. L. Gravano. Geosearch: A geographically-aware search engine. 2003. Available at: http://geosearch.cs.columbia.edu.
7. J. Ding, L. Gravano, and N. Shivakumar. Computing geographical scopes of web resources. In Proc. of the 26th VLDB. 2000; 545–556.
8. A. Daviel. 1999. Available at: http://geotags.com.
9. M. Egenhofer. Toward the semantic geospatial web. In Proc. of the 10th ACM GIS. 2002; 1–4.
10. DCMI Usage Board. Dublin Core Qualifiers. Recommendation of the DCMI, Dublin Core Metadata Initiative. 2000.
11. A. Daviel. Geographic registration of HTML documents. IETF Draft. 2003. Available at: www.geotags.com/geo/draft-daviel-html-geo-tag-06.html.
12. Open GIS Consortium. http://www.opengis.org
13. P. Coppola, V. D. Mea, L. Di Gaspero and S. Mizzaro. The Concept of Relevance in Mobile and Ubiquitous Information Access. Mobile and Ubiquitous Information Access, F. Crestani et al. (Eds.). Springer-Verlag Berlin Heidelberg. 2004; 1–10.
14. P. Coppola, V. D. Mea, L. Di Gaspero, S. Mizzaro, I. Scagnetto, A. Selva, L.Vassena, P. Z. Rizio. MoBe: Context-Aware Mobile Applications on Mobile Devices for Mobile Users. In: Proceedings of ECHISE'05, held in Conjunction with PERVASIVE'05. Munich. 2005; 49–54.
15. G. J. F. Jones and P. J. Brown. Context-Aware Retrieval for Ubiquitous Computing Environments. Mobile and Ubiquitous Information Access, F. Crestani et al., (Eds.). Springer-Verlag Berlin Heidelberg. 2004; 227–243.
16. I. F. Akyildiz, S. Mohanty and X. Jiang. A ubiquitous mobile communication architecture for next-generation heterogeneous wireless systems. Communications Magazine. IEEE. (43):29–36.

17.R. Guha, M. Rob and M. Eric. Semantic search. Proceedings of the 12th international conference on World Wide Web. 2003; 700–709
18.N. W. Group. Hypertext transfer protocol – http/1.1. 1999
19.Q. Zhou, C. Wang, M. Xiong, H. Wang and Y. Yu. SPARK: Adapting Keyword Query to Semantic Search. ISWC/ASWC 2007, K. Aberer et al. (Eds.). Springer-Verlag Berlin Heidelberg. 2007; 694–707

Next-generation DNA sequencing technology, challenges and bioinformatics approaches for sequence alignment

Aleksandra Bogojeska, Slobodan Kalajdziski, Ljupco Kocarev

Department of Computer Science and Informatics,
Faculty of Electrical Engineering and Information Technologies,
Karpos II bb, 1000 Skopje
{aleksandra.bogojeska,skalaj,lkocarev}@feit.ukim.edu.mk

Abstract. The advent of high-throughput sequencing platforms brought bioinformatics to a new level. This, so called 'next-generation' sequencing technology opened the researching doors of every laboratory allowing accomplishment of previously unimaginable scale and expensive experiments. As a result, novel research areas have emerged providing huge amounts of new data ready to be analyzed. Parallel to this progress, a variety of sequencing tools designed for data analysis has been published. Sequence alignment takes the central challenge in data analysis, providing primary representative results for the experiments. Few alignment methods and diversity of tools have been published and developed in the last years. The main goal of all these alignment tools is to fit between performance and accuracy. In this review will be presented the new NGS technologies and platforms, the current alignment approaches applied in data analysis and described some commonly used implementations of the methods.

Keywords: next-generation sequencing, alignment algorithm, short reads

1 Introduction

The process of determining the order of the nucleotides in a molecule of DNA, is called DNA sequencing. Novel approaches in genome sequencing alternate the genomics research rapidly supplanting the former Sanger method of sequencing. This next-generation sequencing (NGS) technology offers low cost and accurate DNA sequencing, which led to appearance of new fields in bioinformatics and molecular biology and advance in the existing ones. Research areas which have emerged and benefit from the use of NGS are metagenomis, Single Nucleotide Polymorphisms (SNPs) detection, gene expression, chromatin immunoprecipitation (ChIP) sequencing, non-coding RNAs discovering, *de novo* and ancient DNA sequencing [1], [2]. Each of these fields is expected to provide new and significant information for future development and research in genomics.

There are two fundamental computational analyses performed over the data after the process of sequencing: assembly and alignment. The assembly is essential for organisms without a sequenced reference genome and represents the process of joining the sequenced reads in whole genome sequence. The alignment on the other hand, remains fundamental analysis that confirms the success of the experiment. This

process will be analyzed later in details. In the end the both types of analysis end with graphical tools for data viewing and representation.

The paper is organized as follow. In Section II the NGS technologies and the sequencing platforms using this technology are going to be presented. Section III gives a closer look to the sequence alignment methods and tools developed for the NGS. The paper is concluded in section IV and future directions for the NSG technology and development of bioinformatics tools are given.

2 Next Generation Sequencing Technologies

Since 2004 when the first NGS platform was presented by Live Sciences, Roche 454, many new approaches and producers of NGS platforms have appeared. For the time being, there exist 8 producers of NGS platforms: Applied Biosystems (www3.appliedbiosystems.com), Complete genomics (www.completegenomics.com), Helicos (www.helicosbio.com/), Illumina (www.illumina.com), Polonator (www.polonator.org/), Roche (/www.roche.com/index.htm), Pacific Bioscience (www.pacificbiosciences.com/) and Ion Torrent (www.iontorrent.com/).

Each of the platforms embodies a complex symbiosis of enzymology, chemistry, high-resolution optics, hardware and software engineering. They are characterized with short read lengths, massive volumes of data generated and low cost sequencing. The NGS platforms produce reads with length between 30-400base pairs(bp) and generate 500(Mega base pairs)Mb-200(Giga base pair)Gb per run, compared to the Sanger method where maximum 70(kilo base pairs)kb per run are being generated with reads length of 900bp. The mentioned characteristics for the listed platforms are presented in Table 1.

Table 1 NGS platforms and their features.

Feature	Technol.	Read Len.(bp)	Data. Gen per run	Time per run	Accuracy
Roche454	Sequencing-by-synthesis	400	500Mb	10h	M:99.5% C:99.99%
Illumina	Revirsible dye termination	2x100 1x35	150-200Gb 18-35Gb	8days 1,5days	/
ABiSOLID	Sequencing by ligation	50	100Gb	6-7days	M:99.94%[1] C:99.999%[2]
Complete genomics	Sequencing by hybridiyaton and ligation	70	8Gb	/	99.95%
Pacific	Sequencing-by-	446	10bases	10-	99.9999%

[1] M – measured accuracy, produced from the sequencing platform
[2] C – consensus accuracy, gained by the alignment tool

Biosciences	synthesis (single mol.)		per sec	15min	
Helicos	Sequencing-by-synthesis (single mol.)	25-55	21-28Gb	8days	C:99.995%
Polonator	Polony sequencing by ligation	26	4-5Gb	4days	98%
Sanger	Capillary sequenciung	800	70kb	3h	99.9%

All platforms mainly differ in the approach of DNA manipulation; they use amplification of DNA molecule or single DNA molecule. The amplification of the molecules in the NGS is done *in vitro* using polymerase chain reaction (PCR). Parallelism is achieved with executing PCR on multiple individual molecules of DNA. This method is used in Illumina, Roche 454, Polonator, IonTorrent and ABISOLiD platforms. For the time being, only Helicos and Pacific Biosystems instruments are regarded as 'single molecule' sequencers. It is expected that the single molecule sequencing will produce read lengths of thousands of bases that will provide simplified and improved data analysis.

The methodology used in the NGS platforms differs in the way of detection and reporting nucleotides. Roche platforms use pyrosequencing method combined with emulsion, Illumina platforms use reversible dye terminators for bases distinguishing, and ABI platforms use sequencing by ligation method. The new Ion Torrent sequencing platform uses pyrosequencing where later the released hydrogen ion is detected by a special semiconductor chip.

Detailed review of the chemistry behind these platforms appears elsewhere [1],[2],[3], as on the companies` websites supplied with up-to-date information for each of the platforms.

3 Alignment, methods and tools used for NGS data analysis

Alignment represents the process of determining the source of the sequenced DNA read. The read can be mapped against a given reference genome or multiple genomes from the species the sequence has come from. The alignment process can also be applied to other genomes, assuming that the evolutionary distance between the species of reads and the genome is appropriate.

The foregoing NGS technologies used for DNA sequencing and their specific characteristics lead to development of new alignment tools. The gold standard tool BLAST used for analysis of sequencing data produced by the Sanger technology and the former generation of programs requires alignments of protein sequences used for performing search through large databases to find the best matched sequences [5]. On the other hand the new alignment tools perform alignment against the genome references of the species of interest. Also, these tools will have to deal with the use of many various technologies with unique error model which has to be implemented in the algorithm design; the specific species rate of polymorphisms that has to be calculated in the design of expected number of mismatches during the alignment.

These design assumptions will result in faster algorithms that will be capable of accurate processing of the massive data volumes produced by the NGS technologies.

In the recent year there have been a growing number of implementations for tools that perform short-read alignment, but the number of significant methodologies implied is much smaller. The alignment tools can be grouped by the methodology used into three categories: hash table-based algorithms, algorithms based on suffix trees together with their modifications and merge sorting based algorithms. There is only one implementation for the third category, the Slider tool [24]. Accordingly this review will focus on the first two techniques. The first discussed method is hash table-based implementation, where one has two possible ways of index creation, using the reference genome, or using the set sequence reads. Also the Burrows Wheeler transform (BWT)-based algorithms will be presented where first an efficient index of the reference genome is created which later results in fast search that has low-memory footprint.

In order to give accurate mapping of the sequence, alignment programs are following a multistep algorithm. In the first step they use heuristic techniques to find the most likely places in the reference where the read can be mapped. After, on this smaller subset of possible mapping locations more accurate and sensitive algorithms are run, like the Smith-Waterman local alignment algorithm [6] and its modifications, giving the top *n* places where the read is mapped against the reference.

3.1 Hash-based alignment methods

The first alignment tools used for the NGS short reads developed and presented, used the same methodology of creating the searching index as the BLAST [5] generation of algorithms - a hash table. This hash table is created form the input query data and is used for structuring the index and scanning trough the database sequences. This method is appropriate for DNA sequencing data where most of the time one has duplicate sequence and all the possible combinations of nucleotides are unlikely to be present. Those two features of the data match the feature of hash tables to index complex and non-sequential data.

The hash table index can be created from the reference genome or the input reads. The difference in the approach is in the gain and loss of the memory and time. The hash tables created from the reference genome have constant memory requirement regardless of the size of the input reads. This memory is usually large depending on the size of the reference genome. Hash tables based on the sequenced reads usually require smaller and variable memory requirements but have slower processing time to scan the entire reference genome against every input read. Again, the memory depends on the number and diversity of the reads.

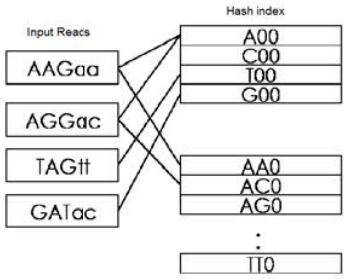


Fig. 1 Hash table-based method. The regions used for seeding are marked with capital letters and the matched reads for the seeds 100 and 110 are shown

Algorithms that use the hash table-based method are: MAQ [7], SOAP [8], SHRiMP [9], SSAHA2 [10], RMAP [11], RazerS [12], SeqMap [13], ZOOM [14], BFAST [15], MOSAIK (http://bioinformatics.bc.edu/marthlab/Mosaik) and the commercial ELAND Illumina's algorithm.

These hash table-based algorithms are implementing the spaced seed paradigm [16] to allow alignment mismatches and gaps. A template in form, example, '101101', represents a seed that requires 4 matches at the '1' position and 2 mismatches at the '0' positions. In Fig. 1, there is one seed 100 and other 110, where the capital letters in the input reads represent the seed region and in the hash index the matched seeds. An improvement of the spaced seed policy is the q-gram approach [17], where multiple spaced seeds per read are used for gaps detection. I am going to give short overview of MAQ and SOAP alignment tools, their hash index implementations, mismatches handling, the algorithm advantages and disadvantages.

MAQ

MAQ [7] (Mapping/Alignment Quality) presents an alignment and variant caller tool that uses hash table for data indexing. Additional characteristic feature for MAQ is the consideration of the quality scores (Phred values) of the reads and later according to this information and number of mismatches it assigns mapping quality value to each mapped read. This feature helps in the process of distinguishing between platform error and true SNP. The algorithm always reports one alignment, and the repeated read is aligned randomly with reporting mapping quality of zero.

To implement the search, MAQ uses multiple hash tables created from the input reads whose number depends on the number of mismatches allowed trough the alignment. By default six hash tables are created allowing up to 2 mismatches in a read using the spaced seed technique. The number of mismatches can be greater, increasing the number of hash tables and processing time. MAQ guaranties to find alignments up to two mismatches in the first 28bp of the reads. It requires smaller memory, less then 1Gb per processor, but longer time to achieve accurate results. Along with the alignment MAQ gives many other possibilities for short-read data analysis as different format conversion, SNPs detection and alignment viewer.

MAQ is capable of processing Illumina and ABISOLiD reads, single or mate pair, no longer of 63bp, which is the greatest disadvantage of the algorithm, as each of the technologies aims to produce longer reads.

SOAP

SOAP [8] (Short Oligonucleotide Analysis Package) is another alignment tool designed for short-reads mapping. It is also hash table-based but does not use the quality information for alignment. To achieve the alignment SOAP uses seed and hash table searching algorithm. Not like in MAQ here the hash table index is created from the reference sequences. The input reads and the reference genome are being encoded to numeric data using 2 bits per base. This value is used as suffix, when searching in the look-up table is performed, in order to find the number of different bases. It reports the alignments with minimal number of mismatches or smaller gaps reported. By default 2 mismatches and gaps between 1 and 3bp are allowed. The number of 2 mismatches allowed is achieved with splitting the read in four fragments, and all 6 combinations of the two fragments where mismatches can exists are taken as a seed. To remove the contaminated end regions of Illumina reads, SOAP can remove few base pairs from the end of the reads. To gain lower memory SOAP loads the reference sequences as unsigned 3-bytes array.

SOAP is also capable of mate pair alignment, but not of color space alignment. It is intended for alignment of short reads up to 60bp. Additional features of the algorithm are specific alignment for mRNA, and small RNA reads.

3.2 Suffix array, FM-index and Burrows-Wheeler transform alignment methods

This category of alignment algorithms has been developed in the recent years and connects the suffix array methods and special FM-index. The connection is done using a data structure that origins form compression theory, Burrows-Wheeler transform. The combination of these methods gives algorithms that perform reads matching 10 times faster than the hash table-based aligners [18], [19].

A suffix array represents a data structure that contains all possible suffixes of a string. It's designed for efficient searching of a large text. Burrows-Wheeler transform [4] is a technique for string compression which implements transformation algorithm and having the resulting compressed string one can easy decode the input string. It uses so called zero string to identify the string end. A table of all possible suffixes of the string is created, and sorted. The last column of the sorted matrix represents the BWT string, Fig. 2. Afterwards using the LF (last-to-first) mapping the initial text can be obtained. These features make BWT suitable for genetic data manipulation where the strings are consisted of four-letters alphabet. The FM(Full-text, Minute space)-index [22] is a data structure based on the BWT and presents an efficient algorithm for data indexing. The creators of FM-index, Ferragina and Manzini, propose special index structure consisted of three parts: superbuckets section (SB), bucket directory (BD), and body of the FM-index. The superbuckets section stores the number of

occurrences of every character in the previous SB, the bucket directory stores the starting position of each compressed bucket in the body of the FM-index, and the body of the index stores compressed image of each bucket. The data structure creation includes two steps. The first step is performing BWT on the reference genome. This process is reversible and within this step sequences that exist multiple times will appear together in the data structure. The second step, memory intensive, represents the final index creation given by the FM-index structure. This data structure reports faster searching time having the same or smaller size index than the input genome size. For the human genome index approximately requires 2-5GB depending on the other algorithm techniques implemented [18-20]. This small index size allows storing the index on disk and loading into memory on any standard computing cluster.

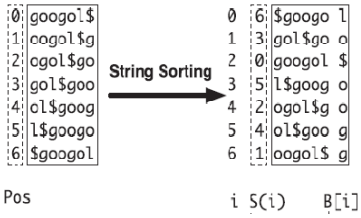


Fig. 2 BWT construction for the string GOOGOL [30]

The tradeoff between speed and algorithm sensitivity in the sense of BWT based algorithms is the number of mismatches allowed. Whereas in the hash table-based methods the seeds were present, here there is no efficient way of dealing mismatches. Each algorithm employs different method for mismatches and gaps detection and management. As the sequencing technology develops and becomes more accurate this limitation will become less important for species with low rate of polymorphisms.

Next I am going to give short description of three BWT-based methods, founders of this generation algorithms.

BWA and BWA-SW

These two algorithms are published by the same developer of the MAQ aligner, using the speed up that BWT gives. The BWA [18] version is intended for short reads up to 200bp length, and the enhanced version BWT-SW [21] is for long reads up to 100kbp. They both use suffix array combined with the BWT and FM-index. As in MAQ the mapping quality is reported for each aligned read.

The BWA algorithm uses few other concepts for Illumina platform reads alignment, considering their characteristics: ambiguous bases, paired-end mapping, determining the maximum allowed number of mismatches and generation mapping quality scores. When comes to exact matching the FM-index backward search is used. For the inexact matching a breadth-first search algorithm (BFS) with a heap-like structure for calculating differences between strings is used. This algorithm will find all alignments within a given *n* number of differences. Acceleration is achieved by using iterative strategy for the top repetitive intervals and unique top intervals with *n*

differences. It is 6-18 times faster than MAQ but reports slightly bigger error rate than MAQ, where one can see the tradeoff between speed and accuracy.

NGS platforms development results in producing reads with 500-1000bp lengths in order to increase the read length. The limitations of BWA to deal with longer data, and data with high error rates, leaded to the development of BWA-SW algorithm. This algorithm is the only one that can process reads with length longer than 1kb, using the BWT. Compared to the short-reads alignment where full-length alignment of the read is required, the long-read alignment prefer local matches because these reads are more likely to contain big number of mismatches and gaps. BWA-SW builds FM-indices for the both reference genome and read. The reference genome is represented as a prefix tree, and the read as a prefix directed acyclic word graph (DAWG). Using dynamic programming the best local match is found. Mismatches and gaps are handled using seeds templates between the two FM-indices. This algorithm reports 5-20 times faster processing time compared to SSAHA2 and BLAST, and same accuracy as SSAHA2 alignment algorithm.

BOWTIE

Bowtie [19] is the first published short-read alignment algorithm that uses the BWT technique. It uses index structure proposed by Kärkkäinen [23], combined with the FM-index structure, which can be configured to tradeoff between speed and memory. For exact alignment pure LF mapping is used. In order to handle sequencing errors and polymorphisms Bowtie introduce two extensions: quality-aware backtracking algorithm and double indexing strategy. The first extension allows mismatches and favors the high-quality alignments using greedy, depth-first search (DFS) to find the best alignment with n mismatches. The second extension avoids unreasonable backtracking for reads with low quality values. Using the MAQ strategy mapping quality is reported, too. Disadvantage is that Bowtie is not able to report gapped alignments, i.e. detection of insertions and deletions. Recent Bowtie versions are able to perform pair-end and color space alignment and process reads up to 1024bp length.

SOAP2

SOAP2 [20] is the improved version of the SOAP alignment algorithm with reduced computer memory usage and increased alignment speed. These improvements are achieved using the BWT compression index instead using the hash table-based seed algorithm. It uses the proposed FM data structure, for example a 13mer on the hash will result in 2^{26} blocks for the reference genome, and in few interactions the exact location of the search string can be found inside the block. For mismatches dealing a 'split-read strategy' is used, same as with the SOAP algorithm. The read is spitted in fragments in order to detect mismatches. As the previous version of this algorithm, the best alignment is reported by the minimal number of mismatches and gaps. Improvement is made in the maximum length of the input reads, where now the algorithm can handle reads up to 1024bp. Using the BWT this algorithm is 12 times faster in the step of index creation for the human genome compared to the first

version, and 20 times faster for the same amount of reads aligned. The SOAP group of algorithms is evolving and there are companion algorithms for *de novo* assembly and SNP's detection.

4 Conclusion and NGS future

Bioinformatics and genomics science has shifted as result of NGS technologies advancement. New projects inspired by the NGS, 1000 Genome Project (http://1000genomes.org), HapMap (http://hapmap.org), will provide significant information for genome structure and functionalities, and open new perspectives in specific diseases and cancer treatment. The bioinformatics community is being intrigued and gives fast response to the challenges coming from the NGS projects. As a consequence, the existing alignment methods have been improved and adjusted to deal with the massive volumes of short-reads data. The competition between alignment tools is still running and there is no answer provided for the question which tool is the most accurate and suitable to use, offering most effective use of computational resources. Table 2 gives overview of the presented alignment tools.

The current NGS improvement, production of longer reads, requires modification of the many developed short-read analysis tools or development of new ones. Long reads will have the primary use when *de novo* assembly is performed, or genomes with high repetitive structure are sequenced. On the contrary, the short-read sequencing will play the main role when sequencing on specific small regions is required, as ChIP sequencing, RNA sequencing and SNPs detection.

Table 2 NGS alignment tools and their features.

Tool	MAQ	SOAP	BWA	SOAP2	BWA-SW	BOWTIE
Method	Hash	Hash	BWT	BWT	BWT	BWT
Platform	Illumina, ABiSOLD	Illumina	Illumina ABiSOLID	Illumina	All	Illumina Roche
Indels	y	y	y	y	y	n
Read Len.	63bp	60bp	200bp	1024bp	100kbp	1024bp

NGS technology is in an early stage of development and the following years will bring improvements and novelties in this researching area and continuing stimulus for researches in bioinformatics. Developers will be continuously challenged as new data types will be presented from new and enhanced NGS platforms. The future development of analysis tools and management systems will have to incorporate information about sequence errors, biases, genome polymorphisms rate. The methods and their algorithm representations described above provide the first approach to the existing and upcoming challenges in the sequencing field.

Acknowledgments. We are thankful to Zlatko Trajanoski and Gernot Stocker from the Institute of Genomics and Bioinformatics, Graz, Austria, for their unselfish sharing of information, resources and experience in the field of bioinformatics.

References

1. Mardis, E.R.: The impact of next-generation sequencing technology on genetics. Trends Gnet. 24,113--141 (2008)
2. Mardis, E.R.: Next-generation DNA sequencing methods. Annu. Rev. Genomics. Hum. Genet. 9, 387—401 (2008)
3. Voelkerding, K., Dames, S.A., Durtschi, J.D.: Next-Generation Sequencing: From Basic Research to Diagnostics. Clin. Chem. 55, 641--658 (2009)
4. Burrows, M., Wheeler, D.J.: A block-sorting lossless data compression algorithm. Tech. Report 124, Digital Equipment Corporation. CA: Palo Alto (1994)
5. Altschul, S.F., Gish,W., Miller, W. *et al.*: Basic local alignment search tool. J. Mol. Biol. 215, 403--410 (1990)
6. Smith,T.F., Waterman, M.S.: Identification of common molecular subsequences. Mol. Biol. 147, 195--197 (1981)
7. Li, H., Ruan, J., Durbin, R.: Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome Res. 18, 1851-1858, 2008
8. Li, R., Li, Y., Kristiansen, K., Wang, J.: SOAP: short oligonucleotide alignment program. Bioinformatics. 24, 713--714 (2008)
9. Rumble, S.M. *et al.*: SHRiMP: accurate mapping of short color-space reads. PLOS Comput. Biol. 5, e1000386 (2009)
10. Ning, Z., Cox, A.J., Mullikin, J.C.: SSAHA: a fast search method for large DNA databases. Genome Res. 11, 1725--1729 (2001)
11. Smith, A.D., Xuan, Z., Zhang ,M.Q.: Using quality scores and longer reads improves accuracy of Solexa read mapping, BMC Bioinformatics, 9, 128 (2008)
12. Weese, D., Emde, A.K., Rausch, T., Doring, A., Reinert, K.: RazerS--fast read mapping with sensitivity control. Genome Res. 19, 1646--1654, (2009)
13. Jiang, H., Wong, W.H.: SeqMap: mapping massive amount of oligonucleotides to the genome. Bioinformatics. 24, 2395--2396 (2008)
14. Lin, H., Zhang, Z., Zhang, M.Q., Ma, B., Li, M.: ZOOM! Zillions of oligos mapped. Bioinformatics, 24, 2431--2437 (2008)
15. Homer, N., Merriman, B., Nelson, S.F.: BFAST: an alignment tool for large scale genome resequencing. PLoS One, 4, e7767--e7767 (2009)
16. Ma, B., Tromp, J., Li, M.: PatternHunter: faster and more sensitive homology search. Bioinformatics, 18, 440--445 (2002)
17. Rasmussen, K.R., Stoye, J., Myers, E. W.: Efficient q-gram filters for finding all epsilon-matches over a given length. J.Comput Biol, 13, 296--308 (2006)
18. Li, H., Durbin, R.: Fast and Accurate Short Read Alignment with Burrows-Wheeler Transform. Bioinformatics, 25, 1754--1760 (2009)
19. Langmead, B., Trapnell, C., Pop, M., Salzberg, S. L.: Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. 10, R25
20. Li, R., Yu, C., Li, Y., Lam, T. W., Yiu, S. M., Kristiansen, K., Wang, J.: SOAP2: an improved ultrafast tool for short read alignment. Bioinformatics. 25, 1966--1967 (2009)
21. Li, H., Durbin, R.: Fast and accurate long read alignment with Burrows-Wheeler transform. Bioinformatics. 26, 589--1595 (2010)
22. Feragina, P., Manzini, G.: Opportunistic data structures with applications. Proceedings of the 41st Symposium on Foundation of Computer Science (FOCS 2000),Redondo Beach, CA, USA. 390--398 (2000)
23. Kärkkäinen, J.: Fast BWT in small space by blockwise duffix sorting. Theor. comput. Sci. 387, 249--257 (2007)
24. Malhis, N., Butterfield, Y., Easter, M., Jones, JMS.: Slider-maximum use of probability information for alignment of short sequence reads and SNP detection. Bioinformatics. 25,6--13 (2009)

Localization in Wireless Sensor Networks

Ustijana R. Shikoska[1] , Dancho Davchev[2]

[1] University for Information Sciences & Technologies „Sv. Apostol Pavle“ – Ohrid, Republic of Macedonia
ustijana@t-home.mk
[2] University „Sv. Kiril i Metodij“, Faculty of Electrical Engineering – Skopje, Republic of Macedonia
etfdav@feit.ukim.edu.mk

Abstract. Due to the close integration of sensor networks with the real world, the categories time and location are fundamental for many applications of sensor networks, to interpret sensing results or to coordinate sensor nodes in Wireless Sensor Networks (WSNs). Time synchronization and sensor node localization are fundamental and closely related services in sensor networks.The most relevant techniques for Localization in Wireless Sensor Networks - lateration, angulation and scene analysis are discussed. A comparison is made between several localization systems. A choice for a certain system or technique depends on the intended application. A possible system for determining the position of a moving device is given. The system's possible advantage over most other existing systems is short deployment time in a new environment and reasonable accuracy. It could be used for indoor and outdoor environments. Improvements can be made with respect to accuracy and mobility.

Keywords: wireless sensor networks, space-time localization, localization methods

1 Introduction

A Wireless Sensor Network (WSN) is a network of many small sensing and communicating devices called sensor nodes. Each node has a CPU, a power supply and a radio transceiver for communication. Interconnection between nodes is achieved via transceiver. A WSN contains one node, the base station, which connects the network to a more capable computer and probably to a network of general purpose computers through it. Sensors attached to these nodes allow them to sense various phenomena within the environment.

The typical purpose of a sensor network is to collect data via sensing interfaces and propagate those data to the central computer, allowing easy monitoring of an environment.

A node is capable of dealing with a variety of jobs. The nodes currently available are battery-operated, they have limited life-time. Memory capacity of a node is also limited.

Life-time, processing and storage restrictions directly affect the algorithms designed for sensor networks. A routing algorithm for WSNs must be energy and memory efficient. Since radio transmissions consume a significant amount of energy, researchers generally seek ways to reduce radio communication as much as possible. When more information is stored and more computation is done as to reduce the communication costs, energy consumption of the processor and memory components are becoming an important issue. Design choices have to be made, they also depend on the intended application.

Currently there is a prototype of a system available, developed using motion sensors to secure an area [1] based on the Smart Dust concept. The idea of the system is to monitor an area or room by a network of sensors with the size of a dust particle. The Smart Dust project is exploring whether an autonomous sensing, computing and communication system can be packed into a cubic-millimeter mote, to form the basis of integrated, massively distributed sensor networks. In the prototype, the size of a sensor is significantly bigger than a dust particle. The moment a sensor detects movement in the area a message is sent to a central server.

The server processes the data and then uses Google Maps to produce a map which shows the detected movement. A GPS receiver is used to determine an absolute position, while RSSI (Received Signal Strength Indicator) is used to locate the sensors relative to the GPS receiver. RSSI uses the decrease in energy of the radio signal as it propagates in space to estimate the distance. Experimentation with the prototype system shows this method becomes unreliable when the batteries of the sensors are getting weaker [1]. Using GPS receivers for all sensors is not an option as GPS cannot function in indoor and many outdoor applications, especially when there is no direct line of sight from nodes to terrestrial satellites. Usage of these devices on sensor nodes is still a challenging issue due to their size, energy and price constraints. As a result, there is a need for reliable localization in WSNs without using of GPS receivers.

A Wireless Sensor Network is deployed to monitor its environment and for disaster response and recovery systems. Applications include health monitoring systems, monitoring of wildlife habitats [2] and nature reserves such as the Great Barrier Reef [3], and forest fire detection systems [4,5,6]. Examples of military applications are battlefield surveillance [7,8] and the previously mentioned securing of an area or room.

An accent is put to localization in mobile indoor WSNs. Localization can be used for tracking objects or people. It is helpful to people navigating indoors where GPS is not available. Mobile devices such as laptops may be tracked within a building in order to locate them easily. Location dependent network services, with application examples ranging from building automation to targeted advertising or augmented reality, also require reliable localization techniques [9].

2 Related work

All the approaches discussed next in this paper are range-based, because the accuracy of range-free algorithms is often limited by requiring dense deployments of sensor nodes. The tracking of moving devices has been studied by Smith et al. [24] under an active mobile and a passive mobile infrastructure using the Cricket location system. Cricket uses the time difference in arrival of concurrent radio and ultrasound signals to estimate distances. In the active mobile architecture, the mobile device actively chirps, and the fixed infrastructure nodes then reply either over a radio channel or a cabled infrastructure, reporting the measured distances to the mobile device or some central processor. In the passive variant, the infrastructure has beacons that periodically transmit signals to a passively listening mobile device, which in turn estimates distances to the beacons.

Priyantha et al. [27] note it is almost impossible to deploy nodes in a typical office or home to achieve sufficient connectivity across all nearby nodes. It is hard to obtain ranging between nodes placed inside and outside a room in a standard building. Capkun et al. introduce the Self-Positioning Algorithm (SPA) [32]. SPA defines and computes relative positions of nodes in a mobile ad-hoc network without using GPS. It is a distributed algorithm that does not use nodes with fixed or known positions. It assumes some method to estimate the distances between nodes and builds a relative coordinate system. A simulation with 400 nodes was performed by the authors. The nodes follow a random movement pattern: they move using a random velocity, wait for a fixed time, and then move again. It is shown that if a larger (three-hop) neighborhood is used instead of a two-hop neighborhood, the mobility of the center of the network decreases. No accuracy information has been provided - reducing the position error is being mentioned as subject of future work.

As the algorithm is focused on providing location information to support basic network functions, accuracy requirements should not be high. Communication costs are relatively high in multi-hop networks as the algorithm requires a broadcast to all the nodes in the network. An Online Person Tracking (OPT) system for an indoor environment is presented by An et al. [20]. OPT employs a passive mobile architecture. The average RSSI of 200 measurements was used to estimate the location. An et al. only used the three strongest received signal strengths because they

claim using more does not guarantee a higher accuracy. Experiments with a static sender and receiver were performed to measure the influence of the antenna orientation on the strength of the received signal. The RSSI value varied up to 15 dBm depending on the antenna's orientation. This leads to bigger error on distance estimation when two motes are farther apart, because the variation in RSSI becomes smaller as the distance becomes larger. The authors applied a bounding box algorithm to select an area in which the optimal position was sought. If there was no overlapping area of circles, the estimation area was expanded to make sure that the potential target position was included in the search area (Figure 1).

Fig. 1. Boundary selection without overlapping area.

The Minimum Mean Square Error (MMSE) algorithm was employed for target location estimation within the selected area. This method is commonly used in statistics and signal processing. In the conventional MMSE method (dubbed C-MMSE) all range estimates were given the same importance in minimizing the position error. A weighted version (W-MMSE) is also proposed by the authors. The higher the slope of the empirical curve between distance and RSSI is, the higher is the assigned weight. Higher RSSI values are considered to be more reliable than low ones.

Lee et al. [9] present an algorithm enabling localization of moving wireless devices in an indoor setting. An active mobile infrastructure is employed; a burst of five packets in 50 ms is sent by the mobile node every 0.6 seconds. Ten nodes were deployed at fixed locations and one mobile node was being localized.

The mobility of the users is modeled by learning a function which maps a short history of signal strength values to a 2D position. During the training phase, ground truth locations of the mobile user are required; however, locations of infrastructure nodes are not needed. The authors used radial basis function fitting to learn a reliable estimate of a mobile node's position given its past signal strength measurements. RSSI measurements were filtered by a box filter and then fed into the learned function to obtain the position of the mobile node.

Nine different trajectories were evaluated: five for training and four for testing. An area of approximately 30 m x 25 m was used for experimentation. Experimental data

shows that the variance due to reflections is particularly severe when either transmitter or receiver was moving, even at low speeds. Several parameters of the algorithm were optimized. The number of past measurements determines how much historical information about the trajectories is available. Using four past values was found to be optimal. In 50% and 97.5% of the cases the accuracy is 1.0 m and 4.1 m, respectively. Since they use past measurements at fixed time intervals, the authors implicitly assume that the speed of the mobile user at a given position is similar during training and localization. Explicitly handling speed differences is subject of future work. In Table 1 the localization systems are compared with respect to accuracy, node density and technique. The table is partially based on a comparison by Kaseva et al. [25], but some values have been corrected after having carefully reviewed the cited papers. The Self-Positioning Algorithm is not included in the overview because it has only been tested in simulations and no accuracy measurements were provided. Although trajectory matching uses RSSI measurements it is considered to perform scene analysis as it requires training for a specific environment.

Table 1. Comparison of Localization Algorithms

Localization system	**Accuracy (*m*)**	**Anchor node density (m^2 per node)**	**Technique**
Ferret [17]	*0.6-1.0 (A)*	*2-4*	*RSSI*/potentiometer
Cricket [24]	*0.02-0.2 (M)*	*2*	Ultrasound time-of-flight
MoteTrack [26]	*2 (M)*	*87*	Scene analysis
RADAR [33]	*2.9 (M)*	*326*	Scene analysis
Online Person Tracking [20]	*2/3.8 (M)*	*8/48*	*RSSI*
Trajectory Matching [9]	*1.0 (M)*	*52*	Scene analysis

A number of comments should be made to put the accuracy of the different systems into perspective. The anchor node density is defined as the number of square meters one anchor node has to cover. It is important for making a comparison because the lower the value, the easier is to obtain a relatively high accuracy. RADAR is an exception in the sense that it uses WLAN technology for anchor nodes as opposed to sensor nodes. In Ferret between five and eleven nodes are used, which explains the variation in accuracy and node density. The accuracy level of Cricket depends on the mobile node's speed. OPT has been evaluated in a corridor and office rooms; as the corridor covers a smaller area and has no interfering walls, a higher accuracy is obtained. Cricket and the Trajectory Matching algorithm are the only systems having tested accuracy of moving nodes; MoteTrack, RADAR and OPT track devices which may change their location but need to be stationary for localization. Ferret and Cricket

were tested in only one room, while other systems were evaluated in office environments having multiple obstructions and realistic error sources. Which system is best depends on the application.

3 Localization

Localization algorithms can be categorized according to a number of different aspects [10, 11, 12]:

- Input data: range-free, range-based - Range-free localization algorithms simply rely on connectivity information. Range-based methods extract distance information from radio signals.
- Accuracy: fine-grained, coarse-grained - A location discovery algorithm should estimate sensor position accurately. Accuracy, or grain size, can be expressed as percentage of sensor transmission range, or in meters. The level of accuracy usually depends on range measurement errors. Range measurements with less error will lead to more accurate position estimates. Certain accuracy is the precision and it is expressed in a percentage. For example, some inexpensive GPS receivers can locate positions to within 10 meters for approximately 95 % of measurements. More expensive units usually do much better, reaching 1 to 3 meter accuracies, 99 % of time. The distances denote accuracy, the percentages precision. If there is less accuracy, there may be a trade for increased precision.
- Dynamics: mobile, fixed - In fixed networks, nodes can establish their location in the initialization phase. Their only task is to report events or relay information sent by other nodes. In mobile net¬works, nodes need to be aware of changes in their position and perhaps of position changes of other nodes. Systems provide more accurate location information when a node is at rest than when it is in motion: tracking a moving node is harder because the inevitable errors that occur in the distance samples are easier to filter out if the node's position itself does not change during the averaging process.
- Beacons: beacon-free, beacon-based - Nodes with known positions are called beacon or anchor nodes. Beacon-based algorithms usually produce an absolute location system where absolute positions of nodes are known - latitude, longitude and altitude. The accuracy of the estimated position is highly affected by the number of anchor nodes and their distribution in the sensor field. The ratio of beacon nodes to blind nodes is small. The location of a beacon node can be determined using an attached GPS device or by manual deployment.

Beacon-free algorithms do not make any assumptions regarding node positions. In this case, instead of computing absolute node positions, relative positioning is used in

which the coordinate system is established by a reference group of nodes. Each object can also have its own frame of reference [13].

- Computational model: centralized, distributed - If an algorithm collects localization related data from the network and processes the data collectively at a single station, then it is said to be centralized. If, on the other hand, each node collects partial data relevant to it and executes an algorithm to locate itself, then the localization algorithm is categorized as distributed. Locally centralized algorithms are distributed algorithms that achieve a global goal by communicating with nodes in some neighborhood only. The sensor network can be divided into local clusters, where each cluster has a head. All the range measurements in a certain cluster are forwarded to the cluster head, where computation takes place.
- Hops : single-hop, multi-hop - A direct link between two neighbor nodes is called a hop. When the distance between two nodes is larger than the radio range but there are other nodes that create a continuous path between them, the path is called a multi-hop path.

3.1 Wireless communication

As sensor nodes use electromagnetic waves to communicate with each other we need to understand the basics of how these waves propagate.

Signal Propagation - a signal emitted by an antenna travels in the following three types of propagation modes: ground-wave propagation, sky-wave propagation and Line-Of-Sight (LOS) propagation. MW and LW radio is a kind of ground-wave propagation, where signals follow the contour of the Earth. Shortwave radio is an example of sky-wave propagation, where radio signals are reflected by ionosphere and the ground along the way. Beyond 30 MHz, line-of-sight propagation dominates, meaning that signal waves propagate on a direct, straight path in the air. Radio signals of line-of-sight propagation can also penetrate objects, especially signals with frequencies just above 30 MHz [14].

Sensor motes support tunable frequencies in the range of 300 to 1000 MHz and the 2.4-GHz band. This means LOS propagation is dominant. The industrial, scientific and medical (ISM) radio bands were originally reserved internationally for the use of RF electromagnetic fields for industrial, scientific and medical purposes other than communications. They have become a part of the radio spectrum that can be used by anybody without a license in most countries.

Multipath Propagation - for visible light we are well aware of the following effects: shadowing, reflection and refraction. In general, electromagnetic waves are also subject to diffraction and scattering [14].

Radio communication is affected by the physical properties of waves; the combined effects may cause a transmitted radio signal to reach a receiver by two or more paths.

Shadowing and reflection occur when a signal encounters an object that is much larger than its wavelength. Though the reflected signal and the shadowed signal are comparatively weak, they in effect help to propagate the signal to spaces where line-of-sight is impossible [14]. Reflections occur from the surface of the earth and from buildings and walls.

- Refraction occurs when a wave passes across the boundary of two media [14]. Compare this to how sunlight refracts when it enters water.
- Diffraction occurs at the edge of an impenetrable body that is large compared to the wave¬length of the radio wave. When a radio wave encounters such an edge, waves propagate in different directions with the edge as the source [15]. Thus, signals can be received even when there is no line-of-sight path between transmitter and receiver.
- Scattering occurs when the medium through which the wave travels consists of objects with dimensions that are small compared to the wavelength, and where the number of obstacles per unit volume is large. Scattered waves are produced by rough surfaces, small objects, or by other irregularities in the channel [16].

If there is line-of-sight between receiver and transmitter, diffraction and scattering are generally minor effects, although reflection may have a significant impact. If there is no clear LOS, such as in an urban area at street level, then diffraction and scattering are the primary means of signal reception.

4 Localization methods

Triangulation, scene analysis and proximity are the three principal techniques for automatic location-sensing [13]. Location systems may employ them individually or in combination. The triangulation location-sensing technique uses the geometric properties of triangles to compute object locations. Triangulation is divisible into subcategories of lateration, using distance measurements and angulations, using primarily angle or bearing measurements. Scene analysis observes features of its surroundings in order to determine the location of an object. In localization based on proximity, an object's presence is sensed using a physical phenomenon with limited range.

4.1 Lateration

Lateration computes the position of an object by measuring its distance from multiple reference positions [13]. Calculating an object's position in two dimensions requires distance measurements from three points that do not lie on a single line (non-collinear points). In three dimensions, distance measurements from four points not lying in the

same plane are required. Domain-specific knowledge may reduce the number of required distance measurements.

The 2D lateration technique works well when the three circles intersect at a single point, but this is rarely the case when estimates are used in ranging. When the range of anchor nodes is sufficiently large, the object to be located falls into a geometric region that is the intersection of three circles. This is called bounded intersection by Terwilliger [17] and is illustrated in Figure 2. It is also possible that the region of intersection is empty. This will occur if at least one ranging estimate is too small. Some methods overcome this problem by selecting the point for localization that gives the minimum total error between measured estimates and distances.

Fig. 2. Bounding the location of a node - the location of `X' is computed by taking the center of the intersection of the three circles.

Lateration is quite expensive in the number of floating point operations that is required. A similar, but computationally less expensive solution is to use a bounding box approach. The main idea is to construct a bounding box for each anchor using its position and distance estimate, and then to determine the intersection of these boxes. The position of the node is estimated to be the center of the intersection box. Figure 3 illustrates the bounding box method for a node with distance estimates to three anchors. The estimated position by the bounding box is close to the true position computed through lateration.

Two general approaches to measuring the distances required by the lateration technique, being attenuation and time-of-flight are discussed.

Fig. 3. The bounding box method for a node with distance estimates to three anchors

4.1.1 Attenuation

The intensity of an emitted signal decreases as the distance from the emission source increases. The decrease relative to the original intensity is the attenuation [13]. The signal strength decays with respect to distance. Under the ideal circumstances, signal power attenuation is proportional to d2, where d denotes the distance between the transmitter and the receiver. This effect is referred to as free space loss [14].

Usage of Received Signal Strength Indicator (RSSI) is one of the most commonly studied ap¬proaches for localization purposes because almost every node in the market has the ability to analyze the strength of a received message [18]. Given a function correlating attenuation and dis¬tance for a type of emission and the original strength of the emission, it is possible to estimate the distance from an object to some point P by measuring the strength of the emission when it reaches P. The widely used radio propagation model, the log-distance path loss model, considers the received power as a function of the transmitter-receiver distance raised to some power. Since this model is a deterministic propagation model and gives only the average value, another propagation model, the log-normal shadowing model, is introduced to describe the RSSI irregularity:

$$RSSI\ (d)[dBm] = RSSIref - 10n \log10\ (d/dref) + Xa \quad (1)$$

In equation (1), d is the transmitter-receiver distance, n is the attenuation constant (rate at which the signal decays), Xa a zero-mean Gaussian (in dB) with standard deviation a and RSSIref is the signal strength value at reference distance dref. Usually, n and a are obtained through curve fitting of empirical data. RSSI is measured in dBm, which is a logarithmic measurement of signal strength. RSSI value does not only depend on the distance, but also on the environment, antenna orientation and the power supply.

A commonly used model for calculating the distance d is given in Equation (2), in which RSSIref is measured at dref = 1 m. It is based on Equation (1), but multipath effects are omitted (Xa is assumed to be zero with probability one).

$$d(RSSI) = 10(RSSIref - RSSI)/10n \quad (2)$$

The attenuation constant is around 2 in an open-space environment, but its value increases if the environment is more complex (walls, large metallic objects, etc.). In environments with many obstructions such as an indoor office space, measuring distance using attenuation is usually less accurate than time-of-flight [13]. An approximation of the attenuation constant for an indoor environment is around 3.5 [16]. There is empirical evidence that due to the unreliability of measurements, accuracy in the scale of meters can be achieved regardless of the used algorithm or approach.

In the localization system Ferret, described by Terwilliger, two different ranging techniques (potentiometer and RSSI) are used to help locate an object to within one meter. In the poten¬tiometer technique, the object to be located - a mobile node begins by transmitting the beacon at the lowest power level and listens for replies from the infrastructure nodes. Increasing the power level with each transmission, once the mobile node gets three replies, it forwards its data to the base station for position computation. A calibration tool needs to be run each time the system is moved to a new environment in order to establish the communication ranges for given transmission power levels. Terwilliger also presents a location discovery algorithm that provides, for every node in the network, a position estimate, as well as an associated error bound and confidence level.

4.1.2 Time of Flight

Measuring distance from an object to some point P using time-of-flight means measuring the time it takes to travel between the object and point P at a known velocity. The object itself may be moving or it is approximately stationary and then, instead the difference in transmission, arrival time of an emitted signal is observed [13]. GPS is a well-known system which uses the time-of-flight technique. The first issue in using time-of-flight is to distinguish direct pulses from reflected ones because they look identical. Reflected measurements may be pruned away by aggregating multiple receivers' measurements and observing the environment's reflective properties. The second issue is agreement about the time. Since the propagation speed of radio signals is very high, time measurements must be very accurate in order to avoid large uncertainties. A localization accuracy of 1 meter requires timing accuracy on the level of $1/(3 \cdot 108) \approx 3.3$ nanoseconds. This means a minimum clock rate of 300 MHz ($3 \cdot 108$ Hz) is required for hardware. As far as time synchronization goes, state-

of-the-art protocols such as FTSP [22] synchronize nodes in the order of microseconds. To avoid this issue, a node could reflect the radio signal back, but this once again requires constant delay for reflecting the signal.

Time difference of arrival can also be measured. Cricket [23, 24], a location-support system for in-building, mobile, location-dependent applications, uses concurrent radio, ultrasound signals and measures the difference between the received times of the two types of signals. As sound waves travel at the speed of sound less precise timing than in the case of RF time-of-flight is required. A difference with radio signals is that an ultrasound signal does not go through walls; a similarity is that ultrasonic reception also suffers from severe multipath effects caused by reflections from walls and other objects. Cricket allows applications running on mobile and static nodes to learn their physical location by using listeners that hear and analyze information from beacons spread throughout a building. A case distinction is made for various situations in order to overcome multipath and interference effects. Practical beacon configuration and positioning techniques are used to improve accuracy up to the centimeter level.

4.2 Angulation

In angulation method angles are used for determining the position of an object. This technique is also called angle-of-arrival. In general, two-dimensional angulations requires two angle measurements and one length measurement such as the distance between the reference points as shown in Figure 4. In three dimensions, one length measurement, one azimuth measurement, and two angle measurements are needed to specify a precise position [13]. Although the definition of azimuth depends on the coordinate system, in this case, the azimuth is the horizontal component of an angle, measured around the horizon, from the north toward the east. Angulations implementations sometimes choose to designate a constant reference vector as 0°.

Fig. 4. Locating object `X' using angles relative to a 0 reference vector and the distance between two reference points.

Proposed solutions require special hardware. Phased antenna arrays are used to measure the angle. Antenna arrays consist of multiple antennas with known

separation in which each antenna measures the time of arrival of a signal. Given the differences in arrival times and the geometry of the receiving array, it is then possible to compute the angle from which the emission originated. If there are enough elements in the array and large enough separations, the angulations calculation can be performed [13]. Other approaches described in literature, Basaran [10], are compass sensors, rotating antennas and rotating light emitters combined with optical sensors.

4.2 Scene Analyses

The scene analysis location-sensing technique uses features of a scene observed from a particular vantage point to draw conclusions about the location of the observer or of objects in the scene [13]. In WSNs the measured feature of the scene is typically the signal strength value at a particular position and orientation. Scene analysis consists of an offline learning phase and an online localization phase. During the offline phase RSSI values to different anchor nodes are recorded at various positions. The recorded RSSI values and the known locations of the anchor nodes are used either to construct an RF-fingerprint database, or a probabilistic radio map. In the online phase, the node to be localized measures RSSI values to different anchor nodes. With RF-fingerprinting, the location of the user is determined by finding the recorded reference fingerprint values that are closest to the measured one. The unknown location is then estimated to be the one paired with the closest reference fingerprint or in the (weighted) centroid of k-nearest reference fingerprints. Location estimation using a probabilistic radio map includes finding the point(s) in the map that maximize the location probability [28].

The Microsoft Research RADAR location system is an example of RF-fingerprinting. RADAR uses a dataset of signal strength measurements created by observing the radio transmissions of an 802.11 wireless networking device at many positions and orientations throughout a building. The location of other 802.11 network devices can then be computed by performing table lookup on the prebuilt dataset. The median resolution of RADAR is in the range of two to three meters.

MoteTrack [26] extends the approach and claims to be more robust than RADAR. Base stations at fixed locations are used and a form of fingerprinting is used for determining the location of mobile nodes. The approach can tolerate the failure of up to 60% of the beacon nodes without severely degrading accuracy.

5 System description

A possible system setup depends on the intended application. On Figure 5, possible hardware setup is shown. One mote is attached to an interface board and acts as a base station. Together with other motes, the mote network is formed. The base station

collects information from the network and relays it to the PC. In turn, the PC processes network events and then updates the information on the server. The server computes the positions of nodes. The PDA asks the server for an update of the nodes' coordinates with a certain interval. The USB interface board provides connectivity to one mote at a time. Two serial ports are emulated over USB, one for communication with a mote and one for programming. A mote can also be reprogrammed over-the-air to receive an update of a program, but has to be programmed through the interface board first with the specific program.

Fig. 5. Possible hardware choice

The PC communicates with the base station over the USB connection and with the server over an Internet connection. It allows for communication between the base station and the server. The external server is used as an application server to which clients, such as the PDA, can be connected. Running the server application on the PC would be possible, but connecting to it from outside the network may be difficult if the network is protected with a firewall. It uses GPRS to connect to the server. This saves deployment time compared to using a PC for connecting to the server because the user does not have to keep walking back and forth to the PC between node registrations. In the localization phase, a mobile node and the PDA can be used together to show the position of the PDA on the map, or the PDA can be used to track another person holding the mobile node. The PDA is not connected to any sensor node. The way motes should be programmed depends on their function. The base mote is connected to the interface board and has to handle the communication between the PC and the mote network. All the non-mobile nodes listen for messages sent by mobile nodes. Each message contains the sender identification, packet number and sequence number. The mobile node sends a packet burst with a regular interval and increases the packet number by one each time this is done. The sequence number

is used to identify a packet within a burst. RSSI information is requested for each packet by the receiver. All the data of one packet burst is aggregated into one message and then sent to the base station. The sending is done using a multi-hop routing protocol, because not every mote may be in range of the base station. One solution for the motes is to use TinyOS. It is an open-source, event-driven operating system designed for wireless embedded sensor networks. It could be written in C programming language. Programs can be built out of components, which are assembled to form whole programs. TinyOS's component library includes network protocols, distributed services, sensor drivers and data acquisition tools.

There are two multi-hop routing protocols in TinyOS available: TYMO and the Collection Tree Protocol. TYMO is the implementation on TinyOS of the DYMO protocol, a point-to-point routing protocol for mobile ad-hoc networks. The current TYMO version is not stable. Therefore, the Collection Tree Protocol (CTP) [29,30] can be chosen. CTP is a tree-based collection protocol. Messages are collected at the roots of trees. Nodes form a set of routing trees to the tree roots. In this case, the only root would be the base station. CTP is the best effort protocol: it does not promise 100% reliable delivery and there are no ordering guarantees. CTP assumes that it has link quality estimates of some number of nearby neighbors. As a link estimator, an implementation of the four-bit wireless link estimation can be used, which can maintain a 99% delivery ratio with a transmission power of 0 dBm over large, multi-hop test-beds [31]. Nodes generate routes to roots, using a routing gradient. The protocol (CTP) uses the expected number of transmissions (ETX) as its routing gradient. CTP represents ETX values as 16-bit fixed-point real numbers with a precision of hundredths. A root has an ETX of 0. The ETX of a node is the ETX of its parent plus the ETX of its link to its parent. In general, CTP chooses the node with the lowest ETX value, unless it has reasons to do otherwise. CTP data frames also have a time has lived (THL) field, which the routing layer increments on each hop. CTP uses the ETX and THL fields to deal with routing loops and packet duplication. Edit, delete and load a map of the environment can be added. When adding or editing a map, the location that the map represents and a general description may be specified. The name and image file location must be specified. The width and height that the map represents in the physical world are also required. When a map is loaded the image is displayed. The position of the mobile node is updated regularly. If a map has been loaded, the user can enter the location of a static node on the map by clicking on it and entering the node number. These locations are saved and displayed. When the map is reloaded or another map is loaded, the nodes and their positions are deleted. The PDA can use a browser with good support of web standards on a mobile device.

6 Conclusions

The categories time and location are fundamental for many applications of Sensor Networks, to interpret sensing results or to coordinate sensor nodes in Wireless Sensor Networks (WSNs).

The most relevant techniques for Localization in Wireless Sensor Networks - lateration, angulation and scene analysis are discussed. A comparison is made between several localization systems. The choice for a certain system or technique depends on the intended application. The application requirements therefore determine for a great deal which hardware and software setup is feasible. The system can adapt itself to its environment within limited time by learning the relation between distance and signal strength. This system's main advantage over most other existing systems could be its short deployment time in a new environment while still achieving a reasonable accuracy. Improvements can be made with respect to accuracy and mobility. To be able to reach the desired accuracy in a large space, a relatively simple solution would be to increase the number of anchor nodes. This would also increase the cost of the system. Another way to improve accuracy would be to use a wall attenuation model, which systems such as RADAR and OPT employ. Such a model compensates for walls between sender and receiver. To better support a moving user, the mobile node could be attached to an accelerometer to detect if it is moving.

The Wireless Sensor Networks offer a huge palette of possibilities referring the localization issues. Of course, improvements referring accuracy, cost, mobility and sustainability are always possible and recommendable.

References

1. Sven Rienstra and Serhat Giildcek. Wireless sensor networks. Bachelor's thesis, Saxion Hogeschool Enschede, Netherlands, June 2008.
2. Alan Mainwaring, David Culler, Joseph Polastre, Robert Szewczyk, and John Anderson. Wireless sensor networks for habitat monitoring. In WSNA '02: Proceedings of the 1st ACM International Workshop on Wireless Sensor Networks and Applications, pages 88-97, New York, NY, USA, 2002. ACM.
3. Stuart Kininmonth, Scott Bainbridge, Ian Atkinson, Eric Gill, Laure Barral, and Romain Vidaud. Sensor networking the Great Barrier Reef. Spatial Sciences Queensland Journal, pages 34-38, Spring 2004.
4. David M. Doolin and Nicholas Sitar. Wireless sensors for wildfire monitoring. In Masayoshi Tomizuka, editor, Smart Structures and Materials 2005: Sensors and Smart Structures Tech¬nologies for Civil, Mechanical, and Aerospace Systems, volume 5765, pages 477-484. SPIE, 2005.
5. Carl Hartung, Richard Han, Carl Seielstad, and Saxon Holbrook. FireWxNet: a multi-tiered portable wireless system for monitoring weather conditions in wildland fire environments. In MobiSys '06: Proceedings of the 4th international conference on Mobile systems, applications and services, pages 28-41, New York, NY, USA, 2006. ACM.

6. Yanjun Li, Zhi Wang, and Yeqiong Song. Wireless sensor networks for wildfire monitoring.The Sixth World Congress on Intelligent Control and Automation (WCICA 2006), June 2006.
7. T. Bokareva, W. Hu, S. Kanhere, B. Ristic, N. Gordon, T. Bessel, M. Rutten, and S. Jha. Wireless sensor networks for battlefield surveillance. Proceedings of Land Warfare Conference2006, October 2006.
8. Tian He, Sudha Krishnamurthy, Liqian Luo, Ting Yan, Lin Gu, Radu Stoleru, Gang Zhou, Qing Cao, Pascal Vicaire, John A. Stankovic, Tarek F. Abdelzaher, Jonathan Hui, and Bruce Krogh. VigilNet: An integrated sensor network system for energy-efficient surveillance. ACM Transactions on Sensor Networks (TOSN), 2(1):1-38, 2006.
9. HyungJune Lee, Martin Wicke, Branislav Kusy, and Leonidas Guibas. Localization of mobile users using trajectory matching. In MELT '08: Proceedings of the first ACM International Workshop on Mobile Entity Localization and Tracking in GPS-less Environments, pages 123¬128, New York, NY, USA, 2008. ACM.
10. Can Basaran. A hybrid localization algorithm for wireless sensor networks. Master's thesis, Yeditepe University, Turkey, July 2007.
11. Andreas Savvides, Mani Srivastava, Lewis Girod, and Deborah Estrin. Localization in sensor networks. In Wireless Sensor Networks, chapter 15, pages 327-349. Kluwer Academic Publishers, Norwell, MA, USA, 2004.
12. Adel Amin Abdel Azim Youssef. SALAM: a scalable anchor-free localization algorithm for wireless sensor networks. PhD thesis, University of Maryland, College Park, MD, USA, 2006.
13. Jeffrey Hightower and Gaetano Borriello. A survey and taxonomy of location systems for ubiquitous computing. Technical Report UW-CSE 01-08-03, University of Washington, Com-puter Science and Engineering, Seattle, WA, USA, August 2001.
14. Pei Zheng and Lionel Ni. Smart Phone and Next Generation Mobile Computing. Morgan Kaufmann, 2005.
15. William Stallings. Data and Computer Communications. Prentice Hall, 8th edition, 2007.
16. Spread Spectrum Scene. An introduction to indoor radio propagation. http://sss-mag.com/indoor.html, June 2001.
17. Mark Terwilliger. Localization in wireless sensor networks. PhD thesis, Western Michigan University, Kalamazoo, MI, USA, April 2006.
18. Youssef Chraibi. Localization in wireless sensor networks. Master's thesis, Royal Institute of Technology, Sweden, November 2005.
19. Tsenka Stoyanova, Fotis Kerasiotis, Aggeliki Prayati, and George Papadopoulos. Evaluation of impact factors on RSS accuracy for localization and tracking applications. In MobiWac '07: Proceedings of the 5th ACM International Workshop on Mobility Management and Wireless Access, pages 9-16, New York, NY, USA, 2007. ACM.
20. Xueli An, Jing Wang, R. Venkatesha Prasad, and I. G. M. M. Niemegeers. OPT: online person tracking system for context-awareness in wireless personal network. In REALMAN '06: Proceedings of the 2nd International Workshop on Multi-hop Ad Hoc Networks: from Theory to Reality, pages 47-54, New York, NY, USA, 2006. ACM.
21. E. Elnahrawy, Xiaoyan Li, and R. P. Martin. The limits of localization using signal strength: a comparative study. In Proceedings of the First Annual IEEE Communications Society Conference on Sensor and Ad Hoc Communications and Networks (SECON '04), pages 406¬414, October 2004.
22. Miklos Maroti, Branislav Kusy, Gyula Simon, and Akos Ledeczi. The flooding time synchronization protocol. In SenSys '04: Proceedings of the 2nd International Conference on Embedded Networked Sensor Systems, pages 39-49, New York, NY, USA, 2004. ACM.

23. Nissanka B. Priyantha, Anit Chakraborty, and Hari Balakrishnan. The Cricket location-support system. In MobiCom '00: Proceedings of the 6th Annual International Conference on Mobile Computing and Networking, pages 32-43, New York, NY, USA, 2000. ACM.
24. Adam Smith, Hari Balakrishnan, Michel Goraczko, and Nissanka Bodhi Priyantha. Tracking Moving Devices with the Cricket Location System. In 2nd International Conference on Mobile Systems, Applications and Services (Mobisys 2004), Boston, MA, USA, June 2004.
25. Ville A. Kaseva, Mikko Kohvakka, Mauri Kuorilehto, Marko Hnnikinen, and Timo D. Hmli-nen. A wireless sensor network for RF-based indoor localization. In EURASIP Journal on Advances in Signal Processing, volume 2008, 2008.
26. Konrad Lorincz and Matt Welsh. Motetrack: a robust, decentralized approach to RF-based location tracking. Personal Ubiquitous Computing, 11(6):489-503, 2007.
27. Nissanka B. Priyantha, Hari Balakrishnan, Erik D. Demaine, and Seth Teller. Mobile-assisted localization in wireless sensor networks. Proceedings of IEEE INFOCOM '05, 1:172-183, March 2005.
28. Ville A. Kaseva, Mikko Kohvakka, Mauri Kuorilehto, Marko Hnnikinen, and Timo D. Hmli-nen. A wireless sensor network for RF-based indoor localization. In EURASIP Journal on Advances in Signal Processing, volume 2008, 2008.
29. Rodrigo Fonseca, Omprakash Gnawali, Kyle Jamieson, Sukun Kim, Philip Levis, and Alec Woo. TinyOS Extension Proposal 123: The Collection Tree Protocol (CTP). http://www.tinyos.net/tinyos-2.x/doc/html/tep123.html, February 2007. Draft-Version 1.8.
30. Rodrigo Fonseca, Omprakash Gnawali, Kyle Jamieson, and Philip Levis. TinyOS Extension Proposal 119: Collection.http://www.tinyos.net/tinyos-2.x/doc/html/tep119.html, Febru¬ary 2006.
31. Rodrigo Fonseca, Omprakash Gnawali, Kyle Jamieson, and Philip Levis. Four-bit wireless link estimation. Technical Report SING-07-00, UC Berkeley, Univ. of Southern California, MIT CSAIL, Stanford Univ., 2007.
32. Srdjan Capkun, Maher Hamdi, and Jean-Pierre Hubaux. GPS-free positioning in mobile ad hoc networks. Cluster Computing, 5(2):157-167, 2002.
33. Paramvir Bahl and Venkata N. Padmanabhan. RADAR: An in-building RF-based user location and tracking system. pages 775-784, Tel-Aviv, Israel, March 2000. IEEE Infocom.

Analysis of an Algorithm for finding Minimal Cut Set for Undirected Network

Marija Mihova[1] and Natasha Maksimova[2],

[1] Ss. Cyril an Methodius University, Faculty of Natural Sciencies and Mathematics, 1000 Skopje, Macedonia
[2]University "Goce Delčev, "Faculty of informatics, 2000 Shtip

Abstract. In this paper we propose an algorithm for obtaining all minimal cut sets for a given two-terminal network. The algorithm works on undirected networks without matter whether they are coherent or not. The difference between this algorithm and the other proposed algorithms is in the fact that there are not received candidates for minimal cut set that are not minimal cut sets. A large part of the paper proves the correctness of the algorithm and analyzes its complexity.

Keywords: Network, minimal cut set, cut set, undirected network, connected network, graph.

1 Introduction

One of the most complex problems in system reliability analysis is finding minimal path and cut sets. Minimal cut sets are mostly used for high reliability systems because this approach received minor rounding errors. A problem for developing algorithms that will give these sets is one of the frequently analyzed problems. But usually there are regarded directed and acyclic networks [1, 2]. Here we regard undirected network.

From the other side, in most of the proposed algorithms that solves this problem, there are obtained candidates for minimal cut sets that are not minimal. These sets must be eliminated by some additional procedure that involves comparison between all pairs of candidates for minimal cut set. This is an expensive procedure, without difference is there are a lot of such candidates or not. The algorithm proposed here gives minimal cut sets only. In fact we propose a technique for determination whether some cut set is a minimal cut set. Moreover we give a proof that the proposed algorithm gives all minimal cut sets. Additionally, we analyze the complexity of the algorithm.

2 General

Let we have a two-terminal undirected network $G(V, E)$, where V is the set of nodes, and E is a set of links. Let s be the source node and t be the sink node. The cut set is defined as a set of links, such that if there no flow through these links, then there no flow from the source to the sink. The cut set C is a minimal cut set if there is not another cut set C' such that $C' \subset C$. For an undirected network the following proposition is clear.

Proposition 1.1 Let $G(V, E)$ be an undirected connected network with source node s and sink node t. Then C is a cut set if and only if by removing the links from C, the graph $G(V, E)$ is divided into two subgraphs $G_1(V_1, E_1)$ and $G_2(V_2, E_2)$ such that $V_1 \cap V_2 = \varnothing$, $V_1 \cup V_2 = V$, the source $s \in V_1$ and the sink $t \in V_2$.

The following proposition specifies the minimal cut sets. The proof of it is given in [3].

Proposition 1.2 Let $G(V, E)$ be an undirected connected network with source node s and sink node t. If the graph $G(V, E)$ is divided into two connected subgraphs $G_1(V_1, E_1)$ and $G_2(V_2, E_2)$ such that $V_1 \cap V_2 = \varnothing$, $V_1 \cup V_2 = V$, the source $s \in V_1$ and the sink $t \in V_2$, then $C = E/(E_1 \cup E_2)$ is a minimal cut set.

Example 1.1 For the two-terminal network with source s and sink t on the Fig.1, the minimal cut set $C = \{\{s,c\}, \{c,b\}, \{b,e\}, \{e,d\}, \{d,t\}\}$ separated the network in to two subgraphs $G_1(\{s, a, b, d\}, \{\{s, a\}, \{s, b\}, \{a, d\}, \{b, d\}, \{a, b\}\})$ and $G_2(\{c, e, f, t\},\{\{c, e\}, \{c, f\}, \{e, f\}, \{e, t\}, \{f, t\}\})$

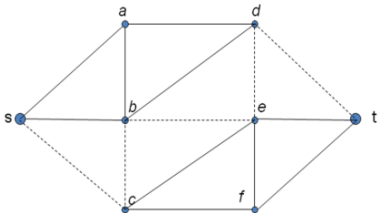


Fig. 1. The two-terminal network with source s and sink t. The set $C = \{\{s,c\}, \{c,b\}, \{b,e\}, \{e,d\}, \{d,t\}\}$is a minimal cut set for this network. The graph is separated in to $G_1(\{s, a, b, d\}, \{\{s, a\}, \{s, b\}, \{a, d\}, \{b, d\}, \{a, b\}\})$ and $G_2(\{c, e, f, t\},\{\{c, e\}, \{c, f\}, \{e, f\}, \{e, t\}, \{f, t\}\})$

3 Description of the algorithm

In this section we will explain the work of the proposed algorithm. From the Proposition 1.2 follows that by deleting the edges from some minimal cut set C, the graph will be separate into exactly two connected components, such that, s belongs in the one of them and t in the other, Fig. 1 So, the idea of the algorithm is to construct all those components.

In order to avoid obtaining the same connected graphs more than once, we define ordering of the nodes. For this ordering we use BFS search starting from s and define a function

P: $V \rightarrow \mathbf{N}$, $P(a)=i$ if a is the i-th node in the BFS search started from s (1)

So, we have that if $P(a) < P(b)$ then the shortest path from s to a is longer or equal then the shorter path from s to b.

In each step of the while loop we use an element $\{B_1, B_2, B_3, T\}$. Let us see the meaning of each element from this list. B_1 is one minimal cut set for the given graph. Corresponding graph G_1 is $G_1(B_3, \{\{u, v\}|\{u, v\} \in B_3\})$ and B_2 is a set of all nodes v from V/B_3 that need to be added in B_3 in one of the next steps. T is a tree rooted in t that connects all nodes that are not in B_3. This tree is kept as a set of nodes, list of decedents of each node call `DescadentList` and the parent function π.

At the beginning we set the initial values of the sets B_1, B_2 and B_3. B_1 is a minimal cut set, consist of all links from s, B_3 is initialize on $\{s\}$, and B_2 is initialize on the set of all nodes connected by edge with s. The initialize tree T is a BFS tree of the graph $G'(V/B_3, E/B_1)$ rooted at t. It is clear that the minimal cut set B_1 divide the graph G into two graphs $G_1(\{s\}, \emptyset)$ and $G_2(\{V/\{s\}, E/B_1)$.

```
1 Program MINIMAL_CUT_SET (G, s, t)
  Output{CutSet (the set of all minimal cut sets)}

  Construct the function P using BFS started from s;
5 B1={{s, a} | {s, a} ∈ E};
  B2={a | {s, a} ∈ E}/{t};
  B3={s};
  T ={DescadentList, π} of the graph G'(V/B3,E/B1) by BFS;
  A = {{B1, B2, B3, T}};
10 MinCutSet={{{s, a} | {s, a} ∈ E}};
     while A≠∅ do
       Take an element B={B1, B2, B3, T} ∈ A and remove it from A.
        for each element b ∈ B2 do
            if first(CONECTION(E, B1, b, T))==True then
15              B3 = B3∪{b}
                B2 ={x|x∈ B2 ∧ P(b) < P(x)} ∪ {a | {a, b} ∉ B1 ∧ (P(b) < P(x) ∨
                                                  ¬(∃{x, a}, {x, a}∈B1)}
                B1 = (B1\ {{a, b}|{a, b}∈B1})∪{{a, b}|{a, b} ∉B1}
```

```
                MinCutSet = MinCutSet ∪ {B1}
20              A = A∪{ {B1, B2, B3 , Last(CONECTION)}}
            else if MIN(CONNECTION[[2]]>b then
                B4 =CONECTION[[2]]
                B3 = B3∪{b} ∪B4
                B2  ={x|x∈ B2 ∧ P(b) < P(x)} ∪ {a | {a, b} ∉ B1 ∧ a ∉ B4 ∧
25                      (P(b) < P(x) ∨ ¬(∃{x, a}, {x, a}∈B1)}
                B1 = (B1\ {{a, b}|{a, b}∈B1})∪{{a, b}|{a, b} ∉B1∧ a ∉B4}
                MinCutSet = MinCutSet ∪ {B1}
                A = A∪{ {B1, B2, B3 , Last(CONECTION)}}
        Print (MinCutSet)
        end{program}
```

In each iteration of the while loop we take an element $\{B_1, B_2, B_3, T\}$ from A. The cut set B_1 divides G into two connected graphs $G_1(B_2, E_2)$ and $G_2(V/B_2, E/(E_2 \cup B_1))$. Then in each iteration of the **for** cycle we take one node from B_2 and check whether his removal from G_2 other nodes remain connected. This can be done by BFS starting from t. But we do this with the procedure `CONNECTION`, which is chipper. If the rest of G_2 is connected, in lines 15-20 we get another member in A and another minimal cut set.

If the rest of G_2 is not connected we do additional checks. Since we have a connected graph, they are connected by some node from $V \backslash V_T$, so V_T and $V \backslash V_T$ separate G into two connected components. Because we want no repetition of the same combination of vertices in B_3, we use strategy to add nodes having larger value of P. But it is not always possible to add nodes in increasing order, so when we receive a set of nodes that are not connected with t such that they have P value greater than $P(b)$, they are added to B_3 (22-28).

The procedure CONNECTION is called from the graph G, the tree T and one node b from T. The output of the procedure is an answer FAULT, when there are nodes in T that, after removing b, are not connected with the sink t and all of them have bigger value of P then $P(b)$. In opposite case, the answer is TRUE, together with a new tree rooted at t and list of unconnected nodes. At the beginning, the procedure CONNECTION (in lines 2 – 8) checks out whether b is a leaf in T. If it is true, than the tree obtained from T by removing b connects remain nodes of T. When b is not a leaf, we have more work. In lines 12-15 each descendant of b is color in red and removed from V_T. Then for each red element it is checked whether there is a link between it and some node from V_T (nodes for which we know that they are connected with the sink). If it is connected, it is colored in gray and added to the tree (d is put as a parent of c and c is put into DescadentList of d, lines 20-21). Now if there not gray nodes, it is clear that nodes in V_T are not connected, otherwise, we are not sure yet. So we use BFS, started from the gray nodes and add to the tree all gray nodes to which we arrive. If you still have red nodes, there are nodes that are not in V_T and we mark *isconect* as True. In opposite, we mark *isconect* as False.

```
1  Procedure CONNECTION (G, T=(V_T, DescedentList,π), b)
     Output{{True, T=(DescedentList,π)};

       If DescedentList(b)==∅ then            // if b is leaf
5        isconect = True
         remove b from the DescadentList(π(b))
         π(b)=NILL
       else
         Gray=∅
10       M=Red=DescedentList(b)
         VT=VT\M
         while M != ∅ do
                  c =First(M);
                  M=(M\{c})∪ DescedentList(c)
                  Red=Red∪ DescedentList(c);
15                VT=VT\ DescedentList(c)
          for all c ∈ Red do
               if there is d such that {c, d} ∈ E and d ∈ V_T
                        then
                          Gray=Gray ∪ {c};
20                        π(c)=d
                          DescedentList(d)= DescedentList(d) ∪ {c}
          while Gray != ∅ do
                f=First(Gray);
25              Gray=Gray/{f};
                VT=VT∪{f};
                DescedentList(f)={u|{u, f} ∈ V and  u ∈Red}
                Gray= Gray ∪ DescedentList(f)
                Red=Red/ DescedentList(f)
                for all u∈ DescedentList(f) do π(u)= f
30        if Red != ∅ then isconnect = False else isconnect = True
        if isconnect=False then output={False, Red, T=(VT, DescedentList,π)) }
          else output={True, T=(VT, DescedentList,π))
    print ( output )
    end{procedure}
```

As a illustration of the algorithm, we give following example.

Example 3.1. Let us consider how the algorithm works on the network given on Fig 2, with source node 1 and sink node 7. The tree rooted at 7 is shown on the Fig. 2 b.

At the beginning A={{B_1={{1, 2}, {1, 3}}, B_2={2, 3}, B_3={1}, T={{7,{2, 3}}, {2, {5, 4}}, {3, {6}}, {4, ∅}, {5, ∅}, {6, ∅}}}.

By the first step, when the procedure CONECTION is call for the first element 2, it is illustrates how the algorithm works with irrelevant links. 2 is not a leaf, so the nodes 4 and 5 are put in Red, and move from T. 4 is connected by 6, so it is color in gray and it is connects by the tree through the node 6. 5 is not connected with some

node from T, and after leaving the procedure CONECTION, Red={5}. New tree is {{7,{3}}, {3, {6}}, {6, {4}}, {4, ∅}} Now, min{5}=5>2, so we obtain the minimal cut set {{1, 3}, {2, 4}, {2, 7}}. In fact the element

{{{1, 3}, {2, 4}, {2, 7}}, {3, 4}, {1, 2, 5},{{{7,{3}}, {3, {6}}, {6, {4}}, {4, ∅}}} (2)

is added to A.

Next we call CONECTION for the element 3, which is a leaf, so the node 6 is put in Red, and move from T. But 6 is connect by 4 and Red becomes ∅. The element

{{{1, 2}, {3, 6}, {3, 7}}, {6}, {1, 3}, {{{7,{2}}, {2, {4, 5}}, {4, {6}}, {6, ∅}, {5, ∅}}} (3)

is added to A. This is an illustration of the case when the node *b* is not a leaf, but all other vertices are connected with the sink.

By the next step we also illustrate one of the characteristics cases. (2) is taken, and CONNECTION is called for 3. The nodes 6 and 4 are added into Red. When we remove 3, these nodes are not connected with the sink. But min{4, 5}=4>3, so 4 and 6 are added in B_3. So as a minimal cut set we obtain {{2, 7}, {3, 7}} and as a B_2, ∅. So from this element we do not obtain another cut set.

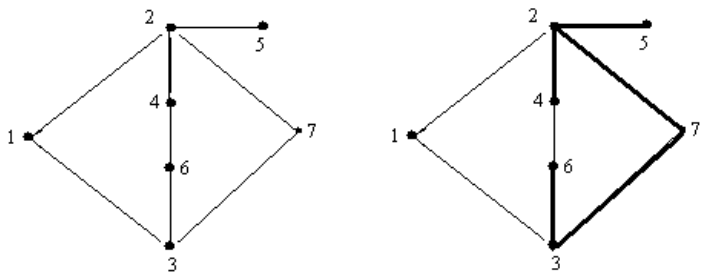


Fig.2. Two-terminal network with source node 1 and sink node 7.

Next step is a characteristics case when we add a leaf in B_3. In fact, CONNECTION is called for 4 which is a leaf in (2), so immediately we add

{{{1, 3}, {6, 4}, {2, 7}}, {6}, {1, 2, 4, 5},{{{7,{3}}, {3, {6}}, {6, ∅}}} (4)

into A.

We are finished with (2) and take (3). We have only one element in B_2, 6, which is a leaf and the cut set we obtained from here is {{1, 2}, {2, 7}, {6, 4}}. Node 4 is connected by node 6, and it is note found as a vertex in the set of links {{1, 2}, {3, 6}, {2, 7}}. So we add 4 in B_2 and put in A the element

{{{1, 2}, {4, 6}, {3, 7}}, {4}, {1, 3, 6}, {{{7, {2}}, {2, {4, 5}}, {4, ∅},{5, ∅}}} (5)

Now we take (4) and add the node 6 into B_3. B_2 becomes empty, since 6 is connect by 4 which is in B_3, and 3 which is a node in B_1. Tne minimal cut set obtained from here is {{1, 3}, {6, 3}, {2, 7}}. Similarly from (5) we obtain {{1, 2}, {4, 2}, {3, 7}} as a minimal cut set and the algorithm ends.

4 Analysis of the algorithm

This section will prove that the proposed algorithm for finding the minimal cut sets for a connected undirected network works correctly.

Proposition 2.1 Let MINIMAL_CUT_SET is terminate on the graph $G(V, E)$. After each iteration of the while loop, the graph $G_1(B_3, \{\{u, v\} | u, v \in B_3\})$ is a connected graph, such that $s \in B_3$.

Proof It is clear that $G_1(\{s\}, \emptyset)$ is a connected graph. In each iteration of the loop we take an element of A={B1, B2, B3, T} and add one new elements in B_3. Suppose that for all integer smaller or equal to k, if $|B_3| < k$, then $G_1(B_3, \{\{u, v\} | u, v \in B_3\})$ is connected. Let B'_3 is obtained from some B_3, such that $|B_3| < k$, by adding new set of nodes $Red \cup \{v\}$. It is clear that the subgraph of G consisting of these nodes is connected. Moreover, since G is connected graph, this subgraph is connected with some node $u \in B_3$. So B'_3 is also connected and $s \in B'_3 \subset B_3$.

Proposition 2.2 Let CONNECT is called from the connected graph G, the tree T of nodes from G, rooted in t and a node $b \neq t$ from T. Let $G_2(\{u | \text{ is in } T\}/\{b\}, \{\{u, v\} | u, v \in \{u | \text{ is in } T\}/\{b\})$. Then CONNECT gives False when G_2 is not connected, and True, together with a tree that connects the nodes from G_2, when G_2 is connected graph.

Proof First suppose that G_2 is not connected. Then there is a node u such that there is no link from t to u in G_2. We claim that all paths in $G'_2(\{u | u \text{ is in } T\}, \{\{u, v\} | u, v \in \{u | u \text{ is in } T\})$ pass through b (if it is not true, there is a path in G_2 from t to u). So, u must be a descendant of b in T and u is colored red in the line 14 and it is deleting from V_T in line 15. Since u is not connected with some node from V_T, u remains red after line 29. Now we claim that there is not a path from some gray node v, to u. If it is true, then there is a path $t \rightarrow a \rightarrow u \rightarrow v$, $\{a, u\} \in E$, $a \in V_T$ such that b is not on that path which is a contradiction. So, in line 30 the set Red$\neq\emptyset$ and CONNECT provides an answer False.

Now suppose that G_2 is connected. We need to proof that after the termination of the procedure CONNECT, $Red = \emptyset$. Since CONNECT is call when b is not a leaf, after line 16, $Red \neq \emptyset$. Let $u \in Red$. Since G_2 is connect, there is a path from t to u in G_2 do not passes through b. Suppose that the first node which lies on the path from t to u and it is a descendant of b in T is a. a will be added in to new tree in the lines 19-21. u will be added into new tree in lines 23-29, since this part of the algorithm is a part of BFS.

Theorem 2.1 Program MINIMAL_CUT_SET is consists of all minimal cut sets of a given graph $G(V, E)$.

Proof From Proposition 2.1 and Proposition 2.2 follows that each element we put into *CutSet* divide the graph G into two connected components, such that s is in one of them and t is in the other. So each element in *CutSet* is a minimal cut set.

It remains to prove that all minimal cut sets will be added to *CutSet*. This will be proved by induction in respect to the number of elements in B_3.

It is clear that the only set B_3 with one element is $\{s\}$. Suppose that all minimal cut sets that separate V into B_3 and V/ B_3, such that $|B_3|\leq k$, are added into *CutSet*. Let C divides G into two connected graphs $G_1(B_3, \{\{x, y\}| x, y \in B_3\})$ and $G_2(V/B_3, \{\{x, y\}| x, y \notin B_3\})$ and let $b \in B_3$ such that $P(b)=\max\{P(x)|x\in B_3$ and there is a link $\{x, y\}\in E$ and $y\notin B_3\}$. Define $S_0=\{b\}$, $S_k=\{x\in B_3| P(x) \geq P(b)$ and $\{x, y\}\in E, y \in S_{k-1}\}$. It is clear that there is finite number of such sets, so let $S = \bigcup_k S_k$ and that S_1 is the set of all neighbors of b in G_1. We will regard two cases:

I case: By removing the nodes from S, $G'_1(B_3\backslash S, \{\{x, y\}| x, y \in B_3\backslash S\})$ is connected. This case will be divided into two subcases, when all neighbors of b in G_1 have greater value of P then b and when there is a neighbor x of b in G_1 such that P(x)<P(b).

Regard how we will obtain C when all neighbors of b in G_1 have greater value of P then b, i.e. the set of neighbors of b in G_1 is S_1. Let a $\in S_1$ is the node with the smallest value of P such that it is connect with some node from $B_3\backslash (S_1\cup \{b\})$. Take the smallest subgraph from G_1, such that b and all neighbors of b in $G_1\backslash\{a\}$ are not in them. This graph and the rest of the graph G, (which is connect, since all nodes in it are connected by b and b is connect with the rest of the graph) separate G into two connected components and all links that are not in it constitute an minimal cut set. From the inductive assumption this minimal cut set is obtained by the algorithm. Moreover, b is in the corresponding list B_2, (it is add by the second part of the formula of B_2). Now, if $S=\{b\}$, C is obtained in lines 15-19, in opposite, in lines 22-27.

In the opposite, the graph G'_1 and the rest of the G separate G into two connected components and b is added into the corresponding list B_2 by the first part of the formula for B_2. So again if $S=\{b\}$, C is obtained in lines 15-19, in opposite, in lines 22-27.

II case: By removing the nodes from S, G'1(B3\S, {{x, y}| x, y ∈ B3\S}) is not connected. Let $\hat{G}'(\hat{V}', \hat{E}')$ is the connected component of G'1 in which s belongs, and $\hat{G}'$ is the graph $\hat{G}(\hat{V}'\cup S, \{\{x, y\} | x, y \in \hat{V}'\cup S\})$. Since for the nodes in G' that have smallest value of P then b, the shorter path in G is shorter then the shorter path in G'. This path must pass true nodes that are not in G', so, such nodes are connected by some other nodes from G2 i.e. $\hat{G}$ and $\hat{G}_1(V /(\hat{V}'\cup S), \{\{x, y\} | x, y \notin \hat{V}'\cup S\})$ separate G into two connected components and from the inductive assumption the appropriate minimal cut set is obtained. The nodes with smaller values of P are added in B_2 by the second part of the union.

Look at the graph $\hat{G}_1$. Let b' is a node in $\hat{G}_1$ such that $P(b')=\max\{P(x)|x\in B_3$, $x \in B_3 \cap (\hat{V}'\cup S)^C$ and there is a link $\{x, y\}\in E$ and y$\notin B_3\}$. Define $S'_0=\{b\}$, $S'_k=\{ x \in B_3 \cap (\hat{V}'\cup S)^C |\ P(x)\geq P(b')$ and $\{x, y\}\in E, y \in S'_{k-1}\}$ and $S' = \bigcup_k S'_k$. There two cases again, by removing the nodes from S', the rest of the graph is connected and by removing the nodes from S', the rest of the graph is unconnected. These two cases are considered in the same way as cases I and II. This can be

repeated until we get connected graph. Since G has a finite number of nodes, the procedure will finish.

At the and, let us regard the complexity of the proposed algorithm. Each minimal cut set is obtained only once, but, as a set of nodes B_3, there obtained some combination that not lead to minimal cut set. The good think is that each set B_3 is obtained at most once. So, the worst case is when all subsets of $V/\{t\}$ are acquired as a set B_3. This is obtained for complete graph only, but in this case, each candidate for a minimal cut set is actually a minimal cut set. From this we have that for a graph with |V| nodes the procedure CONNECTION is call at most $2^{|V|-1}$ times.

In the algorithm proposed in [1] has the same complexity in finding candidates for minimal cut sets, but the candidates for minimal cut set that are not minimal are rejected by comparison with all other candidates. In that comparison it is determine whether there is a set in the list of candidates for minimal cut set which is a subset of the given set of links, and if so, it is rejected from the list. So, in order to reject all candidates from the list of N candidates, it is needed to make $\Theta(N^2)$ comparisons between sets.

In our algorithm we have a procedure for determining whether a given cut set is minimal. In this procedure we make BFS on part of the network. In most of this cases there are very few nodes on which we make BFS. In fact, in dense graphs, the procedure CONNECTION is called frequently, but in most of the cases, the node from which it is called is a leaf or has a small number of descendents, so the complexity of the procedure is constant. For example, for complete graph, the procedure CONNECTION is called only for leaves.

6 Conclusion

The detailed analysis of the proposed algorithm indicates that it correctly gives all the minimal cut vectors for a given two-terminal undirected network. It is characteristic that this algorithm works for undirected networks that are less discussed in the literature. Because the cycles of such networks are inevitable, by minor modification it can be modified to work for directed networks in which the cycles are allowed. Also, this algorithm avoids getting cut sets that are not minimal cut sets. This avoids the additional complexity of the program as a result of rejection of those sets.

References

1. A.Bethel A. Gebre ,Jose E. Ramirez-Marquez, ”Element Substitution Algorithm for General Two-terminal Network Reliability Analyses”, IIE Transactions, Volume 39, March 2007, 265 - 275

2. A. Marteli, “A Gaussian Elimination Algorithm for the Enumeration of Cut Sets in a Graph”, Journal of the Association for Computing Machinery, Vol 23, No 1, January 1976
3. M. Mihova, N. Maksimova, “Minimal Cut Sets for Transportation System”, CD of extended abstracts of the 7th International Conference for Informatics and Information Technology (CIIT 2010);

Session Security

Aleksandar Kotevski[1] and Gjorgi Mikarovski[2]

[1]Faculty of law Science – 7000 Bitola Macedonia
aleksandar.kotevski@uklo.edu.mk
[2]Faculty of Technical Science – 7000 Bitola Macedonia
gjorgi.mikarovski@uklo.edu.mk

Abstract. Sessions are one way of saving the state and some useful user information across subsequent page requests. Sessions usually are used for storing information about registered user, selected language (in situation of Multilanguage portals), item in shopping card (in situation in e-commerce) etc.

Keywords: Session,cookie,fixation, hijacking,poisoning, -site scripting,secure,http,ssl,

1 Introduction

If you working with sessions, you must know that session module cannot guarantee that the information which are store in a session is only viewed by the user who created the session, anyone with access to the server can access the PHP session files. Because of security aspect it is not clever to store critical information in session variables. Therefore, you must take additional measures to actively protect the integrity of the session, depending on the value associated with it.

2 Manipulating with session

Specific issues that may manipulate with sessions are:

- Creation and identification
- Session termination and timeout – what triggers a session termination? How are the resources of the terminated session recycled?
- Concurrency issues

From security aspect, the following considerations must be made:

- It should never be possible for one client to be able to predict the token another client received, or is in the process of receiving, or will receive.
- Furthermore, it is desirable that a client will not be able to predict the next token when he/she will get access to the site. This is useful in minimizing the damage of

stealing the token while it travels (in the clear) to and from, and while it is stored on disk from the client.
- Any token should have a reasonable expiration period – again, to minimize the damage in case of being stolen.

3 What is session variable

Session variables have the following features:

- Unless specified otherwise, session variables expire 20 minutes after a visitor leaves the site.
- Session variables will expire if no activity is detected on the site for 20 minutes by that specific site visitor, or if the visitor quits out of their web browser.
- In order for session variables to work, the visitor's browser must be set to accept cookies.
- The pages for the site all must be located within a single directory on the web server.
- Information stored in session variables is site visitor specific. Different site visitors cannot access each other's session variable information.

4 Types of session attack

There are few types of attacks that you need to be wary about when you are working with session variables:

- Session Fixation
- Session Hijacking
- Session Poisoning (injection)
- Session cross-site scripting

4.1 Session Fixation

Web session security is mainly focused on preventing the attacker from obtaining – intercepting, predicting or brute-forcing - a session ID issued by the web server to the user's browser.
The attack itself is very basic. The hacker forms a link or redirects which sends the user to your site with the session ID present:

```
<a href=http://hostname/index.php?PHPSESSID=1234> Click
</a>
```

When users click on that link or are redirected there, they connect to your site with a session ID that has been set by the attacker. The attacker can now wait for the users to log in and access your site using their credentials.

First, the attacker – who in this case is also a legitimate user of the mail system – logs in to the server and issues a session ID 1234.

Sends a hyperlink `domainane.com/login.php?sessionid=1234` to the user, trying to lure him into clicking on it. The user clicks on the link, which opens the server's login page in his browser. Note that upon receipt of the request for login.php?sessionid=1234, the web application has established that a session already exists for this user and a new one need not to be created. Finally, the user provides his credentials to the login script and the server grants him access to his email account.

Knowing the session ID, the attacker can also access the user's account via `email.php?sessionid=1234.` Since the session has already been fixed before the user logged in, we say that the user logged into the attacker's session.
To prevent this, PHP has to define a `session_regenerate_id()`. This function generates a new session file for the user, gets rid of the old one, and issues a new session cookie if your site utilizes them. Another good practice is setting a session time-out in the php.ini file.
But, unfortunately these methods do not guarantee that an attacker can't get your users' session Ids, so, you could use SSL/TSL. SSL (Secure Sockets Layer) provides one of the most commonly available security mechanisms on the Internet and it's used extensively by web browsers to provide secure connections for transferring sensitive data. An SSL-protected HTTP transfer uses port 443 and is identified with a special URL method - https.

4.2 Session Hijacking

There are three common methods for session defense:

- User agent verification
- IP address verification
- Secondary token

There are two major drawbacks to this method:

- A lot of locations are behind a NAT proxy, so it is possible that the attacker and the user both have the same IP address
- Large ISPs like AOL - a number of them, and AOL specifically, have massive proxy setups that send the user out via a different IP address with every page request.

4.3 Session poisoning

Session poisoning or session data pollution, modification, injection is to exploit insufficient input validation in server applications which copies user input into session variables.

4.4 Session cross-site scripting

Cross-Site Scripting attacks are type of injection problem, in which malicious scripts are injected into the otherwise benign and trusted web sites. Cross-site scripting (XSS) attacks occur when an attacker uses a web application to send malicious code, generally in the form of a browser side script, to a different end user. XSS is just a special case of code injection. In this type of attack, the malicious user embeds HTML or other client-side script into your Web site. The attack looks like it is coming from your Web site, which the user trusts. This enables the attacker to bypass a lot of the client's security, gain sensitive information from the user, or deliver a malicious application. There are two types of XSS attacks: reflected or non-persistent and stored or persistent.

5 Secure practice

At least, each developer needs to follow this secure practice to make the application secure:

- Set a new path for your session data storage
- Do not pass your session identifier in
- If skeptical on the entire HTTP Headers issue, use a security token at all time
- If nothing in the world could determine you to filter every incoming message than the least you can do is to use a security token at all time. A light example of such a token could be: $token = md5(uniqid(rand(), true));

6 Conclusion

- To use SSL when authenticating users or performing sensitive operations.
- Regenerate the session id whenever the security level changes
- Have sessions timeout (time for which sessions expire)
- Store authentication details on the server, not to cookie
- Lock down access to the sessions on the file system, use custom session handling
- For sensitive operations consider requiring the user to provide their authenication details

7 References

1. Ballad, W.: Securing PHP Web Applications
2. Alshanetsky, I.: PHP Architect's Guide to PHP Security, Marco Tabini & associates
3. Murphy, C.: Security for Websites – Breaking Sessions to Hack into a Machine

Self- Describing globally accessible software components

Igorco Pandurski, Marjan Gusev

Institute of Informatics, Faculty of Natural Sciences, Skopje
Republic of Macedonia
pandurski@ii.edu.mk, marjan@ii.edu.mk

Abstract. The last decade, the world had been introduced with multiple and different Web technologies and development paradigms. The main idea-driver behind that boom is to create globally accessible software components. Web Services and the Web agents are the most utilized software components and are considered as self-describing and self-contained. While using meta-data containers interpreted via XML structure they are considered as self-describing and self-contained. But the question is, are they really self-describing and what are the all aspects that needs to be considered while self-describing labeling is taking place?

Keywords: XML, Functional programming, WEB Services, RDF, OWL, Semantic web, fXML, Semantic Web Services

1 Introduction

Globally accessible software components are considered all software components that possess an exposed interface, usually on the Web, and while using specific industry based standards and protocols can be embedded or integrated with other software systems that are independent from the location or the technology behind. With that, the logic or the business processes captured inside these software components are becoming an integral part of the new system.

In order to be able to perform such integration, the software components must possess features to be discoverable by the possible integrators. The model behind making the resources discoverable, today, is consisted of various architectures, technologies and protocols. The most known architecture is SOA (Service Oriented Architecture) and the technologies include Web Services, UDDI (Universal Description Discovery and Integration), WSDL (Web Service Description Language) and the SOAP (Simple Object Access Protocol). The sharing feature of all the after mentioned technologies is the XML (eXtensible Markup Language).

Having this in mind, the scenario goes like this: 1 A new system needs to integrate a function,2 The resources that poses that function are exposed on some of the UDDIs or the location is known 3 The new system lookups for the most appropriate resources that already posses that function checking up the WSDL file which posses the information about that function; 4 Integration is performed and specifically formatted messages are exchanged using SOAP and the paradigm behind globally accessible software components is fulfilled. Independently of the technologies behind every component of the chain uses XML structure as a markup language used for collaboration.

All this sound perfect and possible. And it is. Many of today's web resources are acting on that way very successfully. So, where is the challenge?

Well, the demand today for new web services that are incorporating complex functions in their structure, is rapidly increasing. WEB 2.0 based sites are requiring more than a service that calculates the currency exchanges, or conversion of Celsius degrees in Fahrenheit. They require some intelligence incorporated within, they need to provide you with an info which bank has the best exchange rate so the consumer can gain more benefits or, to suggest when to perform the exchange itself having in stack the stats of the currency flow in the last period.

The demand for exposing the currently incorporated business functions (processes and rules) in the legacy systems and making them globally available is also in search for appropriate exposure, having in mind the factor of global economy and entities mergers increases on a daily basis.

Incorporating a partial functions form one resource and partial functions from another while creating a new one is another hot topic.

Finding the right service that provides the right function is the real challenge. Lately, finding the right set of functions that can provide end-to-end process implementation in the application layer globally available is more suitable to be said.

2 Model for Self-describing Web resources

In the past, the software components written in different languages could have been very hardly integrated with other software components. There were too many obstacles that needed to be overrun in order to make this happen - complexity, preservation models, specific structural requirements, accessibility (communication patterns), proper ways of exposing the functions captured inside etc. There were too many initiatives and ideas how to make this happen. One of the streams was to develop a unified markup language and all the development frameworks and paradigms to develop a support for it. That is how the XML was born. Once the XML emerged on the surface, it was really lightweight document-like language and the rest of the obstacles were still remaining uncovered. The preservation model and the complexity behind the software components had to be addressed if all the business functions behind the specific software components want to be exposed and available for integration [Software as a Service (SaaS) paradigm] . The XML was about to be expanded with support for this as well.

Once this all was covered from one side, and from other side Web emerged as a global area where the software components are distributed and incorporated, mostly

as a Web Services, the real need and challenge become finding the exact function or set of function that you need. That means that the software components need to be self-describing and self-contained and the way how this can be done is adding new feature to the components – metadata. Metadata is binary information describing your software component that is stored either in a common language or is inserted into one portion of the file that is accessible to all the parties. In more technical jargon, metadata stores info about the identity (component, version, public key etc), the types that are exported, security permissions, attributes (base classes, methods, fields, nested types, elements etc).

Once the metadata is processed by the consumers, new challenge is faced again- what is the preferred method of processing of all the consumed web resources, and what is the real meaning, the semantics of the output. To address this challenge, a new model of the self-describing and the self-contained web resources is needed. That model is the Semantic Web Services and the Semantic fXML.

2.1 Self-Describe labeling

The label of Self-describing in order to be put as an attribute to a specific software component or service, the service itself need to go through a specific maturity model. It needs to support flexible explorations and interpretations while the information behind are comprehensive and enough explanatory for the consumers. On the figure 1 is explained the maturity model of a specific software component that need to fulfill before it is considered as a self-describing: to be standard machine readable, to support widely deployed formats, to inhere machine-processable specifications, to be grounded in web, to convey the RDF triple and finally a standard HTTP based algorithm to be used while deploying.

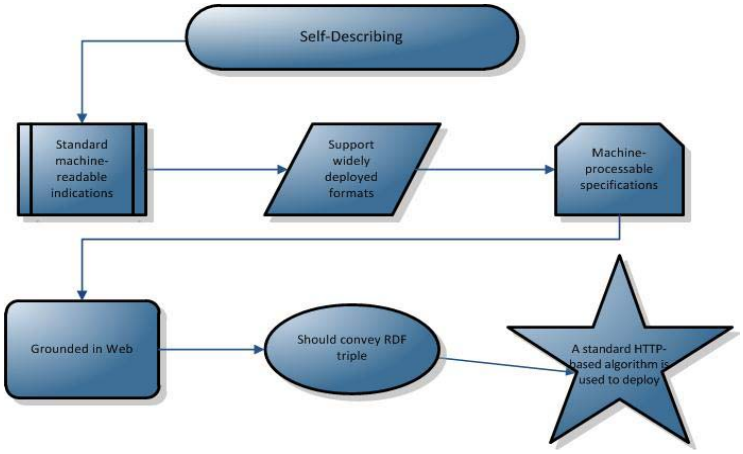


Fig. 1. Self-Describing Maturity Model

In particular, every stage means: [1] standard machine readable indications are the standards and conventions to be used , like XML encoding, HTTP content types

headers etc; [2] support for widely deployed formats- text/html, image/jpg; [3] machine- processable specifications – OWL ontologies; [4] grounded in web – all the specifications needed for proper interpretation of the information should be discoverable over recursively used links starting with the initial URl; [5] should convey the RDF triple (subject-predicate-object) – or simplified, the subject is denotation of the resource, the predicate is the denotation of the aspects of the resource and the object is the relationship between the subject and the object.

2.2 XML, is it really self-describing?

Having in mind the maturity model previously explained, the XML as we know it, is it really self-describing markup language? It is suitable to fulfill all the needs for publishing and linking the Web resources?

The elements and attributes provided in the XML structures are self-describing to some extent. They are providing the information for the usage, data types validated against a XML Schema and with that enforcing specific constrains. But the tags itself, doesn't give enough info about the semantics of the data, validation will pass and the XML document will be valid but you may still not getting the right data. Harry Halpin and Henry S. Thomson [1] are giving very visual explanation: Tag names are not enough. If a document is shared between two or more parties, there is an assumption that the meaning of the document is implicit is some common understanding (Halpin & Thompson, 2006).

Today, with the global market expansion, multiple integration projects are emerging and the parties are not even aware of each other. Agents, on behalf of the requestor go on the Web, searchers for function that he need within all the available Web Resources and the integration starts. So, if the XML based documents (WSDL files or RDDL documents) doesn't include all the information then the requestor will never know if this function or the set of functions are the right one he needs.

Therefore, XML as we know it, it is safe to be said that is not self-describing.

2.3 fXML

To make XML to be a self-describing structure, or at least to extend the level of self-describing, again, Harry Halpin and Henry S. Thomson are suggesting embedding functional programming structures in the XML. This new approach it is suggested a usage of an fx namespace where some of the elements and the attributes are no longer just data carriers but becoming functions or processors. That enables within the XML structure to embed variables, conditions, functions.

The intent of this paper is not to go in details and the ups and downs of this approach but to provide a view to all the efforts in making the XML really self describing. Therefore, we are saying the fXML is an approach that with attaching processors to the XML documents the level of self-describing is raised and the requestors can expect more quality based data that are matching the needs to acceptable altitude.

3 Semantic Web Services

Now, when the XML has reached the acceptable altitude of self-describing, the practice needs this in reality. Web Services, which are currently widely examined, researched and explained, are in their flourishing phase. The business are exposing their business functions (processes and rules) in Web Services and all the clients, independently if they are internal within the enterprise, or external and are located 10 000 miles away or next door, are consuming them just very easily. New systems that increase the productivity are emerging daily and are generating a new value for the owner.

Based on my experience, still these integrations are based on a mutual agreement between the parties, exchange of the necessary files, types and locations independently are we using cutting-edge technologies for implementations or not. Something is still missing. The human factor probably will be never excluded, but the human factor slows down the integration, increases the probability of errors and finally, humans are subjective, always approaching and addressing the challenges on the easiest way. The easiest way, I hope you would agree, is not always the most productive, the most stable en cetera. Actually, my experience is saying that this is almost never the case.

The idea is to expose all the business function into a "pool" of functions that are enough self-describing and then the web crawlers or the web agents to come, perform the search for proper function and do the integration. The human factor will be involved just in exposing the current functions into web resources and deploy them to the "pool". On the top of it, all the corporate, governance and security policies will be applied so all the confidential and business related secrets will be prevented from abuse.

As of now, the portion of available technologies and techniques and the paths of the search for proper concept end on the Semantic Web Services.

The most sustainable definition of the Semantic Web Services and in the same time enough self-explanatory without confusion is the one provided by the Semantic Web Services Initiative Organization [2] which is in the same time their mission - Infrastructure that combines Semantic Web and Web Services technologies to enable maximal automation and dynamism in all aspects of Web service provision and use, including (but not limited to) discovery, selection, composition, negotiation, invocation, monitoring and recovery. The only missing part in this definition is the intelligence. Semantic Web Services are distinguished by the remaining forms of Web Resource according one feature- The intelligence.

What is the relationship between the Semantic Web Services and the other forms of Web Resources?

In order to answer that question on Figure 2 is given an abstract relationship between the technologies and the domains that are considered as a Web Resources. The story starts with the expansion of the Web, when form human-centric environment transcended to software components-centric. This happens when the Static Web with its basics of HTML used for marking up documents and HTTP transferring them over the web, started to be shifted toward dynamic and opening the possibility of semantic way of marking up the documents which is centered on the previously mentioned metadata, ontologies, logic and autonomy. From this shifting;

web services were the result, which in a simplified manner are considered as software components that supports universal interoperability.

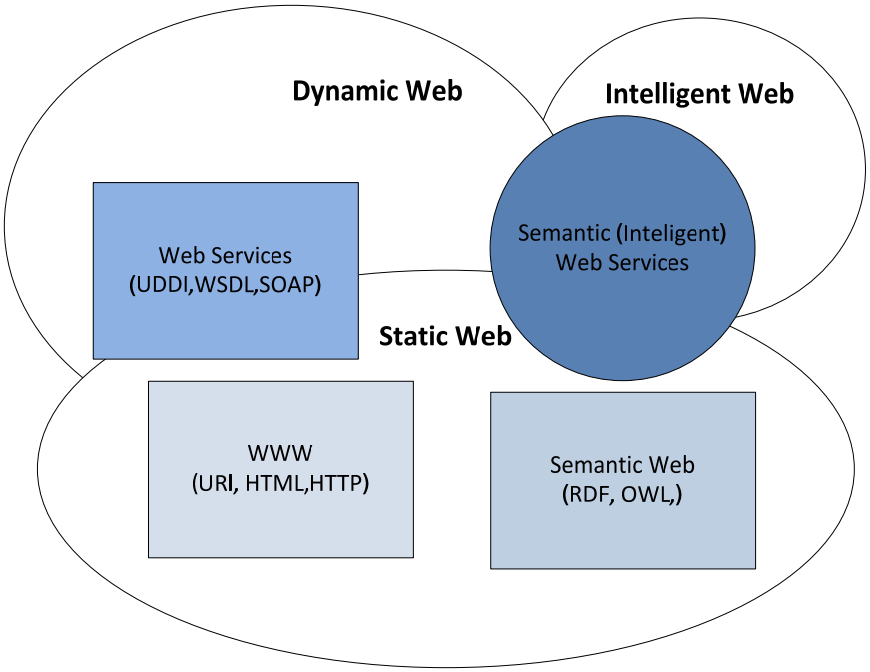


Fig. 2. Semantic (Intelligent) Web Services

Now, the missing link: the intelligence. Once the shifting form static to dynamic domain of the web became stabilized, the intelligence as an attribute was the new challenge.

Jagadeesh Nandigam [3] explains this challenge as synergetic confluence of the Semantic Web and Web Services which have the potential to provide value-added services by autonomously discovering and assembling Web Services to accomplish a domain task.

The way how this challenge is addressed we will explain the two main drivers that enable the addition of the Intelligence to the stack- RDF and OWL.

3.1 RDF

The shifting from Static Web towards Dynamic Web and lately towards Intelligent Web was in need of more suitable way of marking up the documents an enable the data interchange much easily and in the same time enable the implementation of the semantics. RDF (Resource Description Framework) is exactly that- it enables the data interchange even the data schemas are structured, semi-structured or even different.

The most suitable definition for RDF according my perception is the one that RDF is the language used for representing the metadata of the web resources.

The concept behind the RDF is having a URl (Uniform Resource Identifier) for every property and property value of the resource which when interpreted, statements about the resource as graph of nodes is generated as shown bellow.

```
<?xml version="1.0" encoding="utf-8"?>
      <rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-
syntax-ns#"

xmlns:paper="http://www.pandurski.com/paper#">

        <paper:Author
          rdf:about=" http://www.pandurski.com/about">
          <paper:fullName>Igorco Pandurski</paper:fullName>
          <paper:mailbox
rdf:resource="mailto:pandurski@ii.edu.mk"/>
          <paper:personalTitle>Mr.</paper:personalTitle>
          <paper:conference>ICT2010</paper:conference>
        </paper:Author>


      </rdf:RDF>
```

This piece of code can be interpreted as graph with multiple nodes:

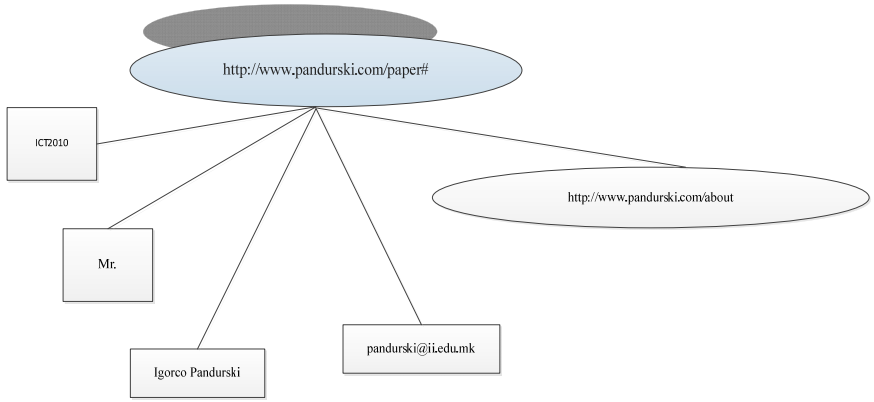


Fig. 3. Semantic (Intelligent) Web Services

This primer actually confirms that the RDF is an extension of the XML and the structure itself is even called RDF/XML.

3.2 OWL

Following the shifting from Static Web to Dynamic and lately to Intelligent Web the metadata is covered with RDF; it remains to cover the ontologies, the logic and the autonomy as paradigms.

Ontology is a visualization of the concept of the domain where the specific resources exists. In this case Web Resources. This visualization of the concept is nothing else but capturing the taxonomy of the environment and expressing it thru a machine readable language.

OWL (Web Ontology Language) is a language developed for expressing the ontologies while using URl for naming and description framework. OWL posses few more distinctive features: ability to be distributed across many systems, it is scalable, it is compatible with the Web Standards (many other ontology languages are not), it is open and extensible. OWL is build on the top of XML and RDF with expanding the vocabulary which can very easily deliver the semantics – relation between the classes' types of properties, equality, cardinality etc.

We are not going to examine the gods of the OWL language but we will show how the ontologies are build- form declaring the namespaces, building simple classes and defining properties:

```
<?xml version="1.0" encoding="utf-8"?>
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-
syntax-ns#"

xmlns:paper="http:"//www.pandurski.com/paper#
         xmlns:owl="http://www.w3.org/2002/07/owl#"
         xmlns:rdfs="http://www.w3.org/2000/01/rdf-
schema#"
         xmlns:xsd
="http://www.w3.org/2001/XMLSchema#">


 <owl:Ontology>
    <rdfs:comment>An example OWL
ontology</rdfs:comment>
       <owl:imports
rdff:resource="http://pandurski.com/document"/>
        <rdfs:label>Sample Ontology</rdfs:label>

        <owl:Class rdf:ID="Paper">
        <rdfs:subClassOf
df:resource="&document;PotableDocument"/>
        <rdfs:label xml:lang="en">Self-Describing
globally accessible software components</rdfs:label>
          <rdfs:subClassOf>
              <owl:Restriction>
```

```
                    <owl:onProperty
rdf:resource="#canBeDistributed"/>
                     <owl:minCardinality
rdf:datatype="&xsd;nonNegativeInteger">1</owl:minCard
inality>
                  </owl:Restriction>
              </rdfs:subClassOf>
            </owl:Class>

  </owl:Ontology>
```

Once the ontologies are built, multiple, autonomous software agents can be build that can crawl over the web resources and provide the requestor with specific function or set of functions. Having in mind the previous ontology it can be a part of a Web Resource that stores scientific or best-practices papers and blueprints. Than the web agents can come autonomously, search over the ontologies and find the right document that treats the self-describing as subject.

4 The vision for Globally Accessible Software Components

The most interesting part from always for me was- how to implement the achievements in the science into the real life world and generate a new value for the clients. Being a part of the happenings within the financial world in US and being a leader of some of the biggest integration software projects , the limitations of the technologies used for building the legacy systems, the time consumption in creating appropriate interfaces for the functions behind and finally integrate the systems is frustrating. The projects itself has a scope, has a timeline and has a stakeholders. When I am saying stakeholders I mean a limited budget. Now, I am facing legacy systems that are robust, non-flexible, non scalable, distributed, practically indefinite scope and a limited time and budget. Whatever you decide to play with, you will face a wall. If you are on time with building the interfaces of the functions than, you have no time and budget for fully implementations and integration with the other systems. Or if I put a balance between all the three peers, you will never succeed in fulfilling the scope and the stakeholders will go crazy.

Therefore, while the projects are ongoing I am in a search for appropriate and cutting-edge technology or technique that will provide me with abilities to safe a time on creating the interfaces, increase the portability and quality and of course, to satisfy the scope.

That is how the ideas of generating a “pool of functionalities” are born- The pool will represent all the functions captured behind the legacy systems and enable them for the software agents to be searchable and integral, independently if the poll is going to be clouded within the company, on the public web or some of the registries available today.

Once this criterion is satisfied, every integration will be much faster, will assure that the scope is fulfilled, the timelines will be preserved and the numbers of

integrations is practically unlimited. Re-usability of the code is another great feature of the Web Services, in this case Semantic Web Services.

The market today goes thru a phase of global merger and consolidation; companies with a similar core business are merging and the data including the functions aggregated within the decades are enormous. Very often, the core-business is not within the software industry, but always are depending on the software and the paradigms of the software – Enterprise Application Integration (EAI) or Inter-organizational systems (IOS).

The globally accessible software components are the most suitable way of performing (1) Legacy-Systems Integration, BPI (Business-Process Integration) and Software as a Service (SaaS) implementations.

4.1 Legacy -Systems Integrations

"Over the years, IT environments have become more complex and more heterogeneous due to diverse customer needs and rapid innovation in the IT industry. Many government and enterprise customers require integration with legacy systems to maintain critical business and operational processes." – Microsoft [4].

This definition explains very visually what the legacy-systems integration all about is. The integration is usually performed, but not limited to, due to potential problems and non-compatibility with the emerging trends, improvements, maintainability, replacement and the lack socially addictiveness. At this time, having in mind the shimmering budgets [5] for the ICT, the Semantic Web Services are the key players in this legacy –systems integrations.

The systems that are integrating are starting from mainframes ending up to corporate websites. In between pleads of applications (both, web oriented and desktop) are passing thru this phase.

4.2 Business Process Integration (BPI)

With the legacy-Systems integration many of the challenges that today's business are facing are addressed- transparent and real-time access to information for the entire enterprise. The functions captured inside are exposed and the business continues to use the functions form all the systems behind. Now, having the functionalities, every new merge requires accommodation of specific business processes and business rules. To avoid the processes duplication and inconsistency which can result in ineffective and inaccurate decision-making, business process integration takes places. How the Semantic Web Services are making this happen?

Well, the "pool of functionalities" first eliminates the silos of stand-alone applications. Now they are all connected. On the top of the pool, BPEL process (Business Process Execution Language) is applied. The BPEL process has the complete same pattern as the Web Service, except that has a possibility to embed graph-like regimes, manipulation and decision-making functions and transactional processes. The BEPL process can embed different Web Services or different Semantic Web Services and apply composition or orchestration over them and with

that specifies and exact order of services invocation and satisfies the desired business process workflow (Figure 3).

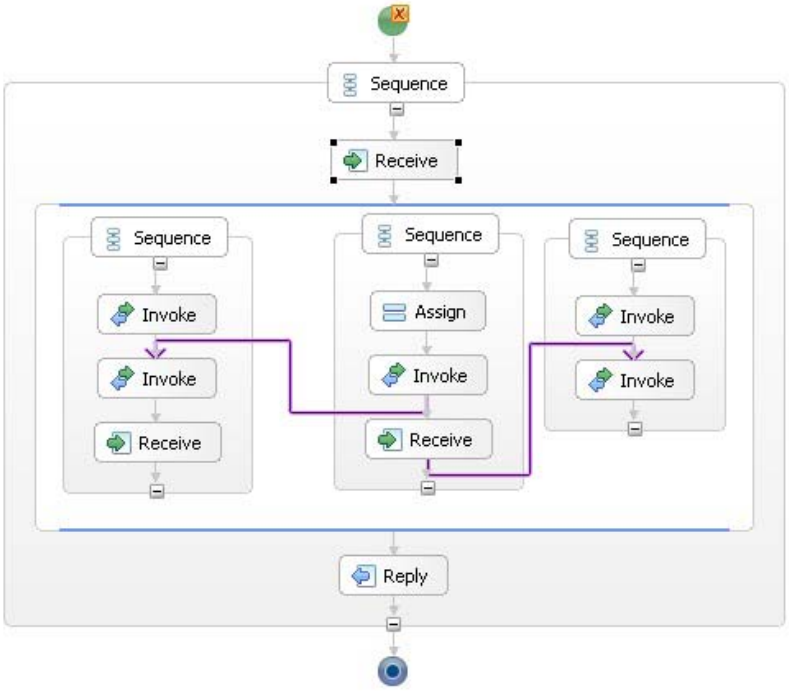


Fig. 4. BPEL Process

4.3 Software as a Service (SaaS)

One of the hottest topics on the Google I/O 2010 [6] was the SaaS topic. In general terms SaaS is software components deployed on the Internet. That concept is nothing new, but what distinguish the SaaS from the other Web Resources is that gives end-to-end execution. The components are accessed remotely and the benefits are that no infrastructure or specific deployment is needed.

On the Google I/O 2010 was exposed a new HTML5 standard that the entire browsers family are going to implement it in the near future. Most of them are already supporting it. For the business SaaS means availability of business applications, including collaboration software and line-of-business applications that require them to run their business without an infrastructure or specific development. Most of the times it SaaS are commercial solutions and all the possible clients need is an Internet and appropriate browser.

This achievement incorporates all what is stated in this paper and it is subject of future deeper analisis.

5 Future work and conclusions

This paper should give us the idea how the Web Resources while using the basis of XML are going to real self-describing attribute and are passing the maturity model. Once the maturity model is satisfied, all the pre-requisites are achieved and the shifting of the Static Web towards Intelligent via Dynamic is possible. This progress flow of the technologies, techniques and software paradigms are implemented in the real life world and are enabling the business to move forward.

With that said, it is safe to be concluded that the Semantic Web Services are the solution in need of a problem.

The Semantic Web Services and the properties they pose can be considered as really self-describing globally accessible software components.

The future works are to examine the benefits of applying SaaS model in the real business and what is the scientific approach behind it.

References

1. Halpin, H., and Thompson, H. S.: One Document to Bind Them: Combining XML, Web Services, and the Semantic Web. Proceedings of the 15th international conference on World Wide Web, 2006
2. The Semantic Web Services Initiative (SWSI), http://www.swsi.org/
3. Nandigam, J.: School of Computing and Information Systems, Grand Valley State University, Allendale, MI, 2005
4. Legacy System Integration, Link
5. Gartner EXP Worldwide Survey, http://www.gartner.com/it/page.jsp?id=1283413
6. Google I/O 2010, http://www.google.com/events/io/2010/

Architecture of a Identity Based Firewall System

Nenad Stojanovski[1], Marjan Gušev[2],

[1] Makedonski Telekom AD, Orce Nikolov BB, 1000 Skopje, Macedonia
[2] Faculty of Natural Sciences and Mathematics, Ss. Cyril and Methodius University
Arhimedova b.b., PO Box 162, 1000 Skopje, Macedonia
nenad.stojanovski@telekom.mk, marjan@ii.edu.mk

Abstract. Classic firewall systems are built to filter traffic based on IP addresses, source and destination ports and protocol types. The modern networks have grown to a level where the possibility for users' mobility is a must. In such networks, modern firewalls may introduce such complexity where administration can become very frustrating since it needs the intervention of a firewall administrator. The solution for this problem is an identity based firewall system. In this paper we will present a new design of a firewall system that uses the user's identity to filter the traffic. In the design phase we will define key points which have to be satisfied as a crucial milestone for the functioning of the whole Identity based firewall system. The design process includes the process of researching a possibility to create an agentless identity based firewall system. During this process we explored the whole logon process in a Microsoft Windows domain. Based on the results from the logon process we designed the architecture of the agentless identity based firewall. As a result from this architecture we are able to define the key components of the identity based firewall solution. These components are the core components of the system and they will provide the functionality of the identity based firewall. Based on the newly architecture, we were able to roughly compare our design with some existing solutions that can be found on the market and based on that comparison we can show what the benefits from our solution will be.

Keywords: Identity based firewalls, user identity, firewalls, network security, computer networks, firewall systems design.

A New Methodology to Analyze Micro Assessment Solutions

A. Bundovski, M.Gusev

University Sts Cyril and Methodius, PMF Institute of Informatics,
Arhimedova b.b. Skopje, Macedonia
bundovski.aleksandar@gmail.com, marjangusev@gmail.com

Abstract. In this paper we analyze web service oriented solutions for micro knowledge assessment. Our research discovered only two existing solutions "Knowledge Accelerator" and "Knowledge Pulse" to provide the customer with the customization program and information needed, especially adjusted for the purposes of the company's field or business. These two solutions are certainly developed by followers of the "lifelong learning" concept, needed both for individual and professional upgrades. As such, the "lifelong learning" tools, do not only enhance individual's involvement in social issues, but they are also enabling them with competitive skills required on their workplace. In this paper we present a new methodology to evaluate micro assessment solutions, based on definition of key performance indicators in three categories (technical, functional and usability requirements). For this purpose we define scoring schema and average quality value estimation.

1 Introduction

1.1 Motivation

The constant development of the information society is surely automating and shortening each business process in one company. Due to the recent economic crisis, most of the managers and employers are trying most of their efforts in reducing the costs and becoming more efficient, thus investing in new software solutions, which will also cut the expenses on company's payroll. Moreover, the efforts of upgrade and strengthening of the whole value chain and supply chain is highly appreciable. Therefore the available workforce should constantly be exposed on acquiring new knowledge on solutions and services. It that can be easily adjusted and utilized for the specific field and business within the company.

The system for "micro knowledge and skills control" allows the manager get information and status on knowledge gained on behalf of the employees, on very short status cuts during the working hours, rather than scheduling special meetings with them in order to asses they gained knowledge. It has been proved that this kind of micro testing saves a lot of time, does not cause stress between employees and allows more time to be dedicated for learning.

Information society development is the result of reducing the disadvantages of ignoring the constant lack of information. With the increasing volume of available information, the main challenge becomes finding the correct and necessary information, and its adaptation to specific needs. Tools discussed in this document, "Knowledge Accelerator" and "Knowledge Pulse", are exactly the tools needed on behalf of the employees in approaching new trends and the required information tailored to the relevant activities of the company. Additionally, these tools do not provide adequate information but only a short test to assess employees' knowledge.

Finally, the two instruments and newly introduced services, are based on the world trends in teaching called "lifelong learning" - a need for continuous personal and professional learning. As such, lifelong learning does not only improves the social inclusion of one individual, but his competitiveness of his job.

1.2 **Definition**

The goal of our definition of a perfect micro assessment system is a system that should be implemented as a software service for micro assessment of skills and knowledge of employees in order to increase and improve the productivity of the company, execution and knowledge of users, adding additional value to their professional development.

According to our research of the "knowledge platform" market today, there are several systems for electronic learning, electronic testing and verification of knowledge, however those who meet the functionalities of Micro knowledge and skills assessment are SAP ® BusinessObjects ™ Knowledge Accelerator and KnowledgePulse ®.

2 Micro assessment solutions

2.1 Perfect micro assessment system

The system for micro assessment of the knowledge and skills, primarily needs to be implemented as a **service set by the company, i.e. set on the vendor's server** managed by the user and the vendor himself. The **educational materials will be deployed with simple human interaction** either through local network or Internet.

The service should be placed on vendor's side which will allow easier maintenance, provided at one place only, and all companies will be served with the latest version that will not require additional installations of the user. On the client (manager, employee and administrator) **access to the service** will be through the web site which will be announced via Web browser. The system should automatically **run** when the user will be announced on the domain of the company. Preferably the system should easily **integrate with another system, service**, and there should be **a possibility for disabling**.

All **changes to existing as well as adding new material** (questions and documents) are available through the Web site. Also **through the Web site all materials can be accessed**, activation of testing can be done, changes of personal

data can be made, shown by test scores, charts, statistics, etc.. The system should allow **adjustment of learning by the student's current knowledge**, as well as **review and download reports on all user activities**, including performance, efficiency, duration, etc..

The main functionality of the system is to be **able to automatically activate** if the process from chosen application, keyboard and/or mouse are inactive for a certain period of time. The service is activated as a **window, plug-in and/or screen saver**, which can contain a question or some useful information. These windows should be **able to delay, prolonging for the specified time** if the user does not want to answer the question at that moment.

The critical condition for choosing between the two solutions, most of the times is based on the level of satisfaction gained on behalf of the users. To make the system closer to users should **multi language support**, and to provide **full technical support and documentation**, video presentations, forum etc.. Assessment of satisfaction from using the system itself includes the opportunity to **survey consumers** as well as inquiring for other topics as defined by managers.

Other functionalities that the system should be able to provide:

- Display of information form an uploaded document (text info), when the system is inactive,
- Answering questions of succession,
- Access to materials through the Web browser,
- Delivery of materials (documents and questions) in a specific format,
- Testing of the entire company with one test,
- Monitoring of inactivity in more than one process,
- The manager can change the parameters for activation of the matter (time of inactivity, a process (s), keyboard, mouse, etc.),
- Testing of the company divided into groups according to the workplace (different tests)
- The test consists of questions from several topics,
- Save and display the answers from questions for each employee.

2.2 SAP® BusinessObjects™ Knowledge Accelerator

SAP

SAP is a multinational consulting corporation, which deals with the development of software solutions for companies and supporting businesses of all sizes. In addition SAP offers a large range of solution for monitoring, understanding and managing business performance. Besides all these services, SAP offers micro assessment solutions and in this paper special focus will be put on SAP BusinessObjects Knowledge Accelerator. [1, 3, 4, 5]

User Interface

The learning environment in SAP BusinessObjects Knowledge Accelerator is divided into different modules and includes quick-reference material based on different learning objectives, business processes and tasks. Each module has specific concept, practice exercises and skill tests. This allows users to focus only on the functions and

concepts they need to know. SAP BusinessObjects Knowledge Accelerator trains users via job-specific skill development scenarios based on sample company data. In addition, it is customizable, so it meets the requirements and relevancy of each user group, including varied training needs and tasks to be performed.
SAP BusinessObjects Knowledge Accelerator is strictly intranet based tool that enables companies to quickly deploy educational materials via the Internet, server or CD. This further makes it easier for users to access manuals and other materials at any time and at any place.

Cost-Effective Learning
The implementation of a new Business Intelligence solution typically impacts a diverse range of user groups. This can make training and support overwhelming and costly. With SAP BusinessObjects Knowledge Accelerator, you can affordably train and support a few hundred or several thousand users at their own computers. Additionally, it is extremely important provision of separate training (included in the tool) available 24 hours a day. It is fully scalable and features extensive flexibility.

Better Productivity and Performance
SAP BusinessObjects Knowledge Accelerator speeds up user productivity, performance and knowledge, helping users add more value. This means faster decisions, better response to issues impacting critical business drivers and more consistent and accurate interpretation of data presented via reports and dashboards.

2.3 RSA KnowledgePulse®

MicroLearning and Knowledge Pulse
MicroLearning is an innovative approach of learning that is integrated with daily activities and allows users continuously taking classes, learning and testing. MicroLearning maintains and implement the Knowledge Pulse tool which is developed based on the belief that "anything that you are aware exists, does not means that you know it". Knowledge Pulse as a software tool provides an opportunity to overcome a different material, such as foreign language, training for project management, training for product development or certificate required for promotion. The idea is that the absorption of knowledge to occur through small steps, which would avoid boring procedure of permanent and extensive learning, which is usually the result of eLearning (e-learning covers all forms of advanced learning technology such as Web-based training) systems. [2, 6]

Further education must pay off
Knowledge Pulse is special in the sense that its customers offer special opportunity to learn just what they need. The content of the material is easily adjusted according to the needs of students and the employer, and is built through small steps of learning.

No additional effort required!
The innovative MicroStep Manager system is installed on the server, which directs and records activities and progress of each client separately. Each client can benefit

from these features, however managers are also benefiting by having opportunity to monitor how the knowledge is absorbed by each employee.

Knowledge Pulse – offers multiple opportunities
Knowledge Pulse has an opportunity for constantly checking the progress and the level of knowledge, proceeding with the lesson to which the client stopped last time, and so on. Knowledge Pulse is easily portable to almost any IT infrastructure, which remains at a modest cost.

3 Comparison methodology

Today there are many solutions that are more or less similar to each other. Both penetrate the market through the shortcomings of others and vice versa.

There is no methodology to evaluate similar solutions and explain introduction of perfect micro assessment solution.

Methodology in this paper will introduce analysis and comparison that involves finding existing tools using web browsers, although we are aware that there aren't ideal tools or in a larger percentage to give what we need. Therefore, in this paper more quality indicators for comparison will be included. Evaluation criteria are grouped into three categories:

- technical requirements,
- functional requirements,
- usability requirements

3.1 Technical requirements

Technical requirements (also known as non-functional requirements) are including technical aspects (hardware and software) that the system should meet. Such claims are as follows:

1. SAAS (Software As A Service) – software as a service set by the company:
    - 0 = not realized as a service,
    - 1 = service that is installed on the machine and has a limited base of materials (not connected to the Internet and can be added to other materials),
    - 2 = web application which does not requires notification of the user,
    - 3 = web application that requires notification of the user,
    - 4 = service installed on the server of the buyer,
    - 5 = realized as a software service.

2. Installation – a way of deployment of educational materials and tools:
    - 0 = requires a full installation as a separate application,
    - 1 = via Web browser, requires additional installation,
    - 2 = thick client installation,
    - 3 = thin client installation,

- 4 = not requires a separate installation of the client, but only a small tool that is placed in the first initialized,
- 5 = requires no installation, just web browser.

3. Need a server - the tool is installed on your own server:
   - 0 = no need for server,
   - 1 = server managed by the vendor,
   - 2 = server managed by the user,
   - 3 = server in vendor managed by vendor,
   - 4 = server in vendor managed by the user.
   - 5 = server in vendor managed by the user and vendor.

4. Web browser - only applies to the Web site that can log the user:
   - 0 = no browser support possible,
   - 1 = application suitable for installation,
   - 2 = working with specified browsers,
   - 3 = not work with older versions,
   - 4 = requires additional installations of the browser,
   - 5 = works with all browsers.

5. Remove, stop and re-install, launch the service:
   - 0 = once installed permanently in your computer
   - 1 = can be removed from the computer, but can't be re-installed and start again,
   - 2 = can be removed from the computer and start again,
   - 3 = can be stopped and start again,
   - 4 = can be removed or stop by the administrator and start again,
   - 5 = can be removed, stopped and start again by the user.

6. Integration with other software - a standalone service, application or part of a bigger software:
   - 0 = does not support data exchange,
   - 1 = manual settings are necessary,
   - 3 = need technical support for integration with other service,
   - 5 = easy integration with other services.

3.2 Functional requirements

Functional requirements represent functions implemented in software that can be calculations, data processing and other specific features that define what the system should achieve. Such features are:

1. Announcing - the user is announcing to the application or the application automatically recognizes the user:
   - 0 = does not support logging,
   - 1 = user must be announced to the system every time its run,
   - 2 = user announces a Web page and registers with the IP address,

- 3 = user is announced only at the beginning of the day,
- 4 = user activates the system and it automatically recognizes it,
- 5 = system is automatically activated when the user will be announced on the domain of the company.

2. Customization - the opportunity to learn just what they need:
   - 0 = cannot be adjusted,
   - 1 = can make selection of materials,
   - 2 = can make choice questions,
   - 3 = can do selection of questions and materials,
   - 4 = full adjustment (time, materials and questions),
   - 5 = allows semantic solution - adapt the system to learn the current knowledge of the student.

3. Activation - how to activate the tool and materials (initiated by the user, screen saver, a process in background, etc.):
   - 0 = system is activated without logging on Web site,
   - 1 = system is activated with a logging on Web site,
   - 2 = system automatically activates at specified period of time,
   - 3 = system is activated with application startup,
   - 4 = system is automatically activated after inactivity on the keyboard and mouse
   - 5 = system automatically activates after inactivity of the process of the selected application, keyboard and mouse.

4. Manipulating with materials (documents and questions) - adding new teaching materials, areas for study and testing:
   - 0 = cannot change the learning content,
   - 1 = changes can make site administrator (vendor),
   - 2 = changes can make site administrator (client),
   - 3 = changes can make manager and administrator with restriction,
   - 4 = changes can make manager and administrator without restriction,
   - 5 = system allows changing the curricula, tracking versions of the curriculum and changes etc..

5. Measuring the activity of the user (reports) - the user is more active in work or in learning:
   - 0 = system does not store the user's actions,
   - 1 = system provides raw (unprocessed) data about the user activities,
   - 2 = system allows simple statistical graphs,
   - 3 = system provides data and graphs related with user answers, accuracy of responses, frequency of questions, etc.,
   - 4 = system allows review of reports on all activities of the user, including performance, efficiency, duration, etc.,

- 5 = system allows review and download reports of all activities of the user, including performance, efficiency, duration, etc..

6. Delay, prolongation of the specified time - if the user does not want to answer the question at this moment, the question should be able to delay for a specified time:
   - 0 = delay not supported, the question is displayed on the screen over other applications (prevents use of other applications) until the answer,
   - 1 = delay not supported, the question is displayed on the screen until the answer, but can use other applications,
   - 2 = can be delayed for a fixed time by the service (eg: 5 min)
   - 3 = can be delayed for one of the opportunities offered by the service (eg 5, 10, 15 or 20 min)
   - 5 = can be delayed for a time entered by the user.

7. Basic functional requirements - each of the above functionalities have difficulty 1 (Ex. If software meet the above 3 functions then total score is 3):
   - Show info from uploaded document, when the system is inactive.
   - Answering questions succession.
   - Access to materials through the Web browser.
   - Download materials (documents and questions) in a specific format.
   - Testing the entire company with one test.

8. Advanced functional requirements - each of the above functionalities have difficulty 1 (Ex. If you meet the above 3 functions then total score is 3):
   - Monitoring inactivity for more than one process.
   - The manager can change the parameters for activation of the matter (time of inactivity, a process (s), keyboard, mouse, etc.).
   - Testing the company divided into groups according to the workplace (different tests).
   - The test consist questions from different topics.
   - Preserve and display answered questions for each employee.

3.3 Usability requirements

The critical condition for choosing between the two solutions, most of the times is based on the level of satisfaction gained on behalf of the users. Such indicators are the following:

1. Survey - an opportunity for the implementation of surveys defined by the manager:
   - 0 = not supported,
   - 1 = supports general surveys contained in the system (using satisfaction, self evaluation)
   - 2 = supports surveys on Web page,

- 3 = supports surveys started as a service (window, screen saver, plug-in),
- 4 = supports surveys defined by manager,
- 5 = supports surveys that are activated for all employees at given time.

2. Language Support – how many languages the solution supports?
   - 0 = no language support,
   - 1 = more than one language,
   - 2 = limited linguistic support,
   - 3 = more language support - the EU,
   - 4 = support of the Macedonian language,
   - 5 = full language support.

3. Maintenance and technical documentation - Documentation for using the service and Web site, and support at any moment:
   - 0 = no technical documentation,
   - 1 = no support,
   - 2 = frequently asked questions,
   - 3 = maintenance and technical documentation,
   - 4 = full technical support, video presentations, forum
   - 5 = live support.

4 Software comparison

We have defined about twenty qualitative indicators that will evaluate software solutions, but not all have the same weight in the decision which of these solutions is better. Indicators that describe the type of software solution (SAAS), the manner of setting the company, access, showing how the questions will be assessed will be graded with factor 2. With the 1.5 factor will be evaluated indicators typical for access to data through Web, adjustment of the system according to the wishes of the user. The basic functionality that should be part of the software solution will be awarded with factor 1.

For the indicators that we cannot assess the value N / A is put. If several indicators are graded with N / A, then those indicators are not taken in the overall assessment. Table 1 represents matching between system requirements and factor groups (1, 1.5and 2).

Table 1. Matching between system requirements and factor groups.

	Factor 1	Factor 1,5	Factor 2
Technical requirements		4, 5	1, 2, 3, 6
Functional requirements	6	2, 7	1, 3, 4, 5, 8
Usability requirements	2, 3	1	

Based on the analisys of the software sollutions metionted above and the application of indicators, which were gained from usage of benchmarks [7] the following comparisons were obtained:

Table 2. Technical requirements.

Technical requirements	Knowledge Accelerator®	Knowledge Pulse®
SAAS (Software As A Service)	3	4
Installation (CD, server, Internet)	4	4
Need a server	2	2
Web browser	5	5
Remove, stop and re-install, launch the service	5	5
Integration with other software	5	3
Average value	4	3.83

Table 3. Functional requirements

Functional requirements	Knowledge Accelerator®	Knowledge Pulse®
Announcing	1	3
Customization	5	5
Activation	3	4

Manipulating with materials (documents and questions)	5	5
Measuring the activity of the user (reports)	3	4
Delay, prolongation of the specified time	2	4
Delay, prolongation of the specified time	4	4
Advanced functional requirements	3	3
Average value	3.25	4

Table 4. Usability requirements

Usability requirements	Knowledge Accelerator®	Knowledge Pulse®
Survey	0	0
Language Support	3	3
Maintenance and technical documentation	5	5
Average value	2.66	2.66

In the following Table 5 the grades of the grouped functionalities with average value are presented:

$$x = \frac{TR + FR + UR}{IN}$$

TR = Technical requirements
FR = Functional requirements
UR = Usability requirements
IN = Instruction numbers

Table 5. Table with grades of the grouped functionalities.

Overall	Knowledge Accelerator®	Knowledge Pulse®	Total by category (factor)
Technical requirements	24	23	30
Functional requirements	26	32	40
Usability requirements	8	8	15

Overall	3.41	3.70	(TR + FR + UR) / IN Maximum value should be 5
In percentage (Maximum value 100)	68%	74%	100%

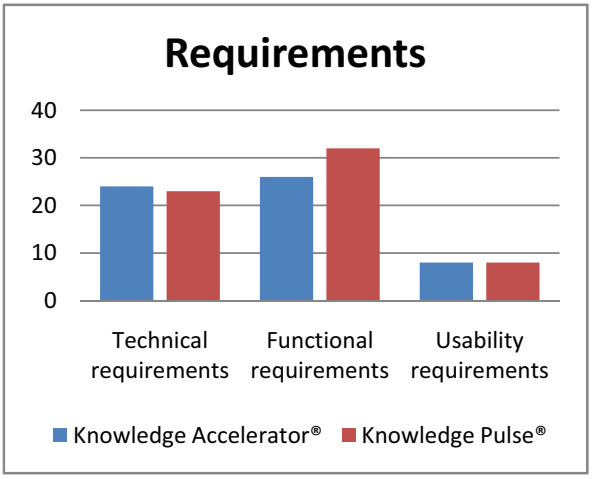


Fig. 1. Column chart with grades for each type of requirements.

In the following Table 6 the grades based on the categories of indicators together with their factors:

$$x = F1 * 1 + F2 * 1.5 + F * 2$$

F1 – sum of grades for indicators from faktor1 group
F2 – sum of grades for indicators from faktor1.5 group
F3 – sum of grades for indicators from faktor2 group

Table 6. Grades based on the categories of indicators with factors.

Overall	**Knowledge Accelerator®**	**Knowledge Pulse®**	**Total by category (factor)**

Factor 1	10	12	15
Factor 1,5	19	17	25
Factor 2	29	32	45
Overall with factors	96.5	101.5	15 * 1 + 25 * 1.5 + 45 * 2 = 142.5
In percentage (Maximum value should be 100)	67%	71%	100%

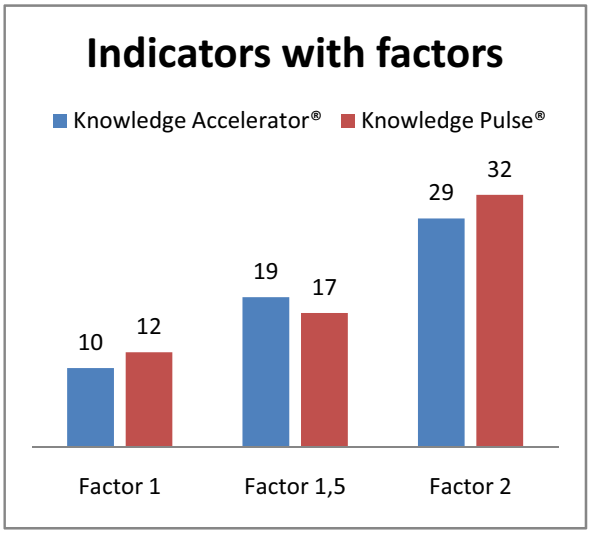


Fig. 2. Indicators with factors

On Figure 3 are presented both grades with and without factor. Based on the figure it can be concluded that both solutions have very similar grades, thus meaning that they also have similar functionalities.

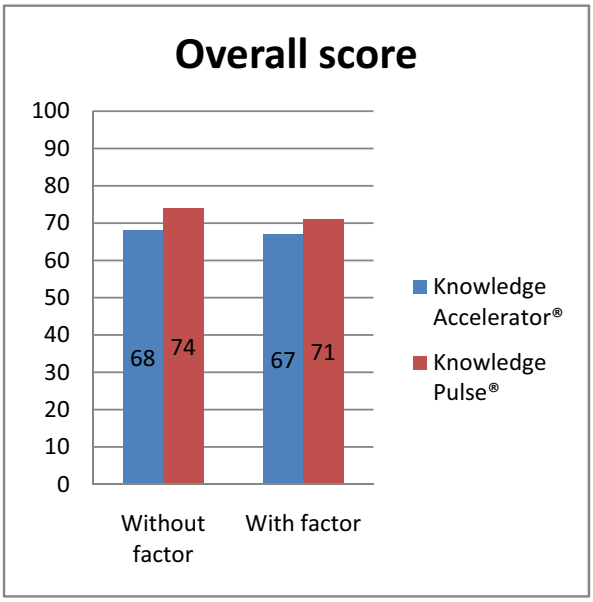


Fig. 3. Overall score with and without factor

5 Conclusion

In this paper we introduced a definition of a perfect system for micro assessment of skills and knowledge. This system should be implemented as a software service that can run independently of the platform, the instructor can easily configure teaching materials and questions for knowledge assessment and skills to be able to join the existing systems for e-learning, etc..

The main aim of this paper is an analysis and comparison of this kind of tools that increase productivity and improve the company's efficiency and knowledge of users, adding additional value to their professional development.

Based on the analysis and comparison on the existing solutions at the market, can be concluded that there are products that enable and facilitate upgraded and perfecting of the employees in a company and whether through web services or using desktop applications. Tools discussed in this document, "Knowledge Accelerator" and "Knowledge Pulse" are primarily tools for large companies that serve large numbers of users who have high hardware requirements and high prices. These tools are untouchable for small businesses, who need readily available and cheaper version of this type of tool that would be simple to implement using Web services.

In this paper we introduced a methodology for assessing this type of systems by defining indicators to compare performance. Using this methodology to compare the existing solutions on the market, it can be concluded that none of them satisfies the requirements set for our system for micro assessment of skills and knowledge.

References

1. SAP AG, Overview, http://en.wikipedia.org/wiki/SAP_AG, 25.12.2009.
2. Microlearning, Overview, http://en.wikipedia.org/wiki/Microlearning, 25.12.2009.
3. SAP Education, SAP Businessobjects Knowledge Accelerator, http://www.sap.com/services/education/softwareproducts/knowledge-accelerator.epx, 25.12.2009.
4. BUSINESSOBJECTS KNOWLEDGE ACCELERATOR, http://www.alteksolutions.com/library/education/ka/ka_datasheet.pdf, 25.12.2009.
5. SAP® BusinessObjects™ Knowledge Accelerator, Maximize Productivity with Effective Software Training, http://download.sap.com/industries/mining/large/brochures/download.epd?context=40B716710AA654CFA15C8C21D796C8F7F3AD1E639B5E33BA40BE48C9D0B950180C32C2D99600DAC318BC3445901B2C2E54309614C367101A, 25.12.2009.
6. The KnowledgePulse®, Overview, http://www.knowledgepulse.com/home_en.html, 25.12.2009.
7. How To Evaluate and Select Order Management Software, Lloyd Merriam (CoLinear Systems, Inc.), http://www.colinear.com/evaluatesoftware.pdf, 15.01.2010.

Distributed Transaction Processing in Mobile Computing Environment

Tome Dimovski and Pece Mitrevski,

University of St. Kliment Ohridski, Faculty of Technical Sciences, Ivo Lola Ribar bb, 7000 Bitola, Macedonia
{Tome.Dimovski,Pece.Mitrevski}@uklo.edu.mk

Abstract. Mobile embedded systems increasingly use transactions in applications like mobile commerce, banking or other commercial applications. A transaction in a distributed environment may involve multiple parties, data servers, where its operations are executed. In these transactions, besides fixed ones, multiple mobile devices can be involved as full participants.
In the execution of transactions the key issue is the protocol that ensures atomicity. Transaction Commit on Timeout (TCOT) protocol decreases the number of wireless messages, but does not consider mobile hosts as active participants in the execution of transactions. Two-Phase Commit Protocol for Mobile Wireless Environment (M-2PC) requires simultaneous connection of all mobile participants at the beginning of a transaction and does not provide adequate management of mobility and failures caused by network disconnection, nor a mechanism to control the competitiveness of distributed transactions. Fault-Tolerant Pre-Phase Transaction Commit (FT-PPTS) protocol, as well, does not provide adequate management of mobility, i.e. when mobile hosts are disconnected from the fixed network they block resources on the fixed participants for an undefined period of time, leading to an increased number of mobile transaction aborts. Similar efforts include "concurrency control without locking" (Moiz and Nizamudin), "scheduling transactions in mobile distributed real-time database systems" (Xiang et al.), etc.
In this paper we propose a communication model in mobile environment where mobile hosts ad-hoc communicate with each other and communicate with the wired network via wireless channels. We consider system model for a mobile distributed environment consisting of a set of mobile hosts (MH) and a set of fixed hosts (FH), where MHs communicate with FHs through Mobile Support Stations (MSS) via wireless channels. In addition, MHs communicate directly with each other via wireless channels, and as soon as they enter a geographical area out of the coverage of any MSS, they try to connect to other MH which is in the covering area of any MSS.
We present the execution of a mobile transaction under the proposed scenario. We expect that mobile transaction execution in our model will reduce the blocking time of resources at the fixed devices, provide fast recovery from failures owing to mobility of mobile devices and increase the number of committed mobile transactions.

Keywords: Distributed transactions, mobile computing environment, ad-hoc communication

Analysis of Cloud Solutions for Asset Management

Goran Kolevski, Marjan Gusev

Institute of Informatics, Faculty of Natural Sciences and Mathematics,
Skopje, Macedonia
{goran.kolevski,marjangusev}@gmail.com

Abstract. One of the key investments in every organization are the assets. Systems for handling data about assets are valuable investment as well. Being able to have an overview of the assets, information about every detail regarding them and moreover being able to analyze that data can be critical for the business. The subject of this paper is to show the importance of the asset management strategy, point the key role that an asset management system has and list out the desired features. Another important aspect of the paper is to show that Software as a Service (SaaS) business model is best suitable model, especially for small and medium sized companies. This paper tries to define a methodology for evaluating asset management systems. Some of the existing solutions are being evaluated against it and compared with each other.

Keywords: asset management systems, asset, management, SaaS, Software as a Service, Cloud computing

1 Introduction

Assets are a key factor as an investment in every organization these days. With few exceptions most of the companies doesn't look at the assets as a valuable investment in their day to day practice. For most of them the concept of asset management is very narrow and use just some of its basics components. Such an example is keeping some list with all items just to pass some external control. This unnatural usage contradicts the original idea. The idea in asset management is to bring some order in the chaos that is living inside the organization and find out the best future steps. The implementation of asset management strategy should imply with benefits on long stages such as cutting down unnecessary expenses whether it's resource or valuable time, maximizing income of currently existing assets and assisting in the development of further investment plans.

Good news on this topic is the increasing number of various size organizations that became aware about the value of the comprehensive system implementation for asset management that will produce complete, precise and dynamic view over their assets but also it will produce indicators for most efficient further development.

One of the most common questions "What is the right time?" stands up here as well. Probably the best answer is "now". Earlier as the organization adopts this

strategy the better it suits all internal processes and avoids any possibility of conflicting with other processes. Having in mind it's critical role in the lifetime of the organization small and medium companies have to adopt it as well. Key point in this discussion is service oriented architecture of asset management system and its SaaS (Software as a Service) nature as the best way to integrate into every company, especially the SaaS in the small and medium sized companies.

With the benefits of contemporary technology organizations have opportunity to organize the assets they own, notice their unnecessary expenses and make most efficient strategy to utilize their current and future investment. Having in mind all of these benefits we put as a strategic goal the implementation of asset management system.

2 Motivation

Organizations with IT asset management strategy are various from multinational corporations to anonymous startups. All those organizations need to put some order in the chaos.

Organizations that implement a good process for efficient asset management reach noticeable savings in their financial reports. According to Gartner its usage can imply with efficient organizing and scheduling of the resources which furthermore can save about 40%-60% of the total IT budget. [3]. Such example show the financial effect of good asset management implementation and that this strategy is critical and deserves attention.

Benefits can be distributed in three general categories like improved efficiency of IT systems, reducing license expenses, reducing of legal and business risks.

Implementing a strategy for asset management means implementing a system for rational spending. Besides implementing a strategy for bookkeeping of asset utilization an additional requirement is executing an analysis over collected data in order to find out the most appropriate strategy for future utilization.

Improved infrastructure for asset management implies with several benefits. As an example most of the spent time with the helpdesk teams is spent on communication regarding the current configuration of the system. Isn't this a subject of asset management? Shouldn't the system administrator be aware of the configuration of the systems in his domain? Another example is easily querying the workstations configuration against wanted parameters. There is no practical limit of the benefits gained with this kind of system. Time after time it's being exploited on various of ways and is generating valuable data.

The implementation of complete system for asset management helps in overcoming of several financial perspective problems. benefit number one is the possibility of generating efficient dispersion of the current assets through the company and discovering the minimum set of assets that has to be bought. And of course this was the fundamental request of a decent asset management system. It is not enough to know what is there in the organization but how it is being used. The key in this kind of system are the analysis of the received data. with applying of

appropriate data analysis tools one can notice the patterns of assets utilization. having in mind such information correct optimizations can be applied on the usage scenarios.

3 Architecture

In the context of asset management solution adding a standalone application might be considered as a paradox. Having in mind that such system would require some additional maintenance overhead for the company. This is the basic reason to consider different solutions that would avoid making the system more complex than it is in the moment but on the other hand solve the problems previously covered.

3.1 Cloud Computing

Cloud computing is probably one of the most popular topics in IT industry these days. Cloud is the next generation computing platform of on-demand information technology services [16]. The general idea with cloud computing is to make a layer of abstraction over overwhelming technological details and bring pure business service for the user. Such step forward would hide the infrastructure like power supplies, data storing devices, processing power, scalability of these systems etc. Cloud computing should shift the computing paradigm to sharing processing power with different applications and expand according to the needs i.e. elasticity, sharing data storage and serving facilities. With such reorganizing of the IT infrastructure smaller business are saved from taking care about problems that aren't their primary business, thus leaving space to concentrate on more important topics for their business. Having this opportunities in mind a lot of vendors are developing and marketing different cloud solutions. There are a lot of benefits with adopting cloud computing both from economical as well as practical perspective [15]:

- Reduced cost: "You don't have to pay for something if you don't use it" is the driving motto of cloud computing. This is saving money on the short run which could be critical for small companies.
- Increased storage: Storing data in a private system could be challenging task if it happens to grow faster than one can handle. Data storage scalability is another aspect of cloud computing that employs latest distributed computing scientific developments.
- Service availability: Storing data in the "cloud" as well as shifting the services and application in the cloud allows one to access them more easier than intranet application.
- Shifting IT personnel focus: With adopting cloud solutions IT personnel is saved from wasting time on maintaining the environment behind the private solution. With such save the IT personnel can focus on more innovative solutions for the business.

Cloud computing also has its own drawbacks. Most obvious problem is the cost of transition and migrating current infrastructure and probably the biggest problem is the security aspect. Privacy issues are large obstacle for offloading data to third party

especially for larger companies. That is why this solution is more suitable for small or maybe medium companies who can't afford dedicated infrastructure but can rent it in order to avoid discussed problems. Having all this arguments in mind the customer is the one who should consider the feasibility of using third party services and make a trade of between the pros and cons [17].

3.2 Service Oriented Architecture and Software as a Service

Service oriented architecture (SOA) is an architecture of software shifted in such way that it employs the Internet as an infrastructure and loosely coupled units of work i.e. services to solve some problem. The service oriented approach delivers task specific reusable software components. This kind of services can be tailored and offered as services to external customers (Software as a Service) or just deliver required results for some internal task. With such architectural change Software as a Service (SaaS) business model will have almost prepared foundation. The only thing left will be tailoring customer oriented services above the already designed infrastructure.

4 Overview of the available solutions

4.1 SAManage

SAManage is an online system for IT Asset management. SAManage offers a detailed overview of the assets in any time from any place via their web application. SaaS (Software as a Service) is the major promoter of this product. The user doesn't have to buy a whole system and make another overhead with it. The system can be used according to the needs and the budget that is available. The one and only requirement is the client application that has to be installed on the computer systems in order to retrieve the required information.

4.2 Express Software Manager

Express Software Manager beside being a non SaaS example is brought into this comparison as a decent representative of such architecture to emphasize the differences between both approaches.

Express Software Manager is a system for managing IT assets such as computers, computer equipment, software tools, licenses and similar helping in making key decisions about the cost of IT assets and also in the process of discovering the potential risks and control what happens within a company network. The product consists of two applications, an administrative console and the client application. The main product is an administrative console through which we can configure the product to manage IT assets. This product has powerful report generating tool, data filtering and searching capabilities, as well as native data populating mechanism.

4.3 Service Now

Service-now.com is so called a pioneer of On Demand IT Service Management. They are combining ITIL v3 guidelines with Web 2.0 technology, and they offer their services with the Software as a Service business model.

The Configuration Management Database (CMDB) is series of tables containing all the assets and business services controlled by a company and their configurations. This includes computers and devices on the network, software contracts and licenses, business services, and more. The IT desk can use the CDMB to understand better their network users' equipment, and the relationships between them. The CMDB can also be referenced by other processes within the system.

The CMDB can be populated using the discovery product. Discovery searches the network for all attached computers and devices, then populates the CMDB with information on each computer/device's configuration, provisioning, and current status. Discovery also reports on any software which is running, and the TCP connections between computer systems, thereby establishing their relationships.

The asset portfolio, asset contracts and configuration applications contain modules which display different tables within the CMDB. Each application is designed with a specific purpose in mind.

The two Asset applications have an Asset Management focus, providing a perspective on the CMDB from a business perspective. The Asset Portfolio application links to CMDB of all assets, hardware, software, assets in stock, as well as records for manufacturers and vendors. The Asset Contracts application contains information about contracts, including leases, service contracts, purchase orders, warranties, and software licenses.[14]

5 Methodology

There are many products that can cover our need or something near that. Some of them are being evaluated in this paper. Different needs and different implementations offer different functionalities that are not comprehensive and equal to our requirements. Sometimes there is a shortage sometimes there are plenty of unneeded features.

In order to make a comparison of the evaluated products in this paper we are going to define some metrics to use in this process. Every indicator has several values that will describe a scale of desirability of that level of implementation of the feature. In order to make some visible comparison every indicator will be assigned a scale of 5 values. The level of implementation will be targeted against those 5 spots according to the desirability of the current implementation. In some of the situations there are not enough implementation options or too many so their values will be scaled appropriately in order to keep the uniformity of the metrics and provide the needed flexibility for new and unpredicted options. In special cases where the feature is not applicable an N/A symbol will be put and it won't be evaluated in the final rankings. Indicators are classified in couple of categories:

5.1 Technical Indicators

Architecture - there are couple of architectural types. The most simple is a local application with these functions, next step is an client-server application with dispersed client application and centralized administration. Next level will be exposing of web services.
1 - local application, 3 - client - server,5 - services

SaaS - Software as a Services is common model for distributing and consuming software services.
1 - Infrastructure provided by the client, 5 - SaaS

Availability over internet - Access to our data in every time from anywhere is a must these days
1 - closed application, 5 - accessible

5.2 Functional Indicators

Reports - One of the most useful features is the ability to generate various reports. Not being able to generate a report and analyze its data discards the idea of asset management.
1 - Doesn't support, 5 - Supports

Data Filtering and Searching - Having in mind we have huge amount of data this requirement is needed in order to find the wanted item(s).
1 - No Search ability, 2 - Characteristics based filters, 5 - Complex filters (almost sql like)

Notifications - in some extreme situations there is need for immediate response.
1 - No notifications, 2 - Notification on the web page, 3 - via email, 4 - SMS, 5 - configurable notifications

Extensibility - Living in a world where everything changes every day we need to be prepared for different kind of scenarios.
1 - No extensibility, 3 - Aggregating available functions, 4 - Programming language support, 5 - Plug-in based expandability with third party applications via services.

Data model abstraction - These kind of applications has to store large spectrum of entities. Being able to easily describe them is wanted feature.
1 - Predefined Classes, 3 - Configuring existing classes, 5 - Defining new classes

Data populating - with the enormous amount of data that is being produced these days automation of the process of population and categorization is required as well.
1 - manual, 3 - automatic population, 5 - automatic population and categorization

6 Comparison of existing solutions

Here is the list with the grades according to the technical requirements.

Table 1. Technical requirements comparison

Technical requirements	SAManage	Express Software Manager	Service-Now
Architecture	5	3	5
SaaS	5	1	5
Availability over internet	5	1	5
Total	5,00	1,67	5,00

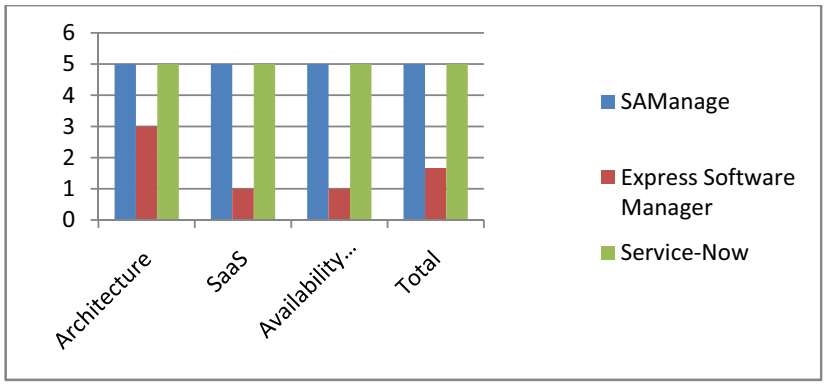


Fig. 1. Technical requirements comparison graph

This is the comparison according to the Functional requirements set.

Table 2. Functional requirements comparison

Functional requirements	SAManage	Express Software Manager	Service-Now
Reports	5	5	5
Data Filtering and Searching	3	3	4
Notifications	1	1	1
Extensibility	1	1	3
Data Model Abstraction	1	1	5
Data Populating	3	5	3
Total	2,33	2,67	3,50

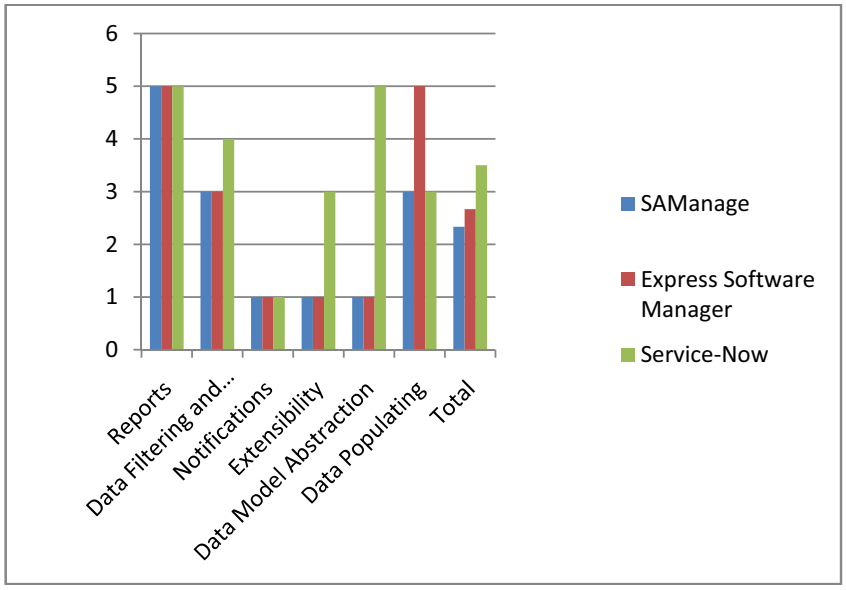


Fig. 2. Functional requirements comparison graph

However some of these features are more important than others. Lets value our desire for some of the features and compare the grades after multiplying them with our desire factor.

Table 3. Factorized functional requirements comparison

Functional requirements	SAManage	Express Software Manager	Service-Now	Factor
Reports	10	10	10	2
Data Filtering and Searching	6	6	8	2
Notifications	1	1	1	1
Extensibility	2	2	6	2
Data Model Abstraction	2	2	10	2
Data Populating	4,5	7,5	4,5	1,5
Total	4,25	4,75	6,58	

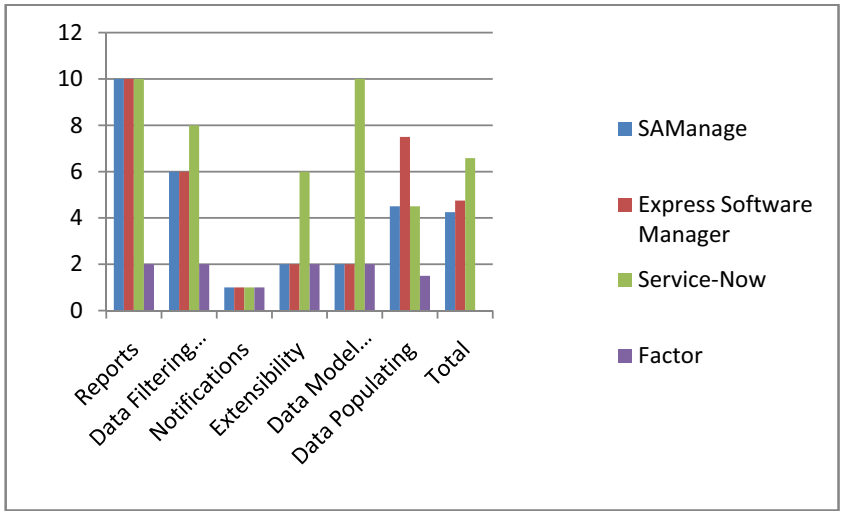


Fig. 3. Factorized functional requirements comparison graph

According to these information the overall summary of the grades is shown in the Table 4 and Figure 4.

Table 4. Overall requirements comparison

	SAManage	Express Software Manager	Service-Now
Technical	5,00	1,67	5,00
Technical - Factor	10,00	3,33	10,00
Functional	2,33	2,67	3,50
Functional - Factor	4,25	4,75	6,58
Total	14,25	8,08	16,58

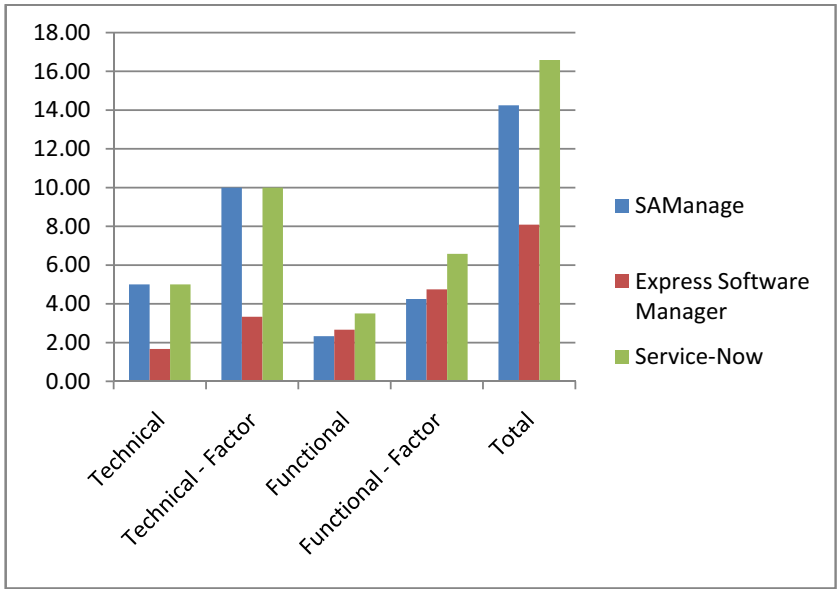


Fig. 4. Factorized functional requirements comparison graph

7 Conclusion

Benefits from asset management strategy appliance are obvious. Also having a good asset management system is one of the most important aspects. Another important aspect was the architecture of the system and the benefits of that choice. Cloud computing and Software as a Service are offered as a solution and placed as an example for appropriate solution and brought as a reference architecture for such system. In this paper we tried to make an overview of desired features and technical preconditions. Also we tried to value our desirability of such functions and value some of the existing software packages against our methodology. General conclusion is that some of the systems managed to gain high scores in some of the sections excelling in some of the features but failed in the rest of the metrics. Examples being chosen are just small subset of all asset management solutions thus another asset management solutions should be analyzed against the offered methodology to gain better knowledge about this field. The metrics list is not comprehensive either and this should be incrementally augmented according to ones needs and newest technology developments.

References

1. Building the business case for it and software asset management - http://www.expressmetrix.com/pdf/roi_6.pdf
2. Microsoft's "implementing sam into your business - http://www.microsoft.com/resources/sam/implementing.mspx
3. Software asset management is on the rise: understanding why - http://www.expressmetrix.com/pdf/itsam_paper.pdf
4. The business case for it asset management - http://success.samanage.com/white-paper---business-case-for-itam/
5. The value of software asset management - http://success.samanage.com/software-asset-management-white-paper
6. Saas vs. Legacy it asset management tools - http://success.samanage.com/saas-vs-legacy-it-management-tools
7. Best practices for implementing software asset management - http://www.expressmetrix.com/pdf/sam_tool.pdf
8. Https://app.samanage.com/contracts
9. Https://app.samanage.com/hardwares
10. Http://www.iso.org/iso/catalogue_detail.htm?csnumber=33908
11. Http://www.expressmetrix.com/products/demo/wbdemo.asp
12. Http://service-now.com/
13. Https://demo.service-now.com/
14. Introduction_to_Assets_and_Configuration: http://wiki.service-now.com/index.php?title=Introduction_to_Assets_and_Configuration
15. R. Buyya, C. S. Yeo, and S. Venugopa, "Market oriented cloud computing: Vision, hype, and reality for delivering it services as computing utilities" In Proceedings of the 10th IEEE International Conference on High Performance Computing and Communications (HPCC-08, IEEE CS Press, Los Alamitos,CA, USA)
16. N. Mirzaei, Cloud Computing,, Pervasive Technology Institute Report, Community Grids Lab, Indiana University
17. Lijun Mei , W. K. Chan , T. H. Tse, A Tale of Clouds: Paradigm Comparisons and Some Thoughts on Research Issues (2008), Proceedings of the 2009 IEEE Asia-Pacific Services Computing Conference (APSCC 2008)

B2B Electronic Service Quality in the Information Systems Discipline: Providing a Research Methodology for Enhanced Scientific Research

Mahmoud Amer [1], Ammar Memari [2], Jorge Marx Gómez [3]

Business Informatics I, Very Large Business Applications
Ammerlaender Heerstr. 114-118, 26129
Carl von Ossietzky University of Oldenburg, Germany
[1] mahmoud.amer@uni-oldenburg.de
[2] memari@wi-ol.de
[3] marx-gomez@wi-ol.de

Abstract. This paper discusses the importance of the electronic service quality measurement in the various domains of Information Systems, Business and Quality Management. The paper proposes a research methodology of studying the electronic service quality in order to provide a comprehensive and rigid research mechanism in this new emerging domain in the Information Systems Field. During our research in the field, we were faced with a problem of the lack of concrete research in the area of Business-to-Business electronic service quality. And the scarcity of methodologies that govern the research in the field. We explore the interdisciplinary nature of the field and propose a research methodology for future research.
Keywords: eCommerce, E-Service, Electronic Service Quality, Business-to-Consumer, Business-to-Business, Service Engineering, Internet.

1 Introduction

The advancement in communications, software and the general knowledge that many customers now possess, along with the emergence of a young generation that is totally dependent on technology, has changed the way many researches are looking at Information Technology. This kind of development made many academics, executives, and entrepreneurs look on a way to enhance the online customer experience by focusing on what kind of metrics or specifications the customer is looking for when visiting a website, trying to learn and categorize these specifications shall lead to enhance customer satisfaction. For the past 50-60 years, the marketing students in the major business schools were taught how to study the human psychology of a customer to make the product more appealing to them. These principles are currently being reformulated as the main medium of selling a product - which used to be a sales person- is being substituted by a website. Of course this

website is a form of software that the customer interacts with, based on a predefined rules programmed in it.

2 What are Services and what is Service Quality?

The simple definition of services can be that, services are deeds, processes, and performances [1]. The services offered by software companies for example are not tangible things that can be touched, seen, or felt, but actually are intangible deeds and performances. These companies offer repair and maintenance services for its equipments, provide IT consulting services for systems and e-commerce applications, web design and hosting, training, and other services. At the end of this process of providing intangible services, we might find some tangible services provided as an end product, for example a final tangible report, a website, or some kind of guides and instruction manuals. If we examine the process of providing service closely, we can see that the service is provided to the customer by some means of problem analysis activities, meetings and interviews with the customer. On the same level, the main offerings of hospitals, hotels, banks, and utilities consist of deeds and actions provided to the end customer. Zeithaml [2] defines the electronic service quality as "*the extent to which a web site facilitates efficient and effective shopping, purchasing and delivery*", which had significant impact on the companies in the service sector. In this sector, the definition usually focuses on meeting customer's needs and requirements, and tries to explain how the service delivered can meet the company's expectations [3].

What do we mean by E-Services?

The topic of E-Services is emerging due to the increased reliance on the internet as a communication medium and as tool in supporting core operations. This utilization can be achieved mainly by two forms. The first is to use information technology as a mean to automate processes in order to reduce human intervention [4]. Something like comparison shopping, and order fulfillment are good examples of that [4]. The other form of utilization can be the actual delivery of service through information technology [5].

The service delivery using electronic mediums is very different from service delivery through traditional channels or "Brick & Mortar" companies, (e.g. Mail order companies, on location retailing and so on). E-Service can be summarized as an interactive information service between the customer and the company through an electronic medium, usually a website. In turn, the service provider gathers and analyzes this information in order to customize future products and other services to be reoffered to the customer [6]. With the emerging advances in the communication fields and the introduction of the internet to support the interaction between the companies and its customers, many of these companies started to provide services to its customers via automated applications. At this level the term eCommerce mainly described the transactions conducted over the internet [6]. Later with the advancement in computer networks and the development in online transactions systems between businesses and suppliers using Electronic Data Interchange (EDI), the term E-

Services started to emerge and evolve to include the new shift in concept by including the customer's experience with the company's interface (i.e.: website), and how he/she interacts with this interface using an online medium. The most suitable definition of E-Services in our analysis is the definition by Hull [7]: "an E-Service is a collection of network-resident software services accessible via standardized protocols, whose functionality can be automatically discovered and integrated into applications or composed to form more complex services." This means that the E-Service is part of a wider concept of Business Service. In summary, E-Service is any service provided on one end via an electronic medium (internet, mobile networks, or interactive TV platform) to a consumer who consumes this service on the other end. Under the above definition we can see for example that withdrawing money from an ATM machine is a form of E-Service as it provides a service using an electronic medium between the customer and a bank.

3 E-Services: Scope and Role

By researching the domain of service quality, we notice that the term is used in multiple disciplines, each of these disciplines adds to the overall value and benefits conceived by the service quality metrics. The main areas that discuss the topic can be classified into three main broader schemes: first is the Information Systems, in this domain the service quality is viewed as a feature of a product -usually a software product- that is harnessed at the end of a process of analyzing, designing and implementing systems. Many of the discussion and papers reviewed in this domain focuses on the service quality as part of an overall software engineering process that tries to deliver a product with a higher level of quality. Figure 1 shows an illustration of the main areas of research around the service quality with some names of the journals and conferences as example in which the term appears in.

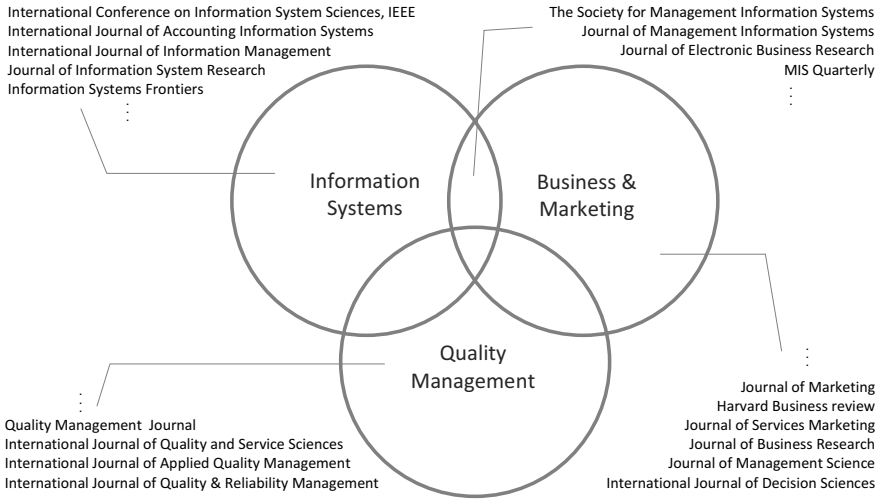


Fig. 1. E-Service as an Interdisciplinary Domain.

The second area is the Business and Marketing domain: in this domain the service quality is viewed from the customer perspective, and part of the consumer buying process which governs his decision in selecting products. The term was first introduced during the 80's and research started to be formulated under a broader theme called Service Marketing [8]. At that time, nearly all the businesses conducted their operations with almost no electronic contact with customer other than telephones and faxes. The concepts were later altered in the 90's to cover the introduction of the internet as part of the consumer buying process scheme. Fig.1 shows the main research journals and conferences that cover the service quality term, as we can see it ranges from purely business related topics, to interdisciplinary subjects between the Business and IS fields. For example Journal of Electronic Business Research and the Journal of Management Information Systems.

The third domain that also discusses the topic can be summarized under the title of Quality Management. Here, many researchers regard the service quality as an overall process that covers an entire production phase [9], making it also a part of the Total Quality Management concept [10], and some researchers also regards it important to enhance the reliability of products.

From the past discussion we can see that when we talk of service quality, it has the same meaning across many disciplines of science and research, but with rather different perspectives depending on the professional and academic background that many researchers possess. But it should not be confused with the term Quality of Service (QoS) that many of the researchers in the computer networks field use. Under that term, we would be discussing the data transfer along the wired or wireless medium in a network [11]. And it is concerned with measuring the technical aspects of the communication medium, rather than measuring the user perception of a service provided by a website or software, in which technical quality is only a part of an overall experience that the user gets by interacting with a company interface, either by website, systems, or employees.

3.1 E-Service Quality as part of the Information System Domain

Some challenges faced when developing E-Services is the need for simultaneous standardization used as cost savings measure and individualization for differentiating the service portfolio. What normally companies do is that they provide a bundle of services with some shared modules, and forming some segments of these modules to provide individual services. But when trying to respond to the ever evolving needs and wants of the customer, and their demand for having multiple service bundled together in a form of "All Inclusive Packages". Companies are faced with a more complex value chain in order to provide these services to the customer. This interaction between the different parts of the value chain using information technologies as mediums to achieve this goal [12]. All this complexity is encompassed into multiple and different systems that work together, and hides this

complexity away from the customer to simplify the service offerings provided to him. Adding to the complexity imposed in the value chain itself is the Information Technology which is easing the way companies enter new markets, and this provides more complexity to the services provided as companies have to deal with different languages, cultural differences and policies.

In the early days of systems development, programmers and computer scientists in charge of developing complex software projects identified a need to regulate the software development process into a structural approach [13]. They embarked on using concepts used to manage complex engineering projects in construction and industry, and benefit from many of the tools and concepts to apply it on the systems they were developing. As these efforts evolved by time, it was grouped and formed under the name of Software Engineering [13]. And hence, providing concrete and structural methodologies in developing software systems with higher specification quality. Various topics, tools, methods and concepts were introduced during the past 40 years since software engineering has been introduced. And one of the latest themes is Service Engineering.

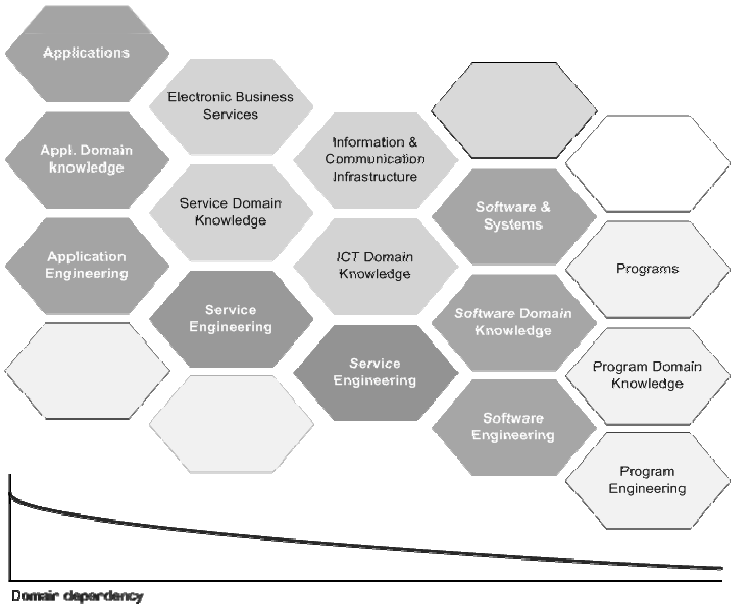


Fig. 2. Service Engineering in the Global Context

Source: Herbert, Weber "From Program Engineering To Software Engineering", TAPSOFT (2005).

Service engineering (SE) is a new approach to the analysis, design and implementation of service-based ecosystems in which organizations and IT provide value to others in the form of services [14]. Service engineering is concerned with

describing a system (i.e. organization) or part of that system from a service perspective to convey the way this system works. This approach transforms the natural language expression of this description to a systematic and structured representation using various models that are understood and accepted. This method reduces the complexity and ambiguity associated with analyzing managing and developing software systems to meet the services needed by the company's customers, departments, or divisions, as many of the organizations see the need for appropriate methodologies and tools to manage and administer their services.

If we talk about the service engineering and its impact on the service sector of the market, we see the need for a greater integration between the services and information technologies, this integration can be reflected by embedding the E-Services in the software development life cycle. This means a combined approach in the effort towards developing systems by combining both software and service engineering in the systems planning, design and implementation phases. This kind of combined integration of software and service aims at developing systems that are efficient, flexible to user's needs and hence effective.

3.2 E-Service Quality as part of the Business and Marketing Domains

The service quality subject has been regarded as the single most researched area in services marketing to date [15], which was due to the increase in the awareness about the importance of service quality in leveraging competitive advantage and long term profitability [16]. And by increasing the level of service quality received by the customer, companies can enhance the overall customer satisfaction perceived by the customers.

Many researchers from the business and marketing fields studied the topic of service quality. One of the first and most recognized studies is the study of Parasurman [17]. This study set out to conceptualize, construct, refine, and test a multiple item scale (E-S-QUAL) for measuring the service quality delivered by websites. The E-S-QUAL was refined, and a subscale of E-S-QUAL called E-Res-QUAL containing items focusing on handling service problems and inquiries was introduced. Other researchers like Collier [18] suggested that customers evaluate the process of placing an order by evaluating the design, information accuracy, privacy, functionality, and ease of the use of a website. Trabold [19] found several interesting differences across the different e-retailing sectors, and mainly focused on analyzing the impacts on the overall E-Service quality dimensions of online retailers, along with several other sectors in the Business-to-Consumer (B2C) environment. These dimensions have resulted in finding differences in service quality among different market sectors.

The study by Fassnacht [20] resulted in a conceptualization with three dimensions, and nine sub dimensions that offer an understanding of quality of electronic service. The three dimensions are environment quality, delivery quality, and outcome quality. The sub dimensions are: graphic quality, clarity of layout, attractiveness of selection, information quality, ease of use, technical quality, functional benefit, and emotional benefit. In this study Jun [21] revealed some important findings about online service

quality. The study identified six key online retailing service quality dimensions as perceived by online customers: reliable prompt responses, access, ease of use, attractiveness, security, and credibility.

In order to deliver and maintain the service quality, an organization must first identify what constitutes quality to those whom it serves [22]. Grönroos [22] classified service quality into two categories: Technical quality, which is primarily focused on what consumers actually received from the service; and Functional quality, focused on the process of service delivery [20].

3.3 Electronic Service Quality as part of the Quality Management Domain

Service quality started to raise some attention during the late 80's and early 90's as some researchers like Lewis [15] who notes in his study about the growing trend and importance of the topic for the domain. This attention was mainly raised as a response to the growing need in the service sector to improve services offered to the customers. This was also influenced by the rapid growth in the service sector and the fierce competition that started to build up during that time. And one of the first papers of Deming [23] that proposed an idea of how statistical thinking can be useful in service organizations and provided a list of ways that these statistical tools can be applied by the service organizations to benefit from them.

During these times, none of the papers or studies proposed dealt nor discussed the intangible portion of the service. As most of the studies were concerned and based on view of applying TQM techniques on a "Service Factory" [24] and [25], this meant that most of the research considered services as tangible form and can be easily measured, controlled, and standardized against variances to the norms. This meant that none of the researchers at that time discussed the challenge imposed by the intangibility of service in their research; it was simply overlooked by many researchers. Some other researchers like [26] and [27] took a more general approach in dealing with total quality. But all efforts were concerned with eliminating inefficient ways and practices used by some organizations that hinders quality, and to increase employee satisfaction. But the main issue of defining the meaning of service quality and how we can improve processes to increase service levels were not discussed at that time.

As the research in the field continued and more concrete and robust studies were conducted to reveal the potential benefits of measuring the service quality, the domain gathered support from the research community and many quantitative tools were introduced. These tools soon were formed into a concise understanding of standards which led to the introduction of ISO 9001:2000 standard under the title of Product and Service Quality Requirements. This standard gained both support and criticism, some critics like Barnes [28] and Naveh [29] argued that applying the standard couldn't achieve higher product or service quality, and that many companies used the certificate for promotional reasons, it also raised some concerns that many products gained the certificate despite providing low quality products.

But all of the criticism about the standards doesn't deny the fact that the market and industry is gaining interest in the matter, but rather that the disagreement on the best

tool to use in order to measure the quality. Currently used tools are being under review, especially when seeing that coupling product and service quality together can be sometime misleading. And that there should be kind of loosely coupled relation between the product quality and service quality, in order for companies to specifically measure the service quality by its own. And to locate the flaws it incurs during its evolution in the service chain.

4 Research Methodology for improving E-Service Quality

Since the early days when Adam Smith proposed the free market economy concept, companies have tried to introduce new business models to gain competitive advantage over their competitors. Whether companies are manufacturing goods, or providing services, there is a service component to every transaction. This means that understanding this component and defining ways to manage, measure, and quantify this component is an essential part of any future development in the service quality field.

Many models discussed in section 3.2 are currently being used, and some are being reviewed for representing changes in the electronic markets that took place in the past 5-10 years, one of them is the increase in online social communities, and hence targeting these social communities by companies is set to provide them with lucrative revenue potential. The other part that is gaining major attention is the Business to Business electronic services that are currently being exchanged between various companies; this exchange of services is formulated in a form of electronic transactions and activities between different parties, starting from the suppliers to manufacturers along to the wholesalers and retailers. Other entities like the financial sector such as banks and insurance companies are also switching its core processes to be integrated with the cyber space. This complex interaction between different parties is governed by each party providing its services to the other in a homogenous manner. A failure in one part means that a bullwhip effect of errors will be echoed along all the value chain and sometimes extending to value systems. And measuring these services in the business to business environment is what we as researchers are interested in.

By looking at Fig. 3 we see that it implies a shift from the traditional Business-to-Consumer (B2C) market which is already covered by research, to the Online Business to Business (B2B) domain that is currently developing and evolving to take its share of attention by the research community. In our efforts to study the Electronic Service Quality (eSQ) in the Business-to-Business (B2B) domain, we were faced with some obstacles of determining how to move from the B2C domain to the B2B one. Our methodology as shown in Fig. 3 can be divided into different phases and steps. The first step is to determine a model of consensus summarizing most important dimension in the various B2C models currently being used. This model will be the first test ground in the B2B domain of the market, which in turn will determine whether the same B2C metrics can be applied to the B2B domain or not. This is done through step 2 by forming a questionnaire representing the B2C model of consensus and testing it in a pilot study consisting of a group of companies from different segments of the market. The results will be represented in step 3 by forming a

preliminary model of B2B electronic service quality model. This model consists of the major metrics that were considered by the companies as important, and these metrics are ranked and grouped according to their weights.

In order to validate the preliminary model, it must be tested on the B2B market (step 4), this test includes companies from different segments of the market, we have concluded that the most feasible way is to divide the market into three segments (Financial sector, Manufacturing sector, and Infomercials which consists of advertising agencies and magazines and newspapers). in this way, we can test whether the model conforms to the needs of different companies in different market segments, And hence minimizing any bias that can develop from one segment of the market influencing the end results. Step 5 indicates that the results of the previous step will clarify the kind of adjustments needed to be done to our preliminary model from step 3, and calibrating the model based on the data results from our market study in step 4. Our final validation test should now take place of the B2B eSQ model, by testing the model on case studies and see how the model conforms to its representations and to have a concrete model that explains the major metrics needed to increase the quality of electronic service provided between businesses.

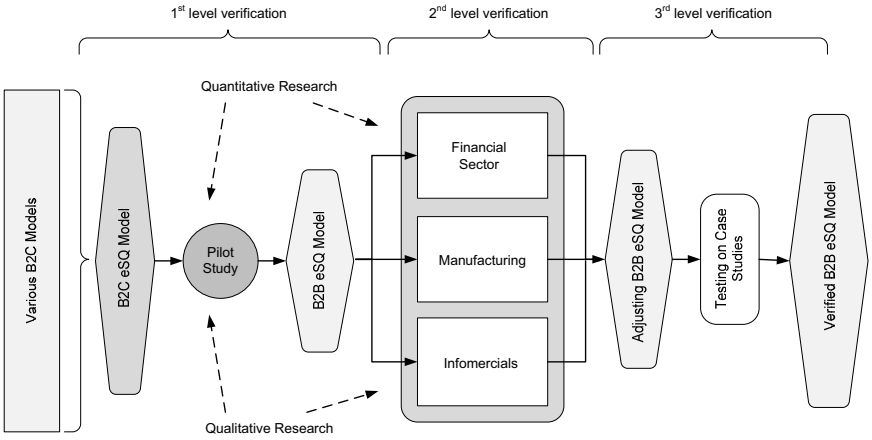


Fig. 3. Research Methodology for Business-to-Business Electronic Service Quality

Source: own preparation

This methodology should be implemented by using both qualitative and quantitative research approaches in step 2 and step 4. By using both questionnaires and interviews to try to see what companies really look for and expect when dealing with a high service quality. Using this multiple approach will not only minimize the bias that can develop during the study, but will also give a chance for the researcher to adjust the

preliminary model for any metrics that might be considered important by the companies which were not listed in that preliminary model. So by going through all the steps mentioned above the final model will have been passed into three levels of verification. The first level is the implied verification of the B2C model by testing it on pilot study and the generation of the preliminary model based on that test. The second level of the verification is testing the model on the market, and the third level of the verification is testing the model on case studies.

6 Summary and Future Research

Everyday many companies are faced with the hard pressing issue of managing and integrating their systems with different partners and organizations. This is usually the way to enhance companies' efficiency in dealing with hard pressures of providing products and services to their customers. This orchestration of activities is choreographed to achieve the best possible output along the value chain. Passing this value at the end of this process to the customers as a form of products and services. By the introduction of electronic devices and software application in the 80's and the integration of these systems with the internet, many companies have leveraged on these benefits, and both customers and companies have demanded more superior services with higher quality. In our paper we have discussed how the topic of electronic service quality relates to the IS discipline, and how this topic has an interdisciplinary nature in the research being conducted. And explaining how different researchers from different backgrounds are researching the topic, each from his own perspective. And during our experience and research in this domain, we have found a lack of a concrete methodology that guides the way researches should attempt researching the matter. With different researchers having their different methods to approach the subject, we have found it of great benefit to provide a methodology that guides future research in the field. A methodology that covers different segments of the market, minimizes bias, and provides rigidly verified models for measuring electronic service quality.

References

1. Zeithaml, Valarie and Bitner, Mary Jo.: Services Marketing. McGraw Hill, New York (1996).
2. Zeithaml, Valarie.: Service excellence in electronic channels. Managing Service Quality, Vol.12 No. 3, pp. 135-139 (2002).
3. Lewis, R. C and Booms, B. H.: The Marketing Aspects of SQ. in Berry L.L, Shostack, G.L., Upah, G.D. (Eds), Emerging Perspectives on Services Marketing, pp.99-104. American Marketing Association, Chicago IL (1983).
4. Singh, Mohini: E-services and their Role in B2C e-commerce. Managing Service Quality, Vol. 12 No. 6, pp. 434-446 (2002).
5. Javalgi, Rajshekhar G., Charles, L. Martin and Patricia, R. Todd: The Export of e-services in the Age of Technology Transformation: Challenges and Implications for International Service Providers. The Journal of Services Marketing, Vol. 18 No. 6/7, pp. 560-73 (2004).

6. Rust, R.T., Kannan, P.: E-service: a new paradigm for business in the electronic environment. Communications of the ACM, Vol. 46 No. 6, pp. 36–42 (2003).
7. Hull, R., Benedikt, M., Christophides, V., Su, J.: E-services: a look behind the curtain. Proceedings of the twenty-second ACM SIGMOD-SIGACTSIGART Symposium on Principles Database Systems, ACM Press, pp. 1–14, New York (2003).
8. Cowell, D.W.: The Marketing of Services. Heinemann, London, (1984).
9. Kotler, P.: Principles of Marketing. Prentice Hall. Englewood Cliffs, NJ, (1980).
10. Asher, J.M.: The Cost of Quality in Service Industries. International Journal of Quality and Reliability Management, Vol. 5 No. 5, pp. 38-46 (1988).
11. Marchese, M.: QoS over Heterogeneous Networks. Wiley, (2007).
12. Tetsuo Tomiyama: Service Engineering to Intensify Service Contents in Product Life Cycles. 2nd International Symposium on Environmentally Conscious Design and Inverse Manufacturing (EcoDesign'01), pp.613 (2001).
13. Ian Sommerville: Software Engineering. Harlow, UK, Addison-Wesley (2008).
14. Herbert, Weber: From Program Engineering To Software Engineering. TAPSOFT (2005).
15. Lewis, B.R.: Quality in the Service Sector: A Review. International Journal of Bank Marketing, Vol. 7 No. 5, pp. 4-12 (1989).
16. Armistead, C.: Quality Assurance in the Uniform Branch of the Police Service. International Journal of Quality and Reliability Management, Vol. 3 No. 3, pp. 8-25 (1986).
17. Parasurman, A., Zeithmal, Valarie and Malhotra, Arvind: E-S-QUAL: a multiple item scale for assessing Electronic Service Quality. Journal of Service Research. Vol. 7 No. 3, pp 213-233 (2005).
18. Collier, Joel and Carol, Bienstock: Measuring Quality in E-Retailing. Journal of Service Research, Vol. 8 No. 3, pp 260-275 (2006).
19. Trabold, Lauren, Heim, Gregory and Field, Joy: Comparing e-services performance across industry sector: Drivers of Overall satisfaction in online retailing. International Journal of Retail Distribution Management, Vol. 34 No. 4/5, pp 240-257 (2006).
20. Fassnacht, Martin and Koese, Ibrahim: Quality of Electronic Services. Journal of Service Research, Vol. 9 No. 1, pp 19-37 (2006).
21. Jun, Minjoon, Yang, Zhilin and DaeSoo, Kim: Customers Perceptions of online retailing service quality and their satisfaction. International Journal of Quality & Reliability Management, Vol. 21 No. 8, pp. 817-840 (2004).
22. Grönroos, C.: A service quality model and its marketing implications. European Journal of Marketing, Vol. 18 No. 4, pp. 37-44 (1984).
23. Deming, W.E.: Out of the Crisis. MIT Press, Cambridge, MA (1986).
24. Maister, D.: Research in Service Operations Management. Proceedings of Workshop on Teaching and Researching Production and Operations Management. London Business School, London (1983).
25. Schmenner, R.W.: How Can Service Business Survive and Prosper?. Sloan Management Review, Vol. 27 No. 3, pp. 21-32, Fall (1986).
26. Berry, L.L, Bennett, D.R and Brown, C.W.: Service Quality: A Profit Strategy for Service Institutions. Dow Jones-Irwin, Homewood, IL (1989).
27. Townsend, P.L. and Gebhart, J.E.: Commit to Quality. John Wiley & Sons, New York, NY (1986).
28. Barnes, Frank.: Good Business Sense Is the Key to Confronting ISO 9000. Review of Business Spring (2000).
29. Naveh. E., Marcus, A.: When does ISO 9000 Quality Assurance standard lead to performance improvement?. IEEE Transactions on Engineering Management, Vol. 51 No. 3, pp. 352–363 (2004).

Comparing Social Bookmarking and Tagging Systems: Towards Semantic Sharing Platforms

Ljupco Jovanoski[1], Vladimir Apostolski[1], Dimitar Trajanov[1]

[1]Faculty of Electrical Engineering and information Technologies, Karpos II bb, 1000 Skopje, Republic of Macedonia
ljupcojovanoski@yahoo.com, vladimir.apostolski@gmail.com, mite@feit.ukim.edu.mk

Abstract. In recent years, collaborative tagging is one of the main ingredient in many sites that provide the user a way to tag bookmarks, photographs and other content. One type of these kind of systems are the well-known social bookmarking sites. Social bookmarking has shown great potential to gather a significant metadata for the web pages it is referring to. The user generated metadata describes the content of the users' bookmarks, despite having its disadvantages, lately is seen as a "goldmine". The use of all this advantages that the social bookmarking uses, results in better social navigation. Application of Semantic web in this area can add a completely new dimension to these applications. In this paper, give an analysis of concepts of social bookmarking and tagging, the most popular social bookmarking systems, including ones that incorporate semantics. We elaborate on their features, advantages and disadvantages. Based on the analysis that we made and our experience in development of application for social semantic bookmarking, we also present the benefits and problems that may arise from the usage of the Semantic web technologies in the field of social bookmarking.

Keywords: social bookmarking, collaborative tagging, semantic web, sharing platforms

1 Introduction

Because of the rapid explosion of resources on the Web, users started to need a way to memorize those resources which are of particular interest to themselves. The process by which Internet users are able to save their web artifacts is called bookmarking. Web browsers offer means for users to bookmark the web resources they consider relevant, and organize them in separate folders, tag them for easy navigation etc. The main problem is that Internet users have become increasingly mobile in the last few years, so their need has grown to the extent that their bookmarks need to be aligned with user's dynamics of movement, i.e. they need to be accessible for the user regardless to whether she accesses it from the same computer or not, or furthermore, regardless to the location of the computer at all. Therefore, different services for storing bookmarks online have been developed. Using descriptive words to mark content of any kind, such as bookmarks, documents etc. is a common way to organize content, so that one can easily retrieve, navigate, filter, search and manipulate it later. The recent popularity gaining by these types of systems has turned focus on itself,

becoming a field of great interest. Even though the concept of assigned keywords is not new, many document repositories and digital libraries provide metadata for their collections of resources. However, the concept of collaboration is not present, as the metadata is provided by an authority, meaning that the custom aspect is not present in this kind of scenarios. In contrast, systems that incorporate collaborative tagging are allowing their users to assign number of keywords, of their choosing. This kind of systems, set the ground for creating semantic web ontologies (formal representation of a set of concepts within a domain and the relationships between those concepts), which are being the highlight of recent researches in this field.

Number of applications on the web use collaborative tagging as their feature. The usual scenario they provide is the following: the user is allowed to publicly tag and share different types of content, so they can not only browse their own content of interest, but share it with other users, and additionally browse through the contents shared by others. The aspect of personalization and social aspect wrapped in one concept called collaborative tagging.

For example, Del.icio.us [1] is a social bookmarking web site allowing collaborative tagging, which has the biggest piece of the pie, meaning that it retains large number of the customers on this market, making it appropriate for research and analysis which will produce generally accepted facts. In following sections, we will analyze the concept of collaborative tagging systems from few points of view. We will concentrate on general user behavior, types of tags provided by the system users, the potential of these types of systems etc.

2 Existing Social Bookmarking Services

In this section, we will review some of the leading sites that incorporate social bookmarking and tagging features: Del.icio.us [1], CiteULike [2], Connotea [3], Diigo [4].

2.1 Delicious

Delicious is a social bookmarking service that allows users to tag, save, manage and share web pages from a centralized source**.** Delicious is the leading application for storing personal bookmarks, adding metadata such as tags, and 2 sharing it publicly with other people. Delicious (formerly del.icio.us), being founded in 2003 by Josua Schachter, today has 5 million users and 150 million bookmarks [5]. Although without an obvious business model, delicious drew attention to major browser extension developers, so that now such addons exist for Firefox, named Foxylicious and Cocoalicious [6] for Mac OS X, accordingly. Delicious's software is written in mod_perl, HTML::Mason, and their data store is MySQL database [7]. Delicious advantages lie in its icebreaker role on this market. Being among the first who contributed to popularization of the social bookmarking concept, Delicious made a suitable ground for becoming great investment by online giants, namely Yahoo. However, Delicious has some weaknesses that competitive applications may take advantage of. First of all, its social side is recessive – meaning interactions between

its members are not very sophisticated. Of course, it should be kept in mind that Delicious was primarily developed and intended to be online bookmark storage, rather than social network with advanced interpersonal features.

Researches that have been conducted, leads to the conclusion that there is a natural convergence of the tags on the Delicious website [8]. The main five cluster hubs of tags revolve around the terms: travel, business, free time, sex, web design. Nonetheless, these tags lack embedded semantics, so when observing Delicious as tagging platform, its folksonomy faces the common ambiguity problems of tagging (such as synonyms or false tags). Mathematical mechanisms have been developed to calculate the interconnection between tags [5] but its implementation might not be feasible in real world high-scaled applications.

2.2 CiteULike

CiteULike is a bookmarking tool for academics. The site includes features such as importing references from desktops and including citations from scientific papers. There are other useful features like finding out who else is reading or has read the same paper. Currently CiteULike is sponsored by Springer [9] [10] and its newest features include one-click extraction of bibliographic references, automated article recommendations etc. CiteULike was developed in October 2004 as a postgraduate research tool by Richard Cameron, but now it is the largest social network for researches available on the Web. This website software was written in Tcl, Common Lisp, Perl, Erlang, and it uses PostgresSQL as its data storage provider [7]. CiteULike's greatest advantage lies in its highly refined targeted group and its partnership with one of the biggest scientific publishers, Springer [11].

2.3 Connotea

Connotea [14], similarly like CiteULike is a web site for sharing references and links, primarily for researchers and clinicians [12]. It also has a good social side for the interaction among the researchers. At first glance, the purpose of Connotea is very similar to the one of CiteULike.

Basically, Connotea supports great variety of file formats for importing, 5, whereas CiteUlike supports only 2 (BibTex and RIS) [13]. On the other hand, Connotea does not support reference lists, compared to CiteULike which supports various formats, such as HTML, RTF, Plain Text, RSS and PDF. Connotea also supports more export file formats rather than CiteULike. There have been discussions over the Internet and lots of comparison lists to what reference management software should be used. However it is worth noting the trend that some social bookmarking applications have emerged into reference management software. Connotea is driven by mod_perl code and stores its data in MySQL database.

2.4 Diigo

Diigo [4] is a new research tool for social bookmarking and knowledge sharing. It has an installable Diigo toolbar which enables users to perform the online bookmarking and annotation of samples of text that one comes across surfing a website. One innovative feature is the ability to highlight certain parts of the hypertext. Diigo also thoroughly covers the aspects of sharing and collaboration in groups. Namely users can create groups for sharing gathered resources upon certain topics. Tagging is also supported for relevant resources, which can be used for topic based search feature. Diigo is also capable of suggesting people with similar interests. One thing that is increasingly annoying for the users, rather than helpful, is the sticky notes feature, which places small sticky notes on every web page the user has opened. Upon hover with the mouse on them, irrelevant information is extracted from diigo and displayed above the sticky note. However, the groups for sharing resources are seemingly sophisticated, with lots of features for notifying members of newly bookmarked resources. Another great capability of Diigo is its integration with the firefox web browser, due to which bookmarking can be performed upon right clicking anywhere on the page. The bookmarking form also contains a snapshot feature, which enables the user to capture a snapshot of the webpage, so other group collaborators can easily see what the intention of the bookmarker was. There is another interesting feature which captures a cached version of the webpage when it was bookmarked, so other group members can clearly see how the web page looked like.

Diigo is one of the richest applications in terms of features, but it's also its greatest weakness. Namely some of the features are annoying, thus having so many features makes less straightforward for usage. It also overnotifies group members about bookmarking activities. Diigo's business model at the time of writing is based on Google Adsense [14].

To summarize, this application bears great potential into becoming an ultimate knowledge sharing and collaboration platform.

3 Collaborative Tagging Potential And Downsides

3.1 Vocabulary problem

The fact that this entire user generated information is publically available, especially tags, leads to the conclusion that the users tagging practice can be influenced, creating closed set of tags over some period of time. However, the fact that users tend to misspell, leads to the biggest problem and 3 disadvantages represented by collaborative tagging systems.

This observation is supported by the probability that two people will provide the same tag for the same content is calculated to be less than 20% [15]. There are two ways of looking at this issues that arise with the vocabulary problem.

The first point of view says that this is not such a problem when it comes to personal bookmarking and sharing with friends. The second point of view states that

when a user wants to take full advantage of collective aspect of social tagging by searching through the shared content, the vocabulary problem is a problem indeed. The fact is that the "selfish" users, which are not thinking about the public aspect that they can involved in or are not aware that they can contribute to form a closed set of tags referring to the shared content, are unlikely to select the same tags as other users.

These patterns of user's tagging practice indicate enhanced findability on the expense of moving towards closed, recognized tag set. These facts also lead towards increased user experience with irrelevant information generated from searches they conduct. The dilemma remains if the del.icio.us users practice social or selfish tagging. Marlow et al. [16] points out that while some people use tags for the purpose of organizing their own bookmarks, others intentionally choose to contribute to the value of "conceptual clusters", or call attention to the pages they bookmark, by adhering to conventions. Golder & Huberman [17] report that over time the relative frequencies of tags applied to a web page stabilize into a pattern such that the most commonly used tags remain so and do not fall out of favor. They elaborate that this is the case because users imitate each other, or because of the nature of the community to tag the same content the same way or because of the relatively limiting content that the web page provides. If the majority of del.icio.us users practice social tagging, than the tag frequencies analysis would not generate the same conclusion[36]. In [15] authors are trying to get facts, instead of making predictions, a research was conducted taking in consideration 349 web pages, including the info provided for them. All of this pages being analyzed were taken from del.icio.us, as one of the biggest players on the market.

Therefore, the main goal of this analysis is to discover if the "vocabulary problem" is really a problem. The results suggest that the del.icio.us users, which are very likely to be similar with all the other users of other social bookmarking sites, tag selfishly. This will only help the users individually, but the problem arises when users of this kind of systems try to use the social side of the social bookmarking services, and then, there will be some disappointing and irrelevant results they'll be getting. Continuing with this tagging practice, instead of benefiting the system by carefully considering the existing set of metadata and providing "the right" tags for their bookmarks, so it can form a open set of metadata, users might end up working against themselves. If the users don't start tagging so they benefit the entire system and its users, so it can form a closed set of metadata, they end up working against themselves.

3.2 How does a tag refer to its content

Now we get to the issue that is the biggest problem in today's social bookmarking services and promises to be of great importance in future development of this kind of systems. When we talk about tagging, we have to understand that tags are a way to organize our content, and additionally add sense to it. The biggest power of the tags provided, lies in the grouping and discovery of similar content.

Let's take a look into the kind of information the tag holds from user perspective. The most common nature of the tags is the following: tags are usually provided in form of nouns that refer to the topics covered in the content being bookmarked i.e. what or who it is about, categorizing content, identify characteristics or assign self

relevance. Users often provide tags related to the type of content being bookmarked e.g. book, blog, gallery, article etc. The perception and the opinion of the user, as well as the emotions provoked by the content, often have a big influence on the tags selection process [17]. It's very likely to run in to tags like funny, scary, interesting, inspirational etc.
One of the worst kind of tags that a user can provide are the self referencing tags (for example tags starting with "my"), emerging by some kind of user's relation with the content. This being the worst tagging practice in today's social bookmarking systems, does not provide any kind of information relative to the user, so he can more easily understand the content being tagged or make an efficient search. A very usual tagging practice indicates that the user bookmarking a content, often connects his tag selection process with a certain tasks he is performing, leading to another personal and selfish way of providing metadata in an unselfish environment.

If we carefully observe the kinds of tags we have mentioned above, we can come to a conclusion that most of them are fully personal, leading towards the concept of a selfish tagging, which doesn't help the social side of things at all.

Anyway, if tags from personal nature do exist i.e. the once only the tagger himself can take benefit from, as shown above, it doesn't mean that they are used by the majority of social bookmarking users, but knowing that a significant number of users tag this way, we can consider this as a downside in current social bookmarking services.

Now, let's take a look on the bright site of this issue. What we are about to state here really, is the potential of the information referring to how the tag describes its content.

Having a clue of how the user interprets his tags for a given bookmark i.e. know the background of the tags is great potential for user activity research, in means of the exploring and determining users practice and habits. But think of the possibilities that arise if the collaborative tagging systems had this kind of information available to use. Imagine the search optimization tools that can be developed based on this information only. The use of all this information about the tags would result in precise and correct results offered by the search engine. This is a new and enhanced way of looking at the concept of tags, and maybe a little over the top, because it's almost impossible to retrieve this type of information from the current format the user provides when tagging his content.

4 Collaboration and resource sharing systems

Because of the specifics of the human nature, the next type of collaboration and resource sharing systems subtly suggest users to prefer tags already used by other users for the same bookmarked resource, leading into more common vocabularies. It is evident that people need better organization rather than relating bookmarked web resources with free text tags. Therefore, every recently developed web application for resource sharing supports autocomplete feature when user is typing the name of the tag she wants to use. The autocomplete usually scans the tags other people used for

the same resource or the whole tag database of the system, channeling tagger's thoughts towards tags already used by other community members. But this is not enough, as tags may have different meanings in different contexts, the folksonomies themselves are not enough for creating knowledge-driven communities.

Braun, Zacharias and Happel [19] point out that among the problems with traditional tagging approach, one can mention: synonyms, multilinguism, polysemy(tags with similar meanings), homonymy(tags with totally different meanings), or mismatch in the level of abstraction(some users prefer narrower, others prefer broader point of view on the same resources). Most of bookmarking/tagging tools described lack the capability to explain what the tag means and how is the tag related to the resource. Semantic Web technologies can aid improving the quality of search for resources through folksonomies, by embedding semantics within each tag.

5 Existing social semantic bookmarking systems

Storing information within tags is time-consuming and requires highly-skilled knowledge engineers, whereas ordinary users refuse to put extra effort in providing metadata in the bookmarking process, without being able to see immediate benefits of it. But despite these obstacles, things began to move to semantically powered communities [20] aiming to migrate from folksonomies to organic ontologies [21]. Among these online systems, that aid information retrieval, knowledge organization, member collaboration and resource sharing, can be mentioned: Faviki [22], Fuzzy [23], Soboleo [24] and GroupMe [25]. This kind of systems are called social semantic bookmarking systems, and the only difference between them and the traditional systems mentioned before, is the key ingredient "semantics".

5.1 Faviki

Not long ago, Faviki [22] was presented like a new social bookmarking service on the web. Faviki simply inherits the concept of social bookmarking, but additionally has a way of retrieving an information describing the tag the user wants to attach to his bookmark. Faviki has a way of suggesting tags to the user when he is about to bookmark his web page of interest. Having the information about the web page the user has intent of bookmarking, Faviki submits its content to Zemanta [26], Wikipedia [27] and Wordnet [28] to identify and retrieve tags [29], which could be of interest to the user. Afterwards, the system presents the tags to the user, who can select some of the suggested tags or he can type his own. However, the one of the issues in the whole concept of Faviki, is the following: when the user types in his tags, it's possible that his input will be rejected. This is pretty annoying for the users, especially when no explanation is provided why they were rejected. The reason that the Faviki tagging tool acts this way, is because Faviki operates over the closed and limited set of concepts provided in Wikipedia. So when a user tries to insert a tag whose definition is not present in the Wikipedia database, the tag is rejected, with no explanation at all. This is the sacrifice that Faviki has to make, in order to form ontology over a closed set of keywords. The good side of Faviki, semantics i.e. every

tag has a meaning attached to it. However, Faviki has a good starting point in the upcoming inevitable "web of meaning".

5.2 Soboleo

Soboleo is a social bookmarking tool that allows concepts that the bookmarks are tagged with to be interlinked and described [19]. It provides AJAX driven bookmarklet for tagging the bookmark and has autocomplete feature. Soboleo offers users ways to edit the ontology itself, allowing them to state that the term "spaghetti" is same as "pasta", drastically improving the semantic search (browsing) through the repository.

The downside of Soboleo is that it requires users to know the concept of ontologies, thus endangers the consistency of the knowledge base by allowing inexperienced users to manipulate the information gathered.

5.3 Fuzzzy

Fuzzy [23] is a bookmark management system that offers users 22 predefined link types for tagging content. Users can also vote for each bookmark, increasing its relevance for that bookmark. In addition, users are allowed to manually choose related tags for a given tag and also to provide wider description about each tag. Fuzzzy introduces the term folktology (as a mesh from folksonomy and ontology)[30]. Fuzzzy uses contextual relevance measurement of semantic distance, similar to the Miller and Charles experiment [31]. Researches over Fuzzy showed that users refuse or avoid annotating content semantically without experiencing immediate results [32], or moreover, they rarely vote for a particular tag as important.

5.4 GroupMe

GroupMe [25] is a semantic application that uses RDF storage to further explain the relationships between tags and the so called groups, which purpose is to aggregate relevant content and share among collaborators on similar topics. GroupMe has an easier approach for management of annotated resources, but all the work of bookmarking and sharing resources has to be done automatically. GroupMe has plenty of space for expanding its semantic features, possibly through semantically powered suggestions about people who might be interested in joining the groups and resources from external sources, such as Freebase [32] or DbPedia [33]. GroupMe is developed by the Semantic Web Group in Hannover, Germany.

5.5 Faveeo

Faveeo.com [37] is a social news aggregator, personalized recommendations system that relies on semantic tagging. This application allows users to import their news stories from digg.com [38] and analyzes their content by their tags. Based on the tags

it attaches, the system recommends relevant content to the reader. It also tracks what stories readers click on, so that they are taken in consideration as well when finding other content for the user. Faveeo also incorporates voting, which leverages Digg API and actually enhances Digg's functionalities with recommendations.

The tagging functionality is powered by OpenCalais [39], a Thomson-Reuters web service that semantically annotates raw text with tags accompanied with their contextual meanings. Faveeo uses the tags which are in context of people, companies and products, and it builds tag clouds for each of these meanings.

In order to retrieve trending topics, Faveeo also tracks twitter updates (referred to as tweets) which mention people, companies or products and displays them alongside with the recommended news stories.

In the time of writing this paper, Faveeo.com is still in beta test and still has to prove its usability level.

5.6 Facebook

Facebook recently turned into social bookmarking website, by presenting its new OpenGraph Protocol [40] Integration. The new "Like" button allows users to save a webpage they like and then retrieve it later from their profile. Moreover, all friends receive notification that a certain person "likes" a given webpage. Being the most popular social network, Facebook goes even further: it tells a visitor which of her friends also like the same webpage prior to bookmarking it, so it subtly suggests that the visitor does the same. The social bookmarking feature is not primarily implemented for advanced browsing through bookmarks, since there is no way to query all of your friends' bookmarks, or query by tag or any convenient way to search through them. However, on individual level Facebook is very precise on what the bookmarks (likes) are about. It now categorizes them on each profile page into categories like: Movies, Sports Teams, Actor, Athlete, University etc with high accuracy. This is because now Facebook allows webmasters to deploy RDFa [41], a semantic web standard to express context of web pages by embedding extra <*meta*> tags containing additional information in the HTML. Having this in consideration, Facebook, now gives webmasters incentives to semantically annotate web pages, since new social relevance ranking algorithms could emerge and secondly, this is a suitable ground for semantic advertising to develop. With these meta-data embedded into webpages, Facebook will build the largest knowledgebase of user interests, individually and in groups. This means that Facebook could wire-up a graph of interests based on the social bookmarking activities of its users.

Although this is the first gigantic company to publicly adopt semantic web technology, it's implementation is flawed in terms of the Semantic Web realization: Facebook does not provide a way to uniquely identify and disambiguate things – if two things have the same names, they are considered the same by Facebook. Second drawback is that only one object can be defined per webpage[42].

However, Facebook breathes new air in the Semantic Web community, providing motivation to further develop social bookmarking applications that are backed up by semantics in the background.

6 Ontology for Sharing Tags Across Platforms: Meaning-Of-A-Tag

It is now evident that tagging platforms are going in direction of embedding semantics and according to one the goals of the Semantic Web, it is supposed to provide means to integrate heterogeneous systems. Tagging activities are going through the same phase. Currently a lot of effort is put to tag resources around the Web: people put tags on blog entries, tweets, video and photo sharing sites, social bookmarking sites etc. But there is no way to reuse these tags across platforms.

In order to overcome this issue, an ontology for describing meaning of tags has been developed [43]. The ontology is called Meaning-Of-A-Tag (MOAT). This ontology provides a framework to describe both local and global meanings of tags towards given resources, tagged by given users. As Passant and Laublet [44][46] denote, this ontology could bridge the gap between the Linked Open Data [45] cloud and the individual systems that contain tags. The MOAT proposes building tag servers: computers that will store tags with their meanings from different systems and provide a SPARQL interface to query tags. Over time, these endpoints are supposed to provide a way to reuse tags created by different platforms, users and meanings [43][46].

7 General impact of integrating semantic web technologies in social bookmarking and tagging environments

Based on the analysis above, a prototype application for social interest sharing has been developed. The semantically powered application would ultimately allow users not only to bookmark resources they find want, but also to propose other bookmarks, people and topics that could be relevant to the users. The application tracks users' bookmarking habits, analyzes the context of the contents and aggregates them into topics with meaning, represented by tags that are used to recommend other content. The development is based on the parole "data will find people, not vice versa anymore". Preliminary results show that on individual level, users' interests can be mapped into a graph, which in turn can be used to propose more sophisticated resource recommendation engines. Because content is tagged with contextual meaning, topic groups can now by themselves choose what content is relevant and should be shared and moreover, who else would be interested to join the group.
The development of this application, has lead us to conclusions that usage of Semantic Web Technologies can have several advantages:

- Solve disambiguates between tags and related problems, such as synonyms and tags in different languages, but with same meaning
- More sophisticated information retrieval; Because the resources are semantically annotated, applications can easily identify relevant content and display it to the users

- Systems that utilize Semantic Web can be interconnected and aligned with lots of other resources described in external ontologies and provide greater value to the users

But usage of Semantic Web technologies has some drawbacks like

- High Cost – applications that utilize semantic technologies inevitably cost more than ordinary applications
- Increased Upfront Investments – comes from the fact that further effort is necessary to formalize the knowledge of given domain
- Increased Delay and Reduced performances – because of the means of searching through the RDF/OWL (Resource Description Framework [34] and Web Ontology Language, respectively)[35] graphs is slower than traditional SQL(Structured Query Language)[36].

7 Conclusion

This paper outlines the social bookmarking services and their shift towards collaboration platforms and semantically fueled applications for interconnecting and sharing metadata. The overview concentrates on the features these systems provide for their users and their shortcomings respectively. Most of the available services offer online storage of bookmarks and resource categorization by tags. The main problems these systems are facing can be summarized as defining vocabulary, context disambiguation, bookmark-tag relation discovery and reusing existing vocabularies among different systems. To overcome these issues, collaboration systems begin to incorporate context-based tags with various approaches for annotation of resources. In general, despite the effort put in developing ontologies, RDF triple-stores, query and rule languages etc., the semantic platforms are facing problems with the unwillingness of people to manually annotate web resources. Possible solutions to these problems range from giving webmasters immediate visible value for embedding metadata or development of web services for automatic semantic annotation through usage of natural language processing algorithms. While neither of them flawless, these approaches lead to the conclusion that global social platforms for knowledge sharing are the future of today's bookmarking and tagging services.

References

1. Delicious Social Bookmarking service, http://www.delicious.com
2. CiteULike Social Bookmarking tool, http://www.citeulike.org
3. Connotea link sharing web site, http://www.connotea.org
4. Diigo Social Bookmarking and Knowledge Sharing, http://www.diigo.com
5. Wikipedia free encyclopedia, http://en.wikipedia.org/wiki/Delicious_(website)
6. Cocoalicious del.icio.us client for Mac OS X, http://www.scifihifi.com/cocoalicious/

7. Tony Hammond, Timo Hannay, Ben Lund and Joanna Scott: Social Bookmarking Tools (I) A General Review (April 2005)
8. Peter Mika: Social Networks and the Semantic Web, Semantic Web and Beyond. (2007)
9. Springer International Publisher Science, Thecnology http://www.springer.com/company/citeulike?SGWID=0-164102-0-0-0
10. Springer International Publisher Science, Thecnology, http://www.springersbm.com/index.php?id=291&backPID=13041&L=0&tx_tnc_news=4739&cHash=56bfa6b56c
11. Springer International Publisher Science, Thecnology, Medicine, http://www.springer.com
12. Connotea link sharing web site, http://www.connotea.org/about
13. Wikipedia free encyclopedia, http://en.wikipedia.org/wiki/Comparison_of_reference_management_software
14. Ad serving application by Google Inc, http://www.google.com/adsense
15. Emilee Rader, Rick Wash: Tagging with del.icio.us: Social or Selfish?, Extended Abstract in Computer Support Cooperative Work (CSCW) Poster Session. (Nov. 2006)
16. Marlow C., Naaman M. and Davis M.: Collaborative Web Tagging Workshop, Proceedings of the 15th International World Wide Web Conference. (May 2006)
17. Scott A. Golder and Bernardo A. Huberman: The Structure of Collaborative Tagging Systems (Aug. 2006)
18. Simone Braun, Valentin Zacharias, Hans-Jörg Happel: Social Semantic Bookmarking, Practical Aspects of Knowlegde Managment 2008. (2008)
19. Simone Braun, Claudiu Schora and Valentin Zacharias: Semantics to the Bookmarks: A Review of Social Semantic Bookmarking Systems, 5th International Conference on Semantic Systems (2009)
20. Ontology (Information science) Wikipedia, the free encyclopedia http://en.wikipedia.org/wiki/Ontology_(information_science)
21. Social bookmarking tool using smart Wikipedia (DbPedia) tags http://www.faviki.com
22. Social bookmarking and networking site for web science academics, web professionals and web enthusiasts http://www.fuzzzy.com
23. SOBOLEO - a web-based system to help users share knowledge http://www.soboleo.com
24. Welcome to Groupme, the Semantic Social Web, http://www.groupme.com
25. Your Content Enhanced – Zemanta Ltd. http://www.zemanta.com
26. Wikipedia, the Free Encyclopedia, http://www.wikipedia.org
27. Wordnet, a Lexical Database in English, http://wordnet.princeton.edu/
28. Vuk Milicic: W3C, Semantic Web Use Cases and Case Studies:Faviki. (May 2009)
29. Roy Lachica and Dino Karabeg: Metadata Creation in Socio-semantic Tagging Systems: Towards Holistic Knowledge Creation and Interchange , Third International Conference on Topic Maps Research and Applications.(Oct. 2007)
30. Sure, Y., Erdmann, M., Angele, J., Staab, S., Studer, R., Wenke, D.: Ontoedit: Collaborative ontology development for the semantic web, In Proceedings of the 1st International Semantic Web Conference. (Jun. 2002)
31. Freebase, http://www.freebase.com
32. DbPedia http://dbpedia.org/About
33. Resource Description Framework , Wikipedia, the Free Encyclopedia http://en.wikipedia.org/wiki/Resource_Description_Framework
34. Web Ontology Language, Wikipedia, the Free Encyclopedia http://en.wikipedia.org/wiki/Web_Ontology_Language
35. Structured Query Language, Wikipedia, the Free Encyclopedia http://en.wikipedia.org/wiki/SQL

36. Furnas, G. W., Landauer, T. K., Gomez, L. M., & Dumais, S. T.: Statistical Semantics: Analysis of the Potential Performance of Key Word Information Systems, Human factors in computer systems. (1984)
37. Faveeo, the Long Tail of the Web, One Person at a Time http://www.faveeo.com/front-temp
38. Digg.com, the Latest News Headlines, Videos and Images http://www.digg.com
39. Home| Opencalais http://www.opencalais.com
40. The Open Graph Protocol, http://opengraphprotocol.org/
41. RDFa, Wikipedia, the Free Encyclopedia http://rdfa.info/wiki/RDFa_Wiki
42. Facebook and the Semantic Web http://www.readwriteweb.com/archives/facebook_the_semantic_web.php
43. Meaning of a Tag | Moat Project http://moat-project.org/
44. Alexandre Passant, Philippe Laublet: Meaning Of A Tag: A Collaborative Approach to Bridge the Gap Between Tagging and Linked Data. (2009)
45. Linked Data | Linked Data http://linkeddata.org
46. Freddy Limpens, Fabien Gandon, Michel Buffa: Linking Folksonomies and Ontologies for Supporting Knowledge Sharing: a State of the Art, Projet ISICIL :Integration Semantique de l'Informationpar des Communautes d'Intelligence en Ligne. (2009)

Increase Learning Success with Game Based Projects

Meltem Yıldırım, Alp Kut

Dokuz Eylül University Computer Engineering Department, Tınaztepe Campus, Buca 35160, İzmir, Turkey
{meltem,alp}@cs.deu.edu.tr

Abstract. Practical studies are significant component in engineering education. Through this importance practical courses must be high quantities rather than the theoretical courses in engineering education process. Application courses can depend on the projects which is prepared as individual or group study by students. Project and game based learning is an effective and powerful method in the learning process. In this study, effect of project and game based methods' with different technique to the course's learning success is examined.

In general each course which project based learning is used may have its own project with related topics. Nevertheless in the same semester these kinds of courses will be causes priorities problems in terms of students. Therefore some courses' success can be decrease. In this sense there is no enough coordination within the courses' instructors. Instead of this classical method, in this approach projects include more than one course's contents are applied. The main idea is to generate specially designed game based projects which include more than one courses contents for easy and interactive learning.

A case-study is proposed as a game based project includes algorithms and programming and digital logic course's topics in this work. Project is produced as a logic game which includes the embedded course's subjects. In this project problem is implementing an interface which enables the user to construct a logical expression.

These projects should also supports incidental learning with game scenarios includes combined several courses contents. This example study is realized with students at first class of computer engineering department, from Dokuz Eylül University.

Grades of experimental students in this study and previous year' grades based on classical learning methods are collected. To explore and survey the learning success of the embedded course, grades of experimental students is compared with the previous year's students' success. This comparison shows that course's success was increased. This study was influenced to increase the learning success of courses.

Keywords: Project based learning, game based learning, learning success

An Expert System for Summer Tourism in Turkey by Using Text Mining and K-Means++ Clustering

Yunus Doğan and Alp Kut

Dokuz Eylul University, Department of Computer Engineering, Tinaztepe Campus, Buca, 35100 Izmir, Turkey
{yunus,alp}@cs.deu.edu.tr

Abstract. This study has an aim to support tourism sector in Turkey by using an expert system; thus, tourists will be able to select the most suitable holiday places for themselves. Before the tourists go to a holiday place which they have not visited before, they make a research about this place. Also, some surprises in this place are learnt before the tourists go and many tourists do not like this situation. Therefore, an operation of text mining is preferred in this study. Thus, tourists do not need a research about the holiday places. All expert system will be returned a decision according to users' preferences. Our expert system has an aim to return more decisions than one. When a tourist uses the system; only one place is not returned, sorted places from the most suitable place to the least suitable one are given. Therefore, a clustering structure is needed. After the system decides the most suitable place for the tourist; the cluster where this suitable place locates finds and the all holiday places in this cluster are recommended in order from the most suitable to the least suitable.

There will be lots of features as attributes from text collection although there will be a low number of holiday places; thus, a large dataset is obtained. Therefore, K-Means clustering algorithm as both simple and fast clustering algorithm is preferred. However, K-Means has problem about deciding the space of clusters, because K-Means can give a different space of clusters with same dataset at each working. The cause of this situation is that K-Means starts clustering with random initial center points. Therefore, K-Means++ clustering is used as a new approach to K-Means without random initial center points and with consistent result spaces.

This study has four steps briefly. Firstly, the most preferable places for summer holiday in Turkey are decided. According to a research on web pages of Cultural and Tourism Ministry of Turkey about tourism, the most important places are Alanya, Ayvalık, Bodrum, Çeşme, Datça, Didim, Dikili, Fethiye, Kaş, Kuşdası, Marmaris, Side and Yalova. These places are preferred by both foreign and regional tourists a lot because of both common and unique features of these places. Therefore, secondly, the features must be determined. For this step, a research with rich documents about these places is done on web and these documents are collected in a text file for each place. These text files will be used for text mining operations in the next steps.

Thirdly, a dictionary is created for each place from the collection of text files. These dictionaries are too large to process, because these dictionaries content stop-words and unnecessary words for tourism. Therefore, some words are

determined to delete from the dictionaries and they are deleted; thus, the satisfactory dictionaries are obtained for each holiday places. A data warehouse must be created from these dictionaries for mining operations. Therefore, preprocess with vector space model is needed; thus, a dataset is obtained with tuples and their attributes. In last step, this dataset is used by K-Means++. It gives a space of clusters where there are the places. Finally, an expert system is ready to use and holiday places are recommended according to these clusters and the expectations of tourists.

Keywords: Expert Systems, Data Mining, Text Mining, Vector Space Model, K-Means++

Windows Filtering Platform, engine for local security

Zoran Spasov, Ana Madevska Bogdanova[1]

[1]Faculty for Natural Sciences and Mathematics, Institute of Informatics
Skopje, Macedonia
zocisp@gmail.com, ana@ii.edu.mk

Abstract. The goal of this paper is to analyze the functionality and usability of the Windows Filtering Platform introduced by Microsoft Corporation as set of API functions that are controlling the network stack of the Windows operating systems. The applications build on it are intended to be used as the windows firewall replacement, with additional functionality in terms of the logic of packet filtering. Some examples demonstrating the usage of the platform will be described. In order to check the durability and exploitability of the systems built on this platform we will use the QualysGuard infrastructure by Qualys, specialized for vulnerability and open ports scanning. The results provided by the tests are analyzed and some conclusions are given.

Keywords: Windows Filtering Platform, QualysGuard, windows security, port scanning, vulnerability testing.

1 Introduction

Windows Filtering Platform (WFP) is a set of API functions and system services that provide a platform for creating applications that control the flow of packets through the networking stack. The code that uses the WFP API, allows developers direct interaction with the packet processing which is done at several layers in the networking stack of the operating system. Network data can be filtered and also modified before the packets reach the destination. WFP offers many application programming interfaces introduced with Windows NT 6.0 that allows applications to tie into the packet processing and filtering pipeline of the new network stack. It provides features such as integrated communication, so the applications coded on top of this interfaces can utilize the powerful processing logic that is based per application or per user bases. It is intended for use by firewalls and other packet-processing or connection monitoring systems.

WFP is a simple development platform, designed to replace previous packet filtering technologies such as Transport Driver Interface (TDI) filters, Network Driver Interface Specification (NDIS) filters, and Winsock Layered Service Providers (LSP). Starting in Windows Vista and Windows Server 2008, the firewall hook and the filter hook drivers are not available; applications that were using previously mentioned drivers should use WFP instead. With the WFP API, developers can implement firewalls, intrusion detection systems, antivirus programs, network monitoring tools,

and some parental controls based on application or user policies. WFP integrates and provides support for firewall features such as installing filters at different stages of the tcp or udp communication, authenticated communication and also there is a dynamic configuration of firewall policy based on applications' use of sockets API (application-based policy). This platform also governs the area of secure communication like controlling and managing the IPSec infrastructure using different policies, raising change notifications, some network diagnostics, and stateful filtering of the connectivity [1].

Windows Filtering Platform as a development platform should not be considered as a standalone firewall engine. The existing firewall application Windows Filtering and Advanced Security (WFAS) built into the newer operating systems, such as: Windows Server 2008, Windows Server 2008 R2, Windows Vista, and Windows 7, is implemented using WFP. So, all these firewall applications utilize the WFP and WFAS features and the powerful processing logic for packet based filtering.

2 WFP: Terms, Definitions and Features

WFP using a set of components acts on the networking stack in order to control the packet flow that goes through the network interfaces on the computer. It is composed from these main components:

- *Filter Engine*. It represents the core multi-layer filtering infrastructure, hosted in both kernel-mode and user-mode, which replaces the multiple filtering modules in the Windows XP and Windows Server 2003 networking subsystem. This engine is responsible for packet filtering based on the firewall policy depending on the data it receives from the shim. Also if there are multiple filter policies that are enforced on a same packet data it provides arbitration between different policy sources and returns the proper action: “permit” or “block”.
- *Base Filtering Engine (BFE)*. It is responsible for controlling the operation of the filtering platform, accepts new filters, creates some statistics and reports the current state of the systems and enforces some configuration on different modules of the system, i.e. the IPSec negotiations.
- *Shims*. These are kernel-mode components that reside between the network stack and the filter engine. They make the filtering decision by classifying against the filter engine. There are many different types of shims: application layer enforcement (ALE), transport layer module, network layer module, etc.
- *Callouts*. They represent a set of special function that are called by a driver and they are used for special filtering, besides the basic “permit” or “block” based on a simple rule. Usually they are used by IPSec processing, adjust stateful filtering behavior, make some intrusion detection processing of packets, etc.
- *Application Programming Interface*. This is a set of data types and functions available to the developers to build and manage network filtering applications.

Windows Filtering Platforms offers many features and extensions to the developers. It also offers connectivity to a third-party filtering systems. Many of the key features that WFP offers are: provides a packet filtering infrastructure where independent software vendors can add some own specialized modules, it works with both IPv4 and IPv6, has a capability for data filtering, modification, and re-injection, performs both packet and stream processing, allows packet filtering to be enabled per application, per user, and per connection in addition to per network interface or per port, provides security during the boot-up process until the firewall engine is transferred to the user environment, enables stateful connection filtering, handles both pre and post IPSec-encrypted data, allows integration of IPSec and firewall filtering policies, provides a policy management infrastructure to determine when specific filters should be activated. This includes mediating conflicting requirements from multiple filters provided by different vendors, has a monitoring and reporting capability of the filtering process and many more [2].

2.1 Windows Filtering Platform architecture

The following picture explains the windows filtering platform architecture in detail:

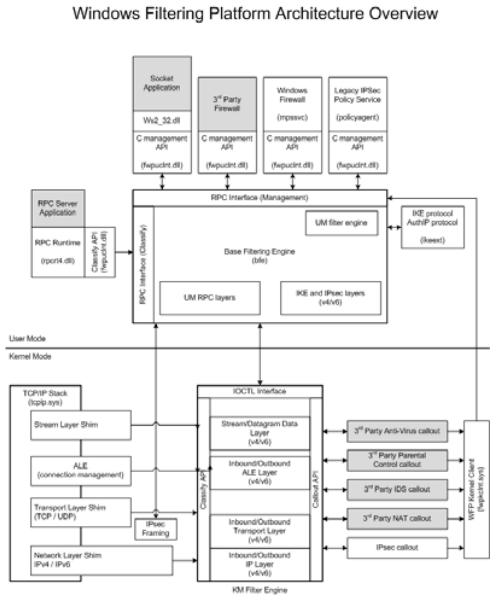


Fig. 1. WFP architecture and communication, courtesy of Microsoft Corp.

The filter engine contains two components that operate in different modes: user-mode and a kernel-mode component, which together perform all of the filtering operations on network packets or streams. The filter engine contains multiple filtering layers for every network layer in the networking stack of the operating system. The

division of the filter engine layers into user-mode layers and kernel-mode layers is based on the filter engine component that owns them.

The user-mode component performs RPC and IPsec filtering. The filter engine contains approximately 10 user-mode filtering layers.

The kernel-mode component performs filtering at the network and transport layers of the TCP/IP stack. This component is also responsible for the calls to the available callout functions during the classification process. The filter engine contains maximum of 50 kernel-mode filtering layers.

The Base Filtering Engine (BFE) belongs to the user-mode component and coordinates the WFP components. The main operation performed by BFE as mentioned earlier is adding and removing filters from the system, storing filter configuration, and enforcing WFP of the security policy. The communication between the application and BFE is done through specific WFP management functions.

Callout drivers provide additional filtering functionality by adding custom callout functions to the filter engine at one or more of the kernel-mode filtering layers. Callouts support deep inspection and packet as well as stream modification. After a callout driver has added its callout functions to the filter engine, filters that specify a given driver's callout function can be added to the filtering process. Such filters can be added by either a user-mode management application or by the callout driver itself.

2.2 Filter Arbitration

When there are multiple filters acting on the same layer there is a need for filter arbitration in order to produce the proper action. Filter arbitration is the logic built into the WFP that is used to determine which filter will have the priority over another in the security policy when making network traffic filtering decisions.

The filter arbitration behavior is done over three rules: no network packet can escape a firewall filter, it passes through all configured filters; packets can be blocked by a Veto from some callout driver even if a filter with higher priority permits it; multiple filters can examine the network traffic at a same layer.

Each filter layer is divided into sub-layers ordered by priority which can be done by attaching weight to a particular layer. Network traffic passes through the sub-layers from the highest priority to the lowest priority. Within each sub-layer, filters are ordered also by weight. Network traffic is indicated to matching filters from highest weight to lowest weight. As we mentioned earlier the filter arbitration algorithm is applied to all sub-layers within a layer and depending on the decision after the packet passes all the filters, a decision is made by this algorithm. This is very suitable if we want to make multiple filters for the same network traffic.

On Fig. 2 is shown a sample diagram of traffic flow through the arbitration algorithm.

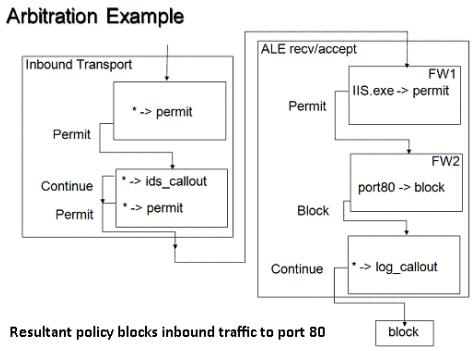


Fig. 2. Filter arbitration example, courtesy of Microsoft Corp.

In this example the firewall policy is intended to block all inbound traffic coming to port 80.

3 Overview of the WFP functionality

Overall the Windows Filtering Platform represents a convenient way to make a security solution based on the operating systems core functions and thus providing the appropriate protection to the users. The fact that this machine directly interacts with the OS networking stack almost on all networking levels eliminates the need for developing generic drivers for the network interface cards that extract the raw network packets.

By introduction of the callout drivers in the WFP architecture opens the possibility for development of a third-party security tools, like intrusion detection systems (IDS), post configuration for secure communication components, network flow monitoring tools, "sniffers" processing on different levels of a network communication, etc. This is also possible due to a good arbitration logic provided by this platform.

Worth mentioning is also the application and the user logic built in the WFP. It provides the means of network traffic filtering based per application or by a specific logged on user. All the filters available through this platform APIs can exist during the boot-up process and after the user profile is up and running they can switch operating in user mode. This transition is atomic and very fast so the local machine is never operating unprotected.

It should be noted that all of this functionality is available in the latest operating systems starting from Windows Vista desktop editions and the Windows Server 2008 family editions. From the developers' point of view, it is important that the API is available only in the C++ programming language [3].

4 QualysGuard solution

In this section we will explore the QualysGuard solution for vulnerability and malware scanning. This is one of the best security scanning solutions on today's market. It searches for open ports on the target machine and tests if its exploitable with all known vulnerabilities and exploits. The scanner will examine the sample application build upon this platform.

The QualysGuard Security and Compliance Suite eliminates network auditing and compliance inefficiencies by leveraging organization's core IT security information. In one consolidated suite, groups with different responsibilities can utilize similar information for their specific needs. The QualysGuard Security and Compliance Suite automates the process of vulnerability management and policy compliance across the enterprise, providing network discovery and mapping, asset prioritization, vulnerability assessment reporting and remediation tracking according to business risk. Policy compliance features allow security managers to audit, enforce and document compliance with internal security policies and external regulations.

QualysGuard vulnerability management enables organizations to effectively manage their vulnerabilities and maintain control over their network security with centralized reports, verified remedies, and full remediation workflow capabilities with trouble tickets. It provides comprehensive reports on vulnerabilities including severity levels, time to fix estimates and impact on business and analysis on security issues.

It has the most up-to-date KnowledgeBase of vulnerability checks in the industry, and the solution comes with external and internal scanners that safely and accurately detect security vulnerabilities across the entire network. As an on demand service, new signatures are delivered weekly, giving users the ability to scan for the latest threats. QualysGuard's is extremely accurate in the scans with almost zero percent false positives, false negatives and host crashes [4].

On Fig. 3 is shown the QualysGuard vulnerability lifecycle.

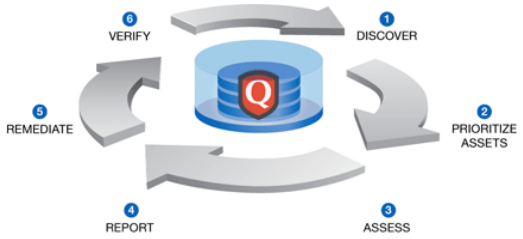


Fig. 3. QualysGuard Vulnerability Management Lifecycle

5 Analysis of WFP applications

In the next section we will present two firewall applications based on the Windows Filtering Platform. The first one is a firewall with two filters at a same sub-layer and the second one will be a demonstration of adding a rule on already started engine. All

the code is written in C++, due to the unavailability in the C# language (.Net framework).

5.1 A firewall application

This firewall application blocks all inbound traffic on the ports less than 10000. The only IP address that is allowed to connect to these ports is the loopback address 127.0.0.1, because it is needed by some processes as a part of the operating system. First we need to register the firewall engine in some user defined dynamic session, as can be seen from the code snippet bellow.

```
FWPM_SESSION0     session;
RtlZeroMemory(&session, sizeof(FWPM_SESSION0));

session.displayData.name = L"FW Session";
session.displayData.description = L"Fltr";
session.flags = FWPM_SESSION_FLAG_DYNAMIC;

error =  FwpmEngineOpen0(NULL,RPC_C_AUTHN_WINNT,
NULL, &session,&engineHandle);
```

After we open the firewall engine for configuration, we must create a sub-layer in which we can add the filters for the network traffic. In our example we must add two filters to correspond to our security requirement mentioned above. The following code represents the filter that allows the loopback address to connect on all the ports.

```
Filter2.subLayerKey=SubLayer.subLayerKey;
Filter2.layerKey= FWPM_LAYER_INBOUND_TRANSPORT_V4;
Filter2.action.type= FWP_ACTION_PERMIT;
Filter2.weight.type=FWP_UINT8;
Filter2.weight.uint8=0x0F;
Filter2.filterCondition= &Condition2;
Filter2.numFilterConditions= 1;

Condition2.fieldKey=
FWPM_CONDITION_IP_LOCAL_ADDRESS;
Condition2.matchType= FWP_MATCH_EQUAL;

Condition2.conditionValue.type=FWP_V4_ADDR_MASK;
Condition2.conditionValue.v4AddrMask =
&AddrAndMask; /// local ip
```

The IP address must be defined as a Hexadecimal number without the dots. Also because we added two filters on the same layer, we must add priority to each filter (weight) in order for the filter arbitration to work properly and the final decision to be correct. From the code above we can also see that we set the firewall engine to process data from the transport layer of the network stack.

5.2 Adding rules

As stated by Microsoft Corp. the rule adding procedure should be very easy and very fast in terms of switching to the new security policy. We will add a rule to our previously mentioned example, which will allow traffic only from one specific address. Bellow is the code snippet.

```
Filter3.subLayerKey=SubLayer.subLayerKey;
Filter3.layerKey= FWPM_LAYER_INBOUND_TRANSPORT_V4;
Filter3.action.type= FWP_ACTION_PERMIT;
Filter3.weight.type=FWP_UINT8;
Filter3.weight.uint8=0x0A;
Filter3.filterCondition= &Condition3;
Filter3.numFilterConditions= 1;

Condition3.fieldKey=
FWPM_CONDITION_IP_REMOTE_ADDRESS;
Condition3.matchType= FWP_MATCH_EQUAL;

Condition3.conditionValue.type=FWP_V4_ADDR_MASK;
Condition3.conditionValue.v4AddrMask =
&AddrAndMask; /// remote ip
```

In order to add the new filter to an existing engine, we need to know the unique sub-layer key. Usually when installing or also known as “pushing” of the new policy, all the policy rules are deleted and rewritten with the new ones. This action should be atomic and done in some sort in a transaction process (in case something goes wrong). WFP API has support for both issues, which was tested and seems to be working, although it need some improvement if one should choose to make serious security solution on top of WFP.

5.3 Testing and analysis

Environment. The example applications were running on Windows 7 operating system, on virtual machine with one gigabyte of RAM and on virtual core 2 duo processor working on 2,20GHz. The native firewall application was turned off.
Tests. Two scans were done with the QualysGuard solution, one with the firewall turned on, and one with the firewall switched off. A third scan was done with a freeware non-commercial application called “NMap” for scanning open ports, in order to have two separate scan results [5]. During the scans the main factor that was observed was the machine performance including CPU usage and memory allocation. The QualysGuard scan profile was set to scan all tcp and udp ports, thus performing a three-way handshake at the process of establishing a tcp connection. This inspection is also known as very intrusive and may cause a denial of service to some services, so it’s recommended to be used with extreme caution. The NMap profile was set to scan all the ports, because it is only a port scanner. It lacks the ability for vulnerability

scans unlike the QualysGuard, but in the firewall application testing scenario, vulnerability scans are not necessary.

Analysis. As mentioned earlier, during the scans, the hardware performance of the machine was monitored, because these tests are known to be very "cruel" to the targets, resulting in their total unresponsiveness. There were CPU usage peaks to maximum of 70% of 10 to 15 seconds period of time. This is not very high, taking into account the scanning profile, particularly the QualysGuard one, and the fact that the firewall was working with virtual hardware.

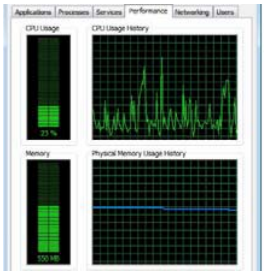


Fig. 4. System performance during the scan process

Memory performance was constant during the whole scanning process. Also all other processes that were running on the target machine did not show any unresponsiveness. Taking into account previously mentioned, the firewall application took the "beating" pretty well. This is one of the major requirements for a firewall application: reliability and availability.

The first scan taken with QualysGuard at the end although showed that the target machine was listening on two tcp ports, it cannot scan for vulnerabilities on these ports.

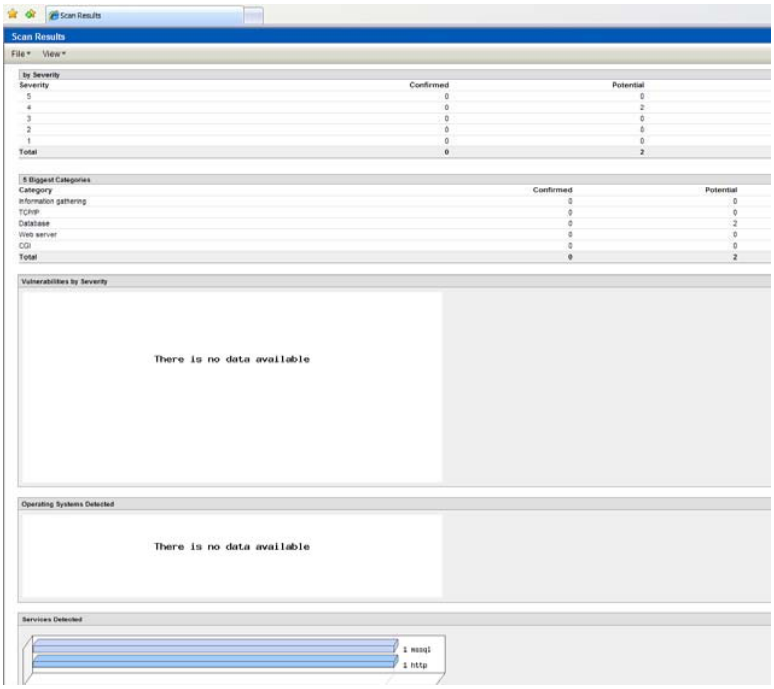


Fig. 5. Scan results from QualysGuard scan

This is mainly because the QualysGuard solution works on probing with TCP RST packets i.e. sends a the first packet for opening a tcp session and goes quite waiting for reset packet from the target machine in order to see if it's listening on that port. Because of the firewall it cannot make a tcp connection and make the necessary vulnerability tests. This port scanning can be block by a callout filter that makes a silent drop at the transport layer using the WFP engine. Blocking of UDP port scanning is done by blocking ICMP destination unreachable messages [6].

The NMap port scanning solution did not find any open ports bellow 10000 when the firewall was turned on.

Adding a rule (deleting rules and installing the new ones) is an atomic procedure, like noted by the platform vendor. The new security policy immediately (bellow one second) was put into operation.

After switching off the firewall the scan taken with the QualysGuard reported a lot of information about open ports and the vulnerabilities on services running on these ports.

6 Conclusion

The Windows Filtering Platform offers solid ground for security software developers in order to quickly produce stable and reliable firewall software. This new technology has proven to have the key requirements needed for serious security solution like: stability, robust engine, easy usage, configurability and option diversibility. Also one of the biggest advantages of this platform is the fact that the platform is available on every new operating system by Microsoft. This means that a piece of software can turn almost any machine in a powerful firewall. Worth mentioning is the support of third-party applications (IDS, VPNs, etc.) through the callout functions.

There are still some disadvantages like: the inability of ARP request processing, some minor bugs in the API functions and the most important one is maybe that until this moment this great platform is not available in the .Net framework.

The platform is in its constant development and even there is announcement that soon will be included in the newest .Net framework. Finally, all these disadvantages are thrown in shadow by the fact that firewall developers will never need to write generic network interface drivers in order to process data for a common application like a firewall.

References

1. Roger A., Grimes, Jesper M. Johansson: Windows Vista Security: Securing Vista Against Malicious Attacks (2007)
2. Microsoft Corporation, online library, http://msdn.microsoft.com
3. Mark E. Russinovich, David A. Solomon with Alex Ionescu: Windows Internals 5th Edition (PRO-Developer) (2009)
4. NMap tool official site, http://nmap.org
5. QualysGuard solution official site, http://www.qualys.com
6. Jan Kanclirz Jr.: How to Cheat at Microsoft Vista Administration (2007)