S. Markovski, M. Gusev (Editors)

ICT Innovations 2012, Web Proceedings

ISSN 1857-7288

S. Markovski, M. Gusev (Editors)

ICT Innovations 2012, Web Proceedings

ISSN 1857-7288

On line edition published on http://ictinnovations.org/2012/

Editors: Smile Markovski, Marjan Gusev
Technical support: Kiril Kirovski, Sasko Ristov
Graphic Design: Innovation, LTD
Illustrator: Alen Aleksovski

Preface

Macedonian Society on Information and Communication Technologies (ICT-ACT) organizes the ICT Innovations conferences since 2009, which is one of its primary activities. In such a way ICT-ACT realizes its mission to support and promote scientific research in the field of informatics and information and communication technologies, as well as their application for building information society.

The fourth ICT Innovations conference "Secure and Intelligent Systems" was held in Ohrid, Republic of Macedonia, on September 12 – 15, 2012. Around hundred presentations and twenty posters were given at the conference from more than 200 authors and coauthors, at the fields of Human Computer Interaction and Artificial Intelligence, Mobile Technologies, Software Engineering, Parallel Processing and High Performance Computing, Computer Networks, Cloud Computing, and Theoretical Computer Science. During the conference five special sessions were organized:

- Quasigroups in Cryptology and Coding Theory
- Bioinformatics
- Multimedia and Presentation of Cultural Heritage
- ICT in Education and E-Learning Platforms
- Improving Quality of Life through Biosignal Processing

as well as two workshops:

- iKnow - Knowledge Management for e-Services in University Management
- German-Macedonian Initiative on Advanced Audio and Speech Signal Processing (GMI-ASP)

There were submitted more than 160 papers of authors from 24 countries. 37 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 53 regular papers and 13 posters presented on the conference. The selection of papers was realized by a program committee of 82 members, 40 of which from the Republic of Macedonia, and the others from 18 countries worldwide. The quality of the conference is based on the hard work done by the reviewers.

Ohrid,
September 2012

Conference chairman
Smile Markovski

S. Markovski, M. Gusev (Editors): ICT Innovations 2012, Web Proceedings, ISSN 1857-7288

Program committee

S. Markovski, M. Gusev (Editors): ICT Innovations 2012, Web Proceedings, ISSN 1857-7288

Organizing committee

S. Markovski, M. Gusev (Editors): ICT Innovations 2012, Web Proceedings, ISSN 1857-7288

S. Markovski, M. Gusev (Editors): ICT Innovations 2012, Web Proceedings, ISSN 1857-7288

Contents

Regular papers

Posters

Workshop

Optimal Resource Scaling for HPC in Windows Azure

Sasko Ristov, Marjan Gusev, Selvir Osmanovic, and Kujtim Rahmani

Ss. Cyril and Methodious University, Faculty of Information Sciences and Computer Engineering,
Rugjer Boshkovikj 16, 1000 Skoipje, Macedonia
sashko.ristov@finki.ukim.mk, marjan.gushev@finki.ukim.mk,
selvir_sk@hotmail.com, kujtim.rahmani@gmail.com

Abstract. Microsoft Windows Azure Cloud offers scalable resources to its customers. The price for renting the resources is linear, i.e. the customer pays exactly double price for double resources. However, not always all offered resources of virtual machine instances are most suitable for the customers. Some problems are memory demanding, others are compute intensive or even cache intensive. The same amount of resources offered by the cloud can be rented and utilized differently to speedup the computation. One way is to use techniques for parallelization on instances with more resources. Other way is to spread the job among several instances of virtual machine with less resources. In this paper we analyze which is the best way to scale the resources to speedup the calculations and obtain best performance for the same amount of money needed to rent those resources in the cloud.

Keywords: Cloud Computing, HPC, Matrix Multiplication

1 Introduction

Cloud computing makes the virtualization and grid computing commercially available. It offers infinite available computing resources (processing power or storage capacity) organized in different virtual machine (VM) instances. Customers can rent as many VMs and use them as long as they need. Scientists can collaborate with each other sharing the data in the cloud [1].

Standard types of instances that are currently offered on the market have linear price / performance ratio. Scaling the resources (CPU, RAM memory) with factor x scales the price with the same factor x. More details about current offers for renting VMs can be found in [9, 3, 2].

However, all problems do not scale well and evenly. Fixed-size speedup (Amdahl's law) bounds speedup to the amount dependent of sequential fraction of the algorithm; the fixed-time speedup is less than scaled speedup bounded to the linear speedup [7]. Also, the performance for particular algorithm does not depend only on CPU and RAM memory. There are a lot of other parameters that impact the performance such as I/O, storage capacity, CPU cache architecture, communication latency runtime environment, platform environment etc. The authors in

[8] discovered several pitfalls resulting in waste of active virtual machines idling. They examine several pitfalls in Windows Azure Cloud during several days of performing the experiments: Instance physical failure, Storage exception, System update. Windows Azure does not work well for tightly-coupled applications [11]. The cloud virtualization generates less cache misses for cache intensive algorithms and thus better performance in both single-tenant and multi-tenant resource allocation for certain workload [5]. Virtualization performance is even better than traditional host for distributed memory, but it provides huge performance drawback in shared memory [4].

The authors in [6] show that dense matrix multiplication algorithm (MMA) runs better on Windows operating system with its C# and threading runtime environment than Linux operating system with OpenMP for parallelization that are hosted in Windows Azure. In this paper we continue the research to analyze the best resource orchestration among the VM instances on the better Windows platform hosted also on Windows Azure. Since the prices of standard VMs grows linearly as resource growth, we determine how to achieve maximum performance for the same price executing dense MMA. We analyze both sequential execution on one core and parallel on maximum 8 cores.

The rest of the paper is organized as follows: Section 2 presents used testing methodology and different hardware environments in Azure Cloud. Section 3 shows the results of the experiments realized to find the performance of the parallelization on particular infrastructure. The final Section 4 is devoted to conclusion and future work.

2 Testing Methodology

This section describes the testing methodology based on 4 different infrastructures with same platform.

2.1 Testing Algorithm

Dense MMA is used as test data. For simplification, we multiply square matrices of same sizes $C_{N.N} = A_{N.N} \cdot B_{N.N}$. Each element c_{ij} of matrix C is calculated as $c_{ij} = \sum_{k=0}^{N-1} a_{ik} * b_{kj}$ where a_{ik}, b_{kj} are correspondingly elements of matrices A, B, for all $i, j, k = 0, \ldots, N-1$.

Matrix elements are stored as double precision numbers with 8 bytes each. One thread multiplies the whole matrices $A_{N.N}$ and $B_{N.N}$ for sequential test cases. For parallel test cases each thread multiplies row matrix $A_{N.N/c}$ and the whole matrix $B_{N.N}$ where $c = 2, 4, 8$ denotes the total number of parallel threads.

2.2 Testing Environments

Testing environment is hosted in Windows Azure. The authors in [10] presents Windows Azure Platform, its components and architecture in details. We use

the same platform environment in each test case with different resource allocation. Windows 2008 Server is used as operating system in each VM instances. Runtime environment consists of C# with .NET framework 4 and threads for parallelization.

The same hardware resources are organized in different Windows Azure VMs:

- 1 x Extra Large VM with total 8 CPU cores;
- 2 x Large VM with total 4 CPU cores per VM;
- 4 x Medium VM with total 2 CPU cores per VM;
- 8 x Small VM with total 1 CPU core per VM.

AMD Opteron 4171 HE processor(s) is used in each VM. It has 6 cores, but maximum 4 of 6 cores are dedicated per VM instance. Each core possesses 64 KB L1 data and instruction caches dedicated per core, 512KB L2 dedicated per core. L3 cache with total 5 MB is shared per chip.

2.3 Test Cases

We realize the experiments in each test case varying matrix size to analyze performance behavior upon different VM resources and variable cache requirements.

Test Case 1: 1 VM with 1 process with 8 (max) threads per process on total 8 cores. In this test case one Windows Azure Extra Large VM is activated allocated with 8 cores as depicted in Figure 1 a). One process in VM executes matrix multiplication with 8 parallel threads. Each thread runs on one core multiplying a row block of matrix $A_{N \cdot N/8}$ and the whole matrix $B_{N \cdot N}$.

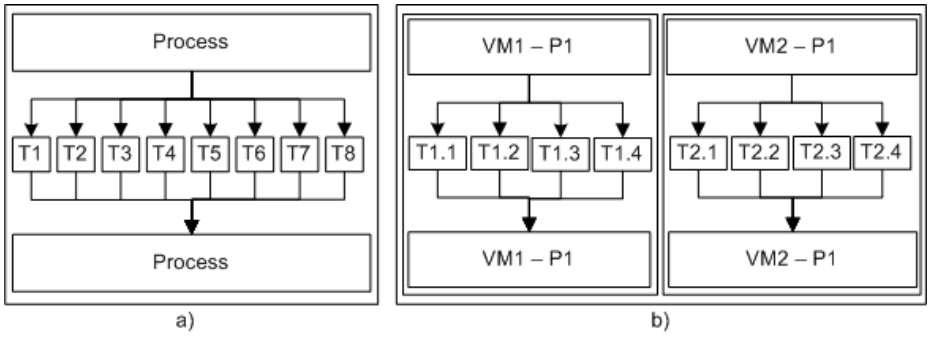


Fig. 1. Test Cases 1 (a) and 2 (b)

Test Case 2: 2 concurrent VMs with 1 process per VM with 4 threads per process on total 8 cores. Two concurrent Windows Azure Large VMs are activated allocated with 4 cores per VM as depicted in Figure 1 b). One process in each VM executes matrix multiplication concurrently with 4 parallel threads

per process (VM). Each process (in separate VM) multiplies the half of matrix $A_{N.N}$ divided horizontally, i.e. a row matrix $A_{N.N/2}$ and matrix $B_{N.N}$. Each thread multiplies a quarter of half matrix A, i.e. $A_{N.N/8}$ and matrix $B_{N.N}$.

Test Case 3: 4 concurrent VMs with 1 process per VM with 2 threads per process on total 8 cores. In this test case four concurrent Windows Azure Medium VMs are activated allocated with 2 cores per VM as depicted in Figure 2 a). One process in each VM executes matrix multiplication concurrently with 2 parallel threads per process (VM). Each process (in separate VM) multiplies the quarter of matrix $A_{N.N/4}$ divided horizontally and matrix $B_{N.N}$. Each thread multiplies a half of quarter of matrix A, i.e. $A_{N.N/8}$ and matrix $B_{N.N}$.

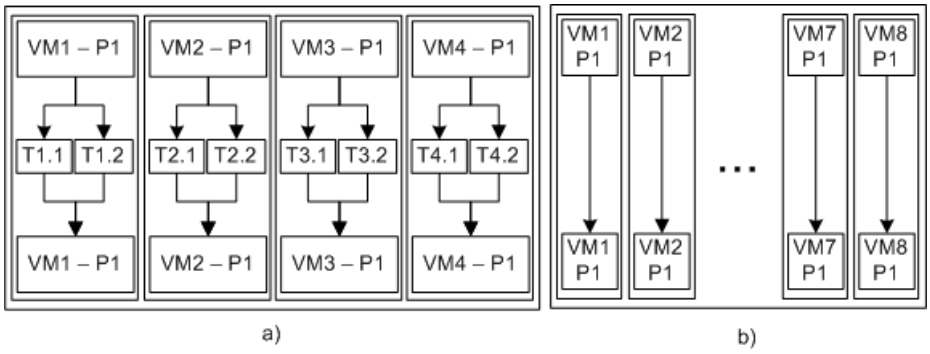


Fig. 2. Test Cases 3 (a) and 4 (b)

Test Case 4: 8 concurrent VMs with 1 process per VM with 1 thread per process on total 8 cores. In this test case eight concurrent Windows Azure Small VMs are activated allocated with 1 core per VM as depicted in Figure 2 b). One process in each VM executes matrix multiplication concurrently with 2 parallel threads per process (VM). Each process (in separate VM) multiplies the quarter of matrix $A_{N.N/4}$ divided horizontally and matrix $B_{N.N}$. Each thread multiplies a half of quarter of matrix $A_{N.N/4}$, i.e. $A_{N.N/8}$ and matrix $B_{N.N}$.

Test Cases 5-8: sequential execution on only one core. Test cases 5-8 execute matrix multiplication sequentially on the testing environments as test cases 1-4 correspondingly. Only one core is used in each of these test cases and all other seven cores are unused and free. The process runs on one core multiplying the whole matrix $A_{N.N}$ with the whole matrix $B_{N.N}$.

2.4 Test Data

Speed V is measured for each test case as defined in (1). Average execution time of all processes per test case is measured to compare the speed of different test cases.

$$V = 2 \cdot N^3 / AverageExecutionTime \quad (1)$$

2.5 Tests Goal

The test experiments has the goal to determine which hardware resource allocation among tenants and threads provides best performance for HPC application in Windows Azure.

Different sets of experiments are performed by varying the matrix size changing the processor workload and cache occupancy in the MMA.

3 The Results of the Experiments

This section presents the results of the experiments that run test cases. We measure the average speed for each test case and analyze their dependencies of different hardware resource allocation, that is, we compare the results of test cases 1 and 5, 2 and 6, 3 and 7, and 4 and 8.

Figure 3 depicts the speed of test cases 1 and 5 for different matrix size N. We determine two main regions with different performance. For $N < 572$ (L_3 region) the whole matrices can be placed in L3 cache and performance are much better than for $N > 572$ (L4 region) where L3 cache misses are generated.

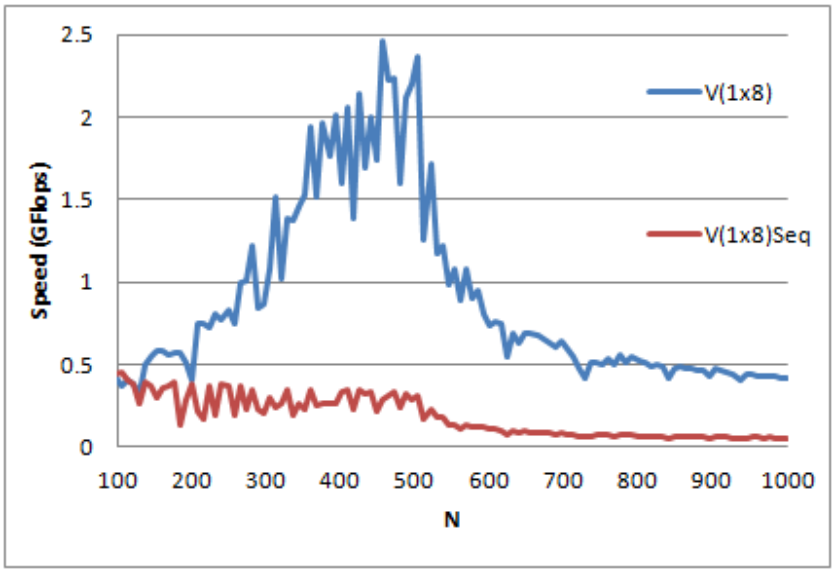


Fig. 3. Speed for test cases 1 and 5

Figure 4 depicts the speed of test cases 2 and 6 for different matrix size N. The same regions with different performance are found also.

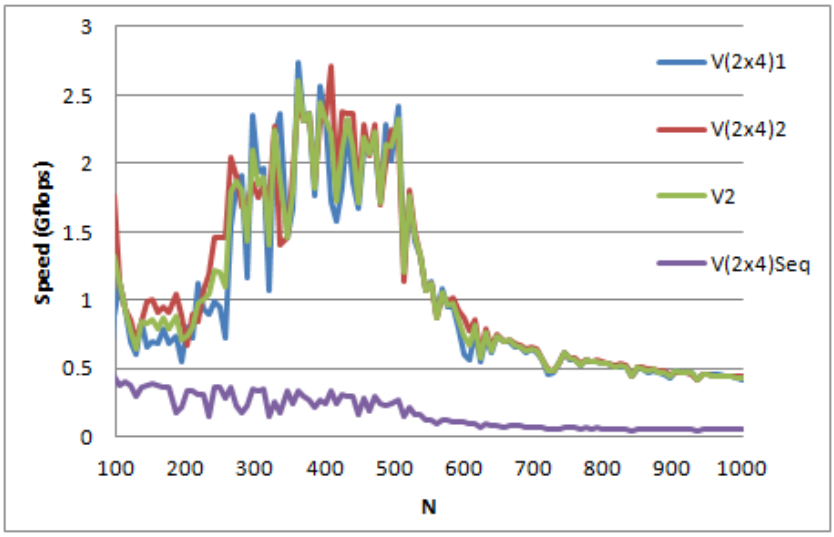


Fig. 4. Speed for test cases 2 and 6

Figure 5 depicts the speed of cases 3 and 7 for different matrix size N. As depicted, there is a huge performance discrepancy in L_3 region among the processes and the average speed.

Figure 6 depicts the speed of test cases 4 and 8 for different matrix size N. We also found a performance discrepancy but more emphasized in L_1 and L_2 regions which are dedicated per process and thread in test case 4 since each process has only one thread, each VM has only 1 core and L1 and L2 caches are dedicated per that core.

4 Conclusion and Future Work

Dense MMA is compute intensive, memory demanding and cache intensive scaled and granular algorithm. Therefore it is optimal algorithm for high performance computing. In this paper we analyze the performance of MMA in Windows Azure Cloud using single-tenant and multi-tenant environments with single-threading and multi-threading hosted on the same hardware resources but differently spreaded among virtual machines.

The results of the experiments for sequential execution are as expected. Extra Large VM achieves maximum speed in front of Large, Medium and Small in L_2 region. MMA algorithm achieves maximum speed when executed parallel on 8

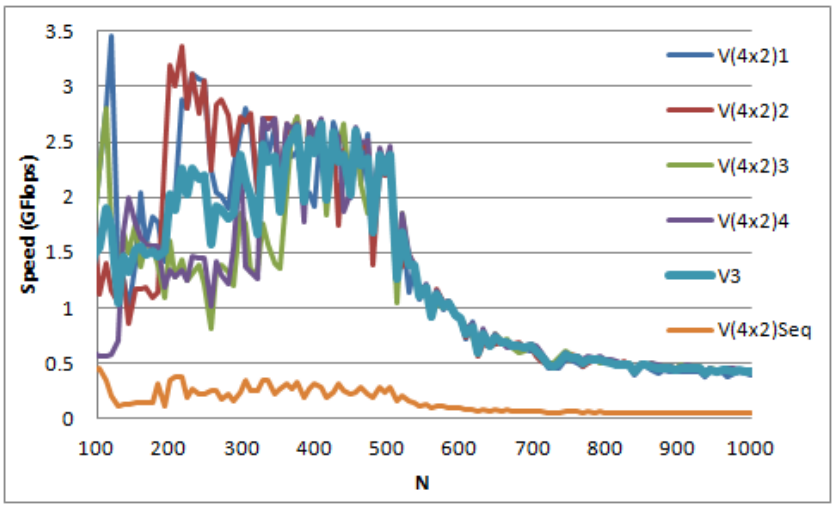


Fig. 5. Speed for test cases 3 and 7

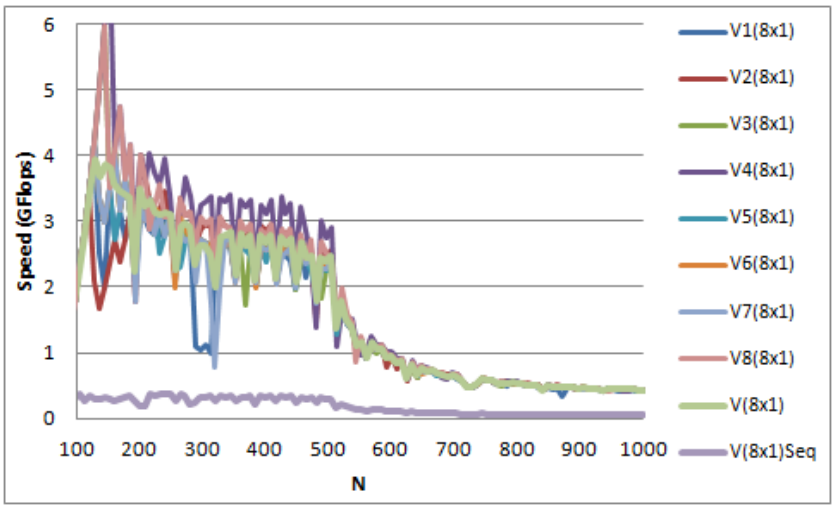


Fig. 6. Speed for test cases 4 and 8

x Small instances, in front of 4 x Medium, 2 x Large, and 1 x Extra Large in L2 and L3 regions, and almost all observed L4 region.

We will continue with research on other hardware architectures and different clouds since MMA depends on CPU and cache memory.

References

1. Ahuja, S., Mani, S.: The state of high performance computing in the cloud. Journal of Emerging Trends in Computing and Information Sciences 3(2), 262–266 (Feb 2012)
2. Amazon: Ec2 (July 2012), http://aws.amazon.com/ec2/
3. Google: Compute engine (July 2012), http://cloud.google.com/pricing/
4. Gusev, M., Ristov, S.: Matrix multiplication performance analysis in virtualized shared memory multiprocessor. In: MIPRO, 2012 Proc. of the 35th International Convention, IEEE Conference Publications. pp. 264–269 (2012)
5. Gusev, M., Ristov, S.: The optimal resource allocation among virtual machines in cloud computing. In: Proc. of 3rd Int. Conf. on Cloud Computing, GRIDs, and Virtualization (CLOUD COMPUTING 2012). pp. 36–42 (2012)
6. Gusev, M., Ristov, S.: Superlinear speedup in windows azure cloud. Tech. Rep. IIT:06-12, University Ss Cyril and Methodius, Skopje, Macedonia, Faculty of Information Sciences and Computer Engineering (July 2012)
7. Gustafson, J., Montry, G., Benner, R.: Development of parallel methods for a 1024-processor hypercube. SIAM Journal on Scientific and Statistical Computing 9(4), 532–533 (July 1988)
8. Lu, W., Jackson, J., Ekanayake, J., Barga, R.S., Araujo, N.: Performing large science experiments on azure: Pitfalls and solutions. In: CloudCom'10. pp. 209–217 (2010)
9. Microsoft: Windows azure (July 2012), http://www.windowsazure.com/pricing/
10. Padhy, R.P., Patra, M.R., Satapathy, S.C.: Windows Azure Paas Cloud: An Overview. International J. of Computer Application 1, 109–123 (Feb 2012)
11. Subramanian, V., Ma, H., Wang, L., Lee, E.J., Chen, P.: Rapid 3d seismic source inversion using windows azure and amazon ec2. In: Proceedings of the 2011 IEEE World Congress on Services. pp. 602–606. SERVICES '11, IEEE Computer Society, Washington, DC, USA (2011)

Nonlinear dynamics of Bose-Einstein condensates by means of symbolic computations

Mihaela Carina Raportaru[1,2] and Alexandru I. Nicolin[1,2,3]

[1] Horia Hulubei National Institute for Physics and Nuclear Engineering, 30 Reactorului St., 077125 Magurele, Romania
[2] Faculty of Physics, University of Bucharest, 405 Atomistilor St., 077125 Magurele, Romania
[3] Faculty of Physics, West University of Timisoara, 4 Parvan St., Timisoara, Romania
alexandru.nicolin@nipne.ro

Abstract. The symbolic computations needed for the variational treatment of high-density Bose-Einstein condensates are described in detail. Two effectively one- and two-dimensional equations based on q-Gaussian functions are derived for the dynamics of cigar- and pancake-shaped condensates. The main result of our symbolic computations is that the variational recipe yields substantially different results for cigar- and pancake-shaped condensates, the variational equations for cigar-shaped condensates being considerable simpler than those for pancake-shaped condensates.

Keywords: Bose-Einstein condensates, q-Gaussian functions, symbolic computations

1 Introduction

The achievement of the first atomic Bose-Einstein condensate (BEC) in 1995 [1] marks the birth of a new research topic which draws from many distinct fields such as atomic and nuclear physics, condensed matter physics, nonlinear and quantum optics and, interestingly enough, even some selected chapters in symbolic and numerical computations. Along with their almost unprecedented experimental maneuverability, BECs are appealing to theoretical physicists due to a very accurate nonlinear partial differential equation, namely the Gross-Pitaevskii equation (GPE), which describes the $T = 0$ nonlinear dynamics of the condensate [2]. The numerical solution of the GPE is well documented by now [3,4,5,6,7,8], but most numerical recipes are time consuming and provide little immediate insight into the dynamics of the condensates, thought there is a long list of numerical and analytical investigations into the dynamics of quantum gases [9,10,11,12].

Variational methods have been extremely attractive in the BEC community because they provide straightforward analytical insight into the properties of the condensates. The most fashionable methods are those tailored around Gaussian functions, which are known to describe the bulk properties of the wave functions

of dilute condensates [13], particularly the frequencies of the collective modes. While Gaussian functions are easy to work with they can not describe accurately the bulk properties of high-density (or, equivalently, highly interacting) condensates where more complex functions have to be considered. Two such functions, namely the S_n [14] and the q-Gaussian [15] functions, have the intriguing property of being able to describe both the low- and the high-density regime. The q-Gaussian functions, in particular, are known to recover analytically the hydrodynamic behavior of high-density condensates and we show here that they can be used to derive effectively one- and two-dimensional equations which describe cigar and pancake-shape condensates [15]. At a more general level, we show in this paper that computer algebra systems are ideally suited for variational investigations into the dynamics of BECs and offer almost immediate answers in regions where traditional paper and pencil calculations are extremely difficult.

2 Non-polynomial Schrödinger Equations

2.1 Cigar-shaped Condensates

The starting point of our investigation is the three-dimensional GPE

$$i\hbar\frac{\partial\psi\left(\mathbf{r},t\right)}{\partial t}=\left[-\frac{\hbar^{2}}{2m}\nabla^{2}+V\left(\mathbf{r}\right)+gN\left|\psi\left(\mathbf{r},t\right)\right|^{2}\right]\psi\left(\mathbf{r},t\right) \tag{1}$$

where $\hbar=h/2\pi$ is the reduced Planck constant, m is the mass of a boson, g measures the strength of the two-body interactions, N is the number of bosons and we consider the trapping potential

$$V\left(\mathbf{r}\right)=\frac{1}{2}m\omega_{\perp}^{2}r^{2}+\frac{1}{2}m\omega_{z}^{2}z^{2}, \tag{2}$$

with $\omega_z \ll \omega_r$, where ω_z (ω_r) represents the longitudinal (radial) frequency of the magnetic trap which confines the condensate. The associated Lagrangian density is given by

$$\begin{aligned}S\left[\psi\left(\mathbf{r},t\right)\right]=\int dtd\mathbf{r}\psi^{*}\left(\mathbf{r},t\right)\left[i\hbar\frac{\partial}{\partial t}+\frac{\hbar^{2}}{2m}\nabla^{2}-V\left(\mathbf{r}\right)\right.\\ \left.-\frac{1}{2}gN\left|\psi\left(\mathbf{r},t\right)\right|^{2}\right]\psi\left(\mathbf{r},t\right),\end{aligned} \tag{3}$$

which we will compute analytically using the trial wave function $\psi\left(\mathbf{r},t\right)=\phi\left(r,t;a,q\right)f$ with

$$\phi\left(r,t;a,q\right)=c\left(1-r^{2}a\left(1-q\right)\right)^{1/(1-q)} \tag{4}$$

and a, q and f are functions of z and t. Imposing that the wave function is normalized to 1, that is $\int dr\left|\psi\right|^{2}=1$, we obtain

$$c = \sqrt{\frac{a\,(3-q)}{\pi}} \tag{5}$$

and after computing the $x-y$ integrals we have that the Lagrangian density (now only with respect to z) is given by

$$\begin{aligned} S\left[f\left(z,t\right)\right] = \int dtdz f^{*}\left(z,t\right) \Bigg[& i\hbar\frac{\partial}{\partial t} + \frac{\hbar}{2m}\frac{\partial^{2}}{\partial z^{2}} \\ & -\frac{gN\left|f\left(z,t\right)\right|^{2}}{2}\frac{a\left(q-3\right)^{2}}{\pi\left(5-q\right)} + \frac{\hbar^{2}}{2m}\frac{2a\left(q-3\right)}{1+q} \\ & -\frac{m\omega_{\perp}^{2}}{2}\frac{1}{2a\left(2-q\right)}\Bigg] f\left(z,t\right). \end{aligned} \tag{6}$$

Despite the unappealing form of the q-Gaussian function (which prohibits any immediate paper-and-pencil calculations) the Lagrangian density can be successfully computed analytically using MATHEMATICA's `Integrate` function. We have tried other computer algebra systems, namely Maple, Matlab and Python (with the SymPy library), and have found that, with the exception of Maple, no other program computed the integrals. We stress that despite their apparent difficulty the $x-y$ integrals yield very simple results (ratios of polynomials in q) which makes it straightforward to compute the *exact* Euler-Lagrange equations associated with $S\left[f\left(z,t\right)\right]$. The only requirement of the symbolic integrator is that $q \in [-1,1]$, because otherwise the q-Gaussian function develops two unphysical branches which make the norm of the wave function (and therefore the number of particles) to diverge. As the physically relevant density regime of the condensate is between $q=-1$ (which corresponds to the Thomas-Fermi limit) and $q=1$ (which corresponds to the low-density limit), the aforementioned divergence is of no physical significance. The ease with which these computations are performed should be contrasted with the extremely restrictive analytical tractability of

$$S_{n} = \exp\left(-\sum_{k=1}^{n}\frac{x^{2k}}{k}\right) \tag{7}$$

for which one can not perform a simple norm such as $I_{n} = \int dx S_{n}^{2}(x)$ for $n \geq 3$, and already I_{2} includes a confluent hypergeometric function.

The equation for $f^{*}(z,t)$ is given by

$$\begin{aligned} i\hbar\frac{\partial f\left(z,t\right)}{\partial t} = \Bigg[& -\frac{\hbar^{2}}{2m}\frac{\partial^{2}}{\partial z^{2}} + gN\left|f\left(z,t\right)\right|^{2}\frac{a\left(q-3\right)^{2}}{\pi\left(5-q\right)} \\ & +\frac{\hbar^{2}}{m}\frac{a\left(3-q\right)}{1+q} + \frac{m\omega_{\perp}^{2}}{4a\left(2-q\right)}\Bigg] f\left(z,t\right), \end{aligned} \tag{8}$$

while the equations for a and q are given by

$$\frac{gN\left|f\left(z,t\right)\right|^{2}}{2\pi}\frac{\left(q-3\right)^{2}}{\left(5-q\right)}+\frac{\hbar^{2}}{m}\frac{\left(3-q\right)}{1+q}-\frac{m\omega_{\perp}^{2}}{4a^{2}\left(2-q\right)}=0 \tag{9}$$

and

$$\begin{aligned}&\frac{gN\left|f\left(z,t\right)\right|^{2}}{2}\left(\frac{2a\left(q-3\right)}{\pi\left(5-q\right)}+\frac{a}{\pi}\frac{\left(q-3\right)^{2}}{\left(q-5\right)^{2}}\right)\\&-\frac{\hbar^{2}}{m}\left(\frac{a\left(3-q\right)}{\left(1+q\right)^{2}}+\frac{a}{1+q}\right)+\frac{m\omega_{\perp}^{2}}{4a\left(2-q\right)^{2}}=0.\end{aligned} \tag{10}$$

The level of difficulty of these equations is similar to that of the equations derived by Salasnich *et al.* [16] for low-density condensates. With specific constraints for the coefficients all computer algebra systems mentioned above are able to provide analytic solutions for equations (9) and (10), but we find it more convenient to restrict the discussion to the high-density regime, that is $N\gg1$, where the approximate solutions are given by

$$q\approx-1+3\left(\frac{2}{a_{s}\left|f\left(z,t\right)\right|^{2}N}\right)^{1/3} \tag{11}$$

and

$$a\approx\frac{m\omega_{\perp}}{8\hbar\sqrt{a_{s}\left|f\left(z,t\right)\right|^{2}N}}. \tag{12}$$

Using the above solutions we have the desired non-polynomial Schrödinger equation

$$\begin{aligned}i\hbar\frac{\partial f\left(z,t\right)}{\partial t}=&\left\{-\frac{\hbar^{2}}{2m}\frac{\partial^{2}}{\partial z^{2}}+2\hbar\omega_{\perp}\left[\sqrt{a_{s}\left|f\left(z,t\right)\right|^{2}N}\right.\right.\\&\left.\left.-\frac{2^{1/3}}{3}\left(a_{s}\left|f\left(z,t\right)\right|^{2}N\right)^{1/6}\right]\right\}f\left(z,t\right)\end{aligned} \tag{13}$$

which was originally introduced in [17] to model the emergence of Faraday waves in high-density cigar-shaped condensates. The main message conveyed by the above computations is that in the quasi one-dimensional case the q-Gaussian radial ansatz allows the *exact* calculation of the Euler-Lagrange equations which *i.*) can either be solved simultaneously by numerical means or *ii.*) one can use the approximate high-density regime solutions of equations (9) and (10) and solve numerically equation (13).

2.2 Pancake-shaped Condensates

For pancake-shaped condensates we apply the same variational treatment using the external potential

$$V(\mathbf{r}) = \frac{1}{2}m\omega_{\perp}^2 r^2 + \frac{1}{2}m\omega_z^2 z^2, \tag{14}$$

now with $\omega_z \gg \omega_r$, and decompose the wave function as $\psi(\mathbf{r},t) = \phi(z,t;w,q)\, f$ where

$$\phi(z,t;w,q) = c\left(1 - \frac{z^2(1-q)}{2w^2}\right)^{1/(1-q)} \tag{15}$$

and w, q and f are functions of r and t. As before, the wave function is normalized to 1, which yields the normalization constant

$$c = \left(\frac{1-q}{2}\right)^{1/4}\left(w\mathbf{B}\left(\frac{1}{2},\frac{q-3}{q-1}\right)\right)^{1/2}. \tag{16}$$

Computing the Lagrangian density using the aforementioned ansatz we find

$$\begin{aligned} S[f(r,t)] = \int dtdr f^*(r,t)\Bigg[& i\hbar\frac{\partial}{\partial t} + \frac{\hbar^2}{2m}\frac{\partial^2}{\partial x^2} + \frac{\hbar^2}{2m}\frac{\partial^2}{\partial y^2} \\ & - \frac{mw^2\omega_{\perp}^2}{7-3q} - \frac{gN|f(r,t)|^2}{2}\cdot\frac{\sqrt{1-q}\mathbf{B}\left(\frac{1}{2},\frac{q-5}{q-1}\right)}{w\sqrt{2}\mathbf{B}\left(\frac{1}{2},\frac{q-3}{q-1}\right)^2} \\ & - \frac{\hbar^2}{m}\frac{U_2\left(\frac{1}{2},2,\frac{3}{2}-\frac{2}{q-1},1\right)}{w^2(3+q)}\Bigg] f(r,t). \end{aligned} \tag{17}$$

Unlike its cigar-shaped sibling, the above Lagrangian density includes both the Euler beta function and the confluent hypergeometric function U [18], which makes it very difficult to work with the exact Euler-Lagrange equation for an arbitrary value of q. One can, of course, derive the exact form of the Euler-Lagrange equations, but the numerous evaluations of the hypergeometric functions preclude simple numerical treatments. Following a detailed analysis of both the Euler beta function and the confluent hypergeometric function U [19], we have found that the Lagrangian density can be well approximated by

$$S[f(r,t)] \approx \int dtdr f^*(r,t)\left[i\hbar\frac{\partial}{\partial t} + \frac{\hbar^2}{2m}\frac{\partial^2}{\partial x^2} + \frac{\hbar^2}{2m}\frac{\partial^2}{\partial y^2}\right.$$

$$-\frac{mw^2\omega_\perp^2}{7-3q}++\frac{\hbar^2}{2m}\frac{1}{w^2}\left(\frac{1}{4}-\frac{3}{2\left(q+1\right)}\right)$$
$$-\frac{gN\left|f\left(r,t\right)\right|^2}{2}\frac{a-b\left(q+1\right)}{w}\Bigg]f\left(r,t\right), \tag{18}$$

where a and b are two numerical constants. Using this approximation one can easily derive the Euler-Lagrange equations which lead to

$$i\hbar\frac{\partial f\left(r,t\right)}{\partial t}=\Bigg[-\frac{\hbar^2}{2m}\frac{\partial^2}{\partial x^2}-\frac{\hbar^2}{2m}\frac{\partial^2}{\partial y^2}+\left(\frac{m}{8}\right)^{1/3}\left(\frac{5\left|f\right|^2 ag\omega_\perp N}{2}\right)^{2/3}$$
$$+\left(3a-40b\right)\left(\frac{g\left|f\right|^2 Nh^3\omega_\perp^4}{2}\right)^{2/9}\frac{m^{1/9}}{4\cdot a^{7/9}3^{1/3}5^{7/9}}\Bigg]f\left(r,t\right).$$

This is equation is not the exact two-dimensional counterpart of (13) due to the approximations used for the Euler beta and the hypergeometric functions.

3 Conclusions

We have investigated by variational means the dynamics of cigar- and pancake-shaped condensates and have derived effectively one- and two-dimensional equations using symbolic computations. Our main result is that the Euler-Lagrange equations obtained for a q-Gaussian ansatz are very sensitive to the geometry of the condensate. Cigar-shaped condensates, in particular, can be well described using relatively simple variational equations, while pancake-shaped condensates are described by more complex equations which involve the Euler beta function and the confluent hypergeometric function.

Acknowledgements For this work M.C.R. was supported by the Ministry of Education and Research under PN 09370104/2012 and by CNCS–UEFISCDI through the project PN-II-ID-PCE-2011-3-0323, while A.I.N. was supported by CNCS–UEFISCDI through the project PN-II-ID-PCE-2011-3-0972.

References

1. Pethick, C.J., Smith, H.: Bose-Einstein Condensation in Dilute Gases. Cambridge University Press, Cambridge (2008)
2. Kevrekidis, P.G., Frantzeskakis, D.J., Carretero-González, R. (eds.): Emergent Nonlinear Phenomena in Bose-Einstein Condensates, Springer, New York (2008)
3. Adhikari, S.K.: Numerical study of the spherically symmetric Gross-Pitaevskii equation in two space dimensions. Phys. Rev. E 62, 2937 (2000)
4. Adhikari, S.K.: Numerical solution of the two-dimensional Gross–Pitaevskii equation for trapped interacting atoms. Phys. Lett. A 265, 91 (2000)

5. Adhikari, S.K., Muruganandam, P.: Bose-Einstein condensation dynamics from the numerical solution of the Gross-Pitaevskii equation. J. Phys. B: At. Mol. Opt. 35, 2831 (2002)
6. Adhikari, S.K., Muruganandam, P.: Bose–Einstein condensation dynamics in three dimensions by the pseudospectral and finite-difference methods. J. Phys. B: At. Mol. Opt. 36, 2501 (2003)
7. Muruganandam, P., Adhikari, S.K.: Fortran programs for the time-dependent Gross–Pitaevskii equation in a fully anisotropic trap. Comp. Phys. Comm. 180, 1888 (2009)
8. Vudragović, D. , Vidanović, I., Balaž, A., Muruganandam, P., Adhikari, S.K.: C programs for solving the time-dependent Gross–Pitaevskii equation in a fully anisotropic trap. Comp. Phys. Comm. 183, 2021 (2012)
9. Balaž, A., Nicolin, A.I., Faraday waves in binary nonmiscible Bose-Einstein condensates. Phys. Rev. A 85, 023613 (2012)
10. Vidanović, I., Balaž, A., Al-Jibbouri, H., Pelster, A.: Nonlinear Bose-Einstein-condensate dynamics induced by a harmonic modulation of the s-wave scattering length. Phys. Rev. A 84, 013618 (2011)
11. Balaž, A., Vidanović, I., Bogojević, A., Belić, A., Pelster, A.: Fast converging path integrals for time-dependent potentials: I. Recursive calculation of short-time expansion of the propagator. J. Stat. Mech. P03004 (2011)
12. Balaž, A., Vidanović, I., Bogojević, A., Belić, A., Pelster, A.: Fast converging path integrals for time-dependent potentials: II. Generalization to many-body systems and real-time formalism. J. Stat. Mech. P03005 (2011)
13. Pérez-García, V.M., Michinel, H., Cirac, J.I., Lewenstein, M., Zoller, P.: Low Energy Excitations of a Bose-Einstein Condensate: A Time-Dependent Variational Analysis. Phys. Rev. Lett. 77, 5320 (1996)
14. Keçeli, M., Ilday, F.Ö., Oktel, M.Ö. : Ansatz from nonlinear optics applied to trapped Bose-Einstein condensates. Phys. Rev. A 75, 035601 (2007)
15. Nicolin, A.I., Carretero-González, R. : Nonlinear dynamics of Bose-condensed gases by means of a q-Gaussian variational approach. Physica A 387, 6032 (2008)
16. Salasnich, L., Parola, A., Reatto, L.: Effective wave equations for the dynamics of cigar-shaped and disk-shaped Bose condensates. Phys. Rev. A 65 043614 (2002)
17. Nicolin, A.I., Raportaru, M.C.: Faraday waves in high-density cigar-shaped Bose-Einstein condensates. Physica A 389, 4663 (2010)
18. Wolfram, S.: The Mathematica book. Cambridge University Press, Cambridge (1999)
19. Nicolin, A.I. : Effective wave equation for the dynamics of high-density disk-shaped Bose-Einstein condensates. Rom. Rep. Phys. 61, 641 (2009)

Increasing students' motivation by using social networks in and out of the classrooms

Vladimir Trajkovic[1], Marina Vasileva[2], Olga Samardzic Jankova[3], Slavica Karbeva[4] and Maja Videnovic[5]

[1]Faculty of Computer Science and Engineering, Skopje
[2]Primary School Sv. Kiril i Metodij, Skopje
[3]Civic Accossiation Open the Window, Skopje
[4]Primary School Petre Pop Arsov, v. Bogomila, Municipality Chashka
[5]Primary School Krste Misirkov, Skopje

Abstract. Putting his Majesty computer into schools made tectonic shift of the central pillar in every classroom - the teacher. The threat became bigger with the use of social networks like Facebook and Twitter by the students, as a result of that these sites became forbidden in the schools.

But when the students are already there, available in any time and place, why can't we try to direct these users for the benefits of innovative learning and to improve their motivation?

The following text will try to investigate the possible answers of this question concerning the benefits from using Facebook and twitter in education as a media for communication between students and teacher.

After reviewing the subject of research from the aspect of planned distance education or distance learning which will continue and after the end of the school year, desired data will be obtained. Using the WebQuest method will provide approach to learning from the source of the information distant from the students. Qualitative and quantitative indicators will be presented by the measurement of the achievements and by the students and teacher attitudes.

Keywords: social networks, Facebook, twitter, WebQuest method.

1 Introduction

Putting his Majesty computer into schools made tectonic shift of the central pillar in every classroom - the teacher. The threat became bigger with the use of social networks like Facebook and Twitter by the students, as a result of that these sites became forbidden in the schools.

But when the students are already there, available in any time and place, why can't we try to direct these users for the benefits of innovative learning and to improve their motivation?

We can try to do the opposite - Threats to translate into Opportunities. Than we can seek benefits from using social networks: Facebook and Twitter, in and out of the

classrooms as a media for communication between students and teacher. The course will use the WebQuest method and available Open Educational Resources (OER).

1.1 Facebook

Facebook is the world's largest social network, with more than 900 million users as of May 2012, most of them using the mobile devices. This web service is free to join and open to anyone over 13. It provides a place for social connection via the sharing of photos, videos and text updates. Users create personal profiles and establish relationships with other people and companies.

1.2 Twitter

Twitter is an online social networking service and microblogging service that enables its users to send and read text-based messages of up to 140 characters, known as "tweets".

It was created in March 2006 by Jack Dorsey and launched that July. The service rapidly gained worldwide popularity, with over 500 million active users as of 2012, generating over 340 million tweets daily and handling over 1.6 billion search queries per day. Since its launch, Twitter has become one of the top 10 most visited websites on the Internet, and has been described as "the SMS of the Internet." Unregistered users can read tweets, while registered users can post tweets through the website interface, SMS, or a range of apps for mobile devices.

1.3 Web Quest .

A WebQuest is an inquiry-oriented lesson format in which most or all the information that learners work with comes from the web. The model was developed by Bernie Dodge at San Diego State University in February, 1995. WebQuests can be created using various programs, including a simple word processing document that includes links to websites. A WebQuest is distinguished from other Internet-based research by three characteristics. First, it is classroom-based. Second, it emphasizes higher-order thinking (such as analysis, creativity, or criticism) rather than just acquiring information. And third, the teacher preselects the sources, emphasizing information use rather than information gathering. Finally, most WebQuests are group work with the task frequently being split into roles.

A WebQuest has 6 essential parts: introduction, task, process, resources, evaluation, and conclusion.

1. The introduction provides background information and gives meaning to the exercise. The introduction should have a motivational component to excite the students.
2. The task is the formal description of what the students will produce in the WebQuest. The task should be beautiful, meaningful, and fun. Creating the task is the most difficult and creative part of developing a WebQuest.

3. The process is consisted of precisely defined steps the students should take to accomplish the task. It is frequently profitable to reinforce the written process with some demonstrations.
4. The resources the students should use. Providing these helps focus the exercise on processing information rather than just locating it. Though the instructor may search for the online resources as a separate step, it is good to incorporate them as links within the process section where they will be needed rather than just including them as a long list elsewhere. Having off-line resources like visiting lecturers and sculptures can contribute greatly to the interest of the students.
5. The evaluation is the way in which the students' performance will be evaluated. The standards should be fair, clear, consistent, and specific to the tasks set.
6. The conclusion part should be used for reflection and discussion of possible extensions.

1.4 Open educational resources

Emphasizing that the term Open Educational Resources (OER) was coined at UNESCO's 2002 Forum on Open Courseware and designates "teaching, learning and research materials in any medium, digital or otherwise, that reside in the public domain or have been released under an open license that permits no-cost access, use, adaptation and redistribution by others with no or limited restrictions. Open licensing is built within the existing framework of intellectual property rights as defined by relevant international conventions and respects the authorship of the work". Users of Open Educational Resources have free (no-cost) access to the materials and free (no-cost) permission to engage in the "4R" activities when using them, including:

Revise: adapt and improve the Open Educational Resources so it better meets your needs;

Reuse: use the original or your new version of the Open Educational Resource in a wide range of contexts;

Remix: combine or "mashup" the Open Educational Resource with other OER to produce new materials;

Redistribute: make copies and share the original Open Educational Resource or your new version with others.

2 Research

2.1 Goal of the research

In the course of this research we tried to investigate are the students more motivated to learn through the use of social networks in and out of the classroom.

Hereby, the suitability shall be examined through the following research questions:

1. Does the communication between teacher and students, using the social networks, increase the students' motivation for learning?

2. Does the use of the WebQuest method during the above mentioned communication additionally increase the students' motivation to be proactive and implement hands-on experience?
3. Does the communication between teacher and students, using the social networks motivate the students to use actively open educational resources?

2.2 Description, participants, methods

In this digital world, opportunities for education are available like never before. Though teachers using online tools are empowering students to take part in their education, they may also expose them to inappropriate material, sexual predators, and bullying and harassment by peers.

Teachers who are not careful with their use of the sites can fall into inappropriate relationships with students or publicize photos and information they believed were kept private. For these reasons, critics are calling for regulation and for removing social networking from classrooms – despite the positive affects they have on students and the essential tools they provide for education in today's digital climate. The positive effects of social networking sites in education are profound and the students who are already engaging in social networking could benefit from incorporating it into curriculum. Through utilizing teaching techniques that incorporate social media, teachers should be able to increase students' engagement in their education, increase technological proficiency, contribute to a greater sense of collaboration in the classroom, and build better communication skills.

Having in mind that according to the State Statistical Office, 82.0% of the Internet users, from which the majority are the students, used the Internet for the social networking purpose, Facebook and Tweeter mostly, it is important to explore the possible benefits that such communication can offer to the modern education.

Use of social networks starts in the classroom, by giving modified WebQuest assignment with specific tasks, links to the relevant resources and clearly defined process. Using the social networks, the communication between the teacher throughout and the students does not end when school year ends, on the contrary, it becomes a form of a distant learning.

Students from three different classes and three different schools (two urban and one rural school) participated in the research: fourth, fifth and sixth grade students (age nine to eleven). Besides the student, their parents and teachers were also included in the research activities. Three different subjects were identified to be used during the research: mother tongue, mathematics and IT. The students and their teacher used their Facebook and Tweeter profiles for communication.

Prior to this research, all parents were informed about the use of Facebook and Tweeter for communication and they all signed that they approve their children to be part of this activity.

The task for the mother tongue assignment was presented as authentic situation when the school library is closed and the students are not able to borrow the necessary books for reading during the summer period. The teacher using the modified WebQuest method gave the assignment to students. The suggested resources varied

from printed to e-books, videos and multimedia. The students were given the opportunity to choose the way of the presentation of the completed tasks. During the whole process, the teacher and the students communicated using Facebook and Tweeter.

Math assignment was presented to the students as WebQuest authentic situation in which the students prepare themselves for the regional geometry completion using geometry application. The teacher used her Facebook profile to give geometry tasks and problems to her students and the students used their profiles to post the solutions on their walls.

For the purpose of IT assignment, each day the standardized tests were given to the students using Facebook.

The students and their parents were given questionnaires to provide feedback on the methods used for the purpose of this research. The group interview with the students and their parents was also valuable resource of information. The analytical descriptive method was used and qualitative and quantitative indicators presented by the measurement of the achievements and by the students, parents and teacher responses.

3 Results and data analysis

On the basis of the analysis the Opinions of the respondents covered with the surveys and interviews it was determined that:

According to the computer usage, 18% of the students reported that they use computer at school only, while the remaining 82 % use the computer at school and at home. In regards the use of social networks, 10% of students (and these are the youngest ones) do not have Facebook or Tweeter profile. 40% of the students have their Facebook profiles opened by their parents, 30% opened their profiles by themselves but with the approval from their parents, and 20% opened their profiles by themselves without approval from their parents. 70% of the students uses Facebook more than a year, 20% uses few months, and (as mentioned before) do not use Facebook. 18% of the students use Tweeter in the last few months and 72% do not use it at all.

About the effect of the communication between teacher and students, using the social networks, increase the students' motivation for learning

80% of the students stated that they communicated regularly with their teacher during the assignment, and 20% communicated from time to time. 80% of the students communicated with their classmates using Facebook, but using other communication tools (as phones etc.), and 20% used only Facebook. The students answered that the teacher's responses were in a timely manner and 18% waited for the response. 90% consider that the teacher provided valuable information and 88% confirmed that the communication with the teacher helped them to complete the given task. 85 % of the students consider that collaboration between them and the teacher enforces the learning. 80% would like to have other subjects be organized in a similar way.

The parents responded that they are very satisfied with such communication between their children and the teacher and that they would like the communication to continue during the summer break, because it is a good motivation for the students to have their responsibilities always on their minds.

All students responded that they would like to have the similar approach in learning in other subjects too. Students and their parents are very satisfied with the used method of learning and communication and they would like teacher to continue such communication with the students in the future.

The students' and parents' responses lead to the conclusion that: The communication between the teacher and the students using social networks increases the students' motivation for studying, which was expected. But wanting to go in depth, it was very important to find out the additional elements that increases the motivation during the teacher – student communication.

About the effect of the WebQuest method during the mentioned communication and the increased students' motivation to be proactive and implement hands-on experiences tasks.

The use of WebQuest method significantly draws the students' attention. It allows them to solve real life problems using Internet based educational resources. Besides the dilemma which way is the best to motivate students to do the reading during the summer period, part of the research was focused on comparing the traditional printed readings vs. digital editions found on Internet. It is known fact that the students consider the readings bored, so it was important to find creative and motivating tactic to encourage them to learn during spending time in front of the computer.

The given WebQuests motivated the students to spend their spare time and searching through the given resources to increase their knowledge. The students feel that the offered method of individual research provides more permanent knowledge which is not the case when they have to memorize the facts only. Student activities are very organized by the WebQuest and they can stay focused on using information rather than finding it. It extends students' thinking to the higher levels of Bloom's Taxonomy; analysis, synthesis, and evaluation and support critical thinking and problem solving through authentic assessment, cooperative learning, scaffolding, and technology integration. But also it encourages independent thinking and to motivate students. Differentiate instruction by providing multiple final product choices and multiple resource websites. Multiple websites as reading content allow students to use the resource that works best for their level of understanding. It helps the students to bridge the gap between school and "real world" experiences and encourages students to become connected and involved learners.

The received feedback from students allows us to conclude that: The use of the WebQuest method during the teacher – student communication increases students' motivation to be proactive and implement hands-on experiences tasks.

The communication between teacher and students, using the social networks motivate the students to use actively open educational resources.

The answers to the question regarding the previous experience in the use of the Internet based resources leads to the fact that majority of the students (90%) use Internet as a resource for their homework or project based activities within the last year or less.

76% of the students consider that the web based resources such as e-books and/or videos helped them better to understand the given tasks and complete them successfully, while 11% consider that the Internet has no influence in completing the assignments.

84% of the students that the Internet resources are appropriate to their age and understanding and 64% were motivated to look for additional Internet and digital resources.

The parents pointed that they like the idea to use the Internet educational resources very much and that they will be very happy if the quality as well as the quantity of the offered educational resources increase and is available for all subjects.

Having in mind the results received from the questionnaire, but from the group interview with the students and their parents it is fair to say that the communication through the social networks motivates the students to actively use open educational resources.

4 Conclusion

Nobody would dispute that the risks of children using social media are real and not to be taken lightly. But there are also dangers offline. The teachers and parents who embrace social media say the best way to keep kids safe, online or offline, is to teach them. The educational benefits of social media far outweigh the risks, and they worry that schools are missing out on an opportunity to incorporate learning tools the students already know how to use.

Traditional education tactics often involve teacher-given lectures, students with their eyes on their own papers, and not talking to your neighbor. However, social media as a teaching tool has a natural collaborative element. Students critique and comment on each other's assignments, work in teams to create content, and can easily access each other and the teacher with questions or to start a discussion.

Social networking could be a useful tool and the educational system should allow the use of these sites because students benefit from it in many ways. They can build skills that they may need once they finished their education. Overall, the social networking has great benefits, but is very important that the teacher implements its use properly and that it contains educational value.

References

1. Paris OER Declaration, (June 20-22, 2012), World Open Educational Resources (OER) Congress, UNESCO, Paris,

http://www.unesco.org/new/fileadmin/MULTIMEDIA/HQ/CI/CI/pdf/Events/English_Paris_OER_Declaration.pdf
2. Social Networking In Schools: Educators Debate The Merits Of Technology In Classrooms, Huffington Post March 27th 2011 http://www.huffingtonpost.com/2011/03/27/social-networking-schools_n_840911.html
3. Usage of information and communication technologies in households and by individuals (2011), 21.10.2011, Year XLIX, No: 8.1.11.25, State Statistical Office, Skopje, Macedonia. Retrieved from http://www.stat.gov.mk/pdf/2011/8.1.11.25.pdf
4. Ten ways schools are using social media effectively, e_School News, Technology News for Today's K-20 Educator, October 1st 2011 http://www.eschoolnews.com/2011/10/21/ten-ways-schools-are-using-social-media-effectively/5/
5. The Case For Social Media in Schools, Machable Tech, September 29th 2010 http://mashable.com/2010/09/29/social-media-in-school/

Partial Learning Model of Dyslexic Learner

Velimir Graorkoski[1], Ana Madevska-Bogdanova[2], Marjan Gusev[2]

[1] Z-SoftNet, P.O. Box 1685, Portland OR, 97207, USA

[2] Faculty of Computer Science and Engineering, University of "Sv. Kiril i Metodij", Skopje, Macedonia

{veljo@z-softnet.com, ana.madevska.bogdanova@finki.ukim.mk, marjan.gushev@finki.ukim.mk}

Abstract. According to the new definition of the learning components in the advanced adaptive learning in comparison with the basic adaptive learning, the process of learning is different and closely related to the human learner. In order to demonstrate the key improvements, we present the developed model reflecting the dyslexia state of a human and its behavior in the adaptive learning environment. Although not all dyslexia symptoms are covered, the model helps understanding the purpose of these learning environments for subjects with learning disabilities.

Keywords: basic adaptive learning - BAL, partial learning – PL, blank concept – BC, advanced adaptive learning – AAL, adaptive learning environment – ALE, adapted learning environment – IALE, learning mechanism – LM, in concept – IC, test set - TS

1 Introduction

After the development of several strategies for testing and instntiation using simulated learner models [7], we entered the phase where we can simulate a human learner on a way different than the one made in the basic adaptive learning conditions. During our previous research stages, we mostly used artificial learners typical for the machine learning methods [8] in order to prove the justification of the instantiation process. However those learners, although still suitable for experimenting within the advanced adaptive learning [2], can not show the improvements the partial learning brings over the traditional adaptive learning interpretation. This is the reason of the need for simulating different learner models closely related to the human behavior.

The first choice for a learner model was the one with the dyslexia syndrome and although we do not treat all of the symptoms, the simulation strives to show the potential usage and benefits of the advanced adaptive learning methods. Its success will be a huge step for our further research with other human related learner models which is a small part of our motivation for more efficient overall learning.

S. Markovski, M. Gusev (Editors): ICT Innovations 2012, Web Proceedings, ISSN 1857-7288

2 Key Facts About Dyslexia

Dyslexia [4], [5] is a syndrome of learning disability in the part of the information processing. The most frequent case of dyslexia is the visual processing impairment, especially affecting the reading, writing and calculating skills of an individual. Difficulties in reading, initially alter the learning process of a human, since textual expressions are very hard to recognize properly and thus their processing could result in totally wrong reasoning or conclusion.

Our analysis of the characteristics of dyslexia and effects it has on a learner, resulted in focusing only on the cases of acquiring information from textual expressions. The reason for this is the Awareness system [1] itself, since it is based on our research of the adaptive learning, so far using only string data types. In future research we expect this to change and to cover wider range of dyslexia effects with different input types layered in different stages of information processing cycle [14].

We derived two main questions to answer in order to successfully proceed with the development of a learner model with dyslexia:

- Which characteristics of the individual suffering from dyslexia can be represented with LM?
- Is it possible for the LM of a dyslexic learner to be recognized by the ALE? If yes, what TS has to be constructed for that purpose?

3 LM of Dyslexic Learner

The simulation of LM of dyslexic learner is intentional in its nature. Unlike the intentional simulations examined in [7], this one is not an easy task because of two reasons:

- difficulty in representing different types of dyslexia characteristics in PL conditions;
- difficulty in sustaining the PL process with the existing methods;

Most characteristics and effects of text processing impairment, such as inadequate phonological processing abilities [4] or visual discomfort [4], can not be described through the LM in PL because they are not related with the knowledge units.

For the following, however, it is possible to make a representation in the PL:

- short-term memory [4] – can not be directly expressed in LM but we propose an assumption that the concepts are learned by smaller number of ICs so that they can be easier to remember;
- lack of ability to associate individual words with their correct meanings [5] – the concept in LM can be learned by either non existing ICs or by ICs non related to it in ALE;

The short-term memory representation in LM can have a tolerance factor since the number of ICs for a concept in ALE can vary. Therefore in our simulation of the first dyslexia characteristic we can use a reference number so that the concepts which have IC set with equal number of ICs or less than its value are included in LM. We cannot forget the initial step of choosing a random number from 1 to $|V_{ALE}|$ as the total

number of LM's concepts - $|V_{LM}|$. Thus we have the following main strategies for LM simulation by including the concepts:

- having an IC set with the smallest possible number of ICs in the ALE (LM 1 on Fig. 1);
- which include an IC set with number of ICs equal to or less than a chosen number (LM 2 on Fig. 1);

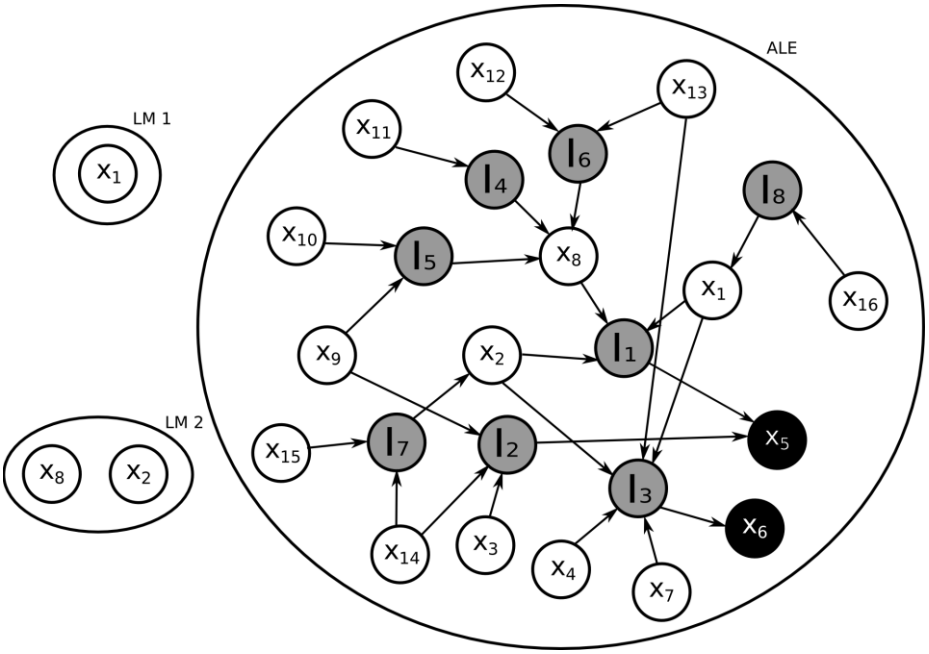


Fig. 1. LM simulation of short-term memory

In order to explain the example on Fig. 1, we state that LM 1 is simulated by choosing the concepts with IC set consisting of only one IC since 1 is the smallest possible number of ICs of all the IC sets in ALE. Only the concepts x_1 and x_8 satisfy this condition, but x_1 is included since the chosen random number of $|V_{LM}|$ is 1.

On the other hand, LM 2 is simulated by choosing the concepts with IC set consisting of 2 or less ICs. Besides the previously mentioned concepts, the concept x_2 satisfies this condition too, but x_1 is omitted because the chosen random number of $|V_{LM}|$ is 2, and the concepts x_2 and x_8 are chosen instead.

The second characteristic of dyslexic learner (not being able to associate some words with their correct meanings), is even harder to simulate because it occurs less frequent and a right choice has to be made among the incorrectly described concepts.

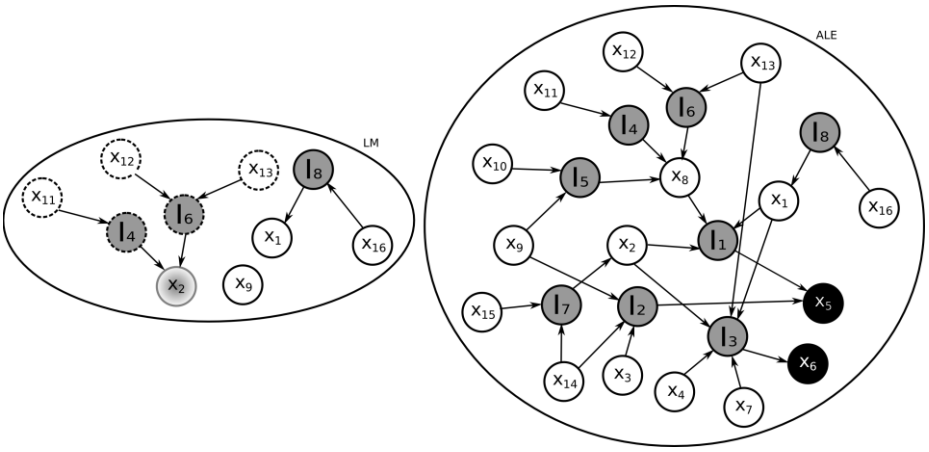


Fig. 2. LM simulation of the occurrence of concept with incorrect meanings

Associating a word with incorrect meanings can be represented as having a concept with incorrect IC sets and ICs, which provide knowledge for different concept in the ALE. Usage of this representation will result in LM which breaks the rule of BAL – in order to achieve AL, the LM must be structured as a subgraph of ALE's knowledge graph. However, in PL conditions we will treat the "confused" concepts only to recognize the LM as the one belonging to a dyslexic learner. These concepts will not take part in the further learning process.

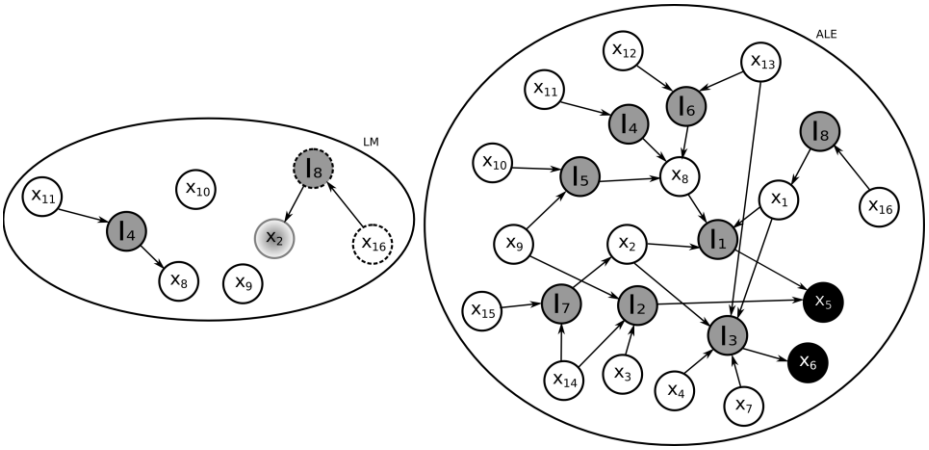


Fig. 3. LM simulation of both dyslexia symptoms

The simulation is done by random choosing of one or more pairs of concepts with the existing IC sets. Then for each of the pairs, a single concept is chosen to be

included in the LM. We choose one subset of the set of IC sets of the concept that is not included in the LM, to provide the knowledge to the chosen concept in the LM. The LM can also include concepts with correct relations.

As shown in Fig. 2, the LM contains one concept (x_2) with knowledge passed from the IC sets determined by the BCs I_4 and I_6. However these two BCs pass knowledge for the concept x_8 in ALE. This means that the learner's LM does not provide knowledge properly i.e. the learner is falsely aware of the concept x_2 (learned through BC I_7 in ALE). The advantage in this situation is that the ICs providing knowledge for I_4 and I_6 can be used as a base for easier learning of other concepts, in this case the correct concept x_8.

The whole picture of LM representing the dyslexic learner with both characteristics can be given by combining the two simulations into one. Fig 3. shows the LM containing concept x_8 as the concept learned by only one IC and concept x_2 incorrectly learned by concept x_{16}, besides the root concepts x_9 and x_{10}.

4 Detection of the LM of Dyslexic Learner

Important task in the AAL when experimenting with dyslexic symptoms is to be certain that the LM belongs to a learner suffering from dyslexia. This is very hard to determine because some learners can show the short-term memory and/or incorrect meanings effects, but they might not suffer from dyslexia. This kind of miss chiefs are rare but possible especially when it comes to relatively small ALEs. That is why our model so far can detect the LM as dyslexic only with certain probability.

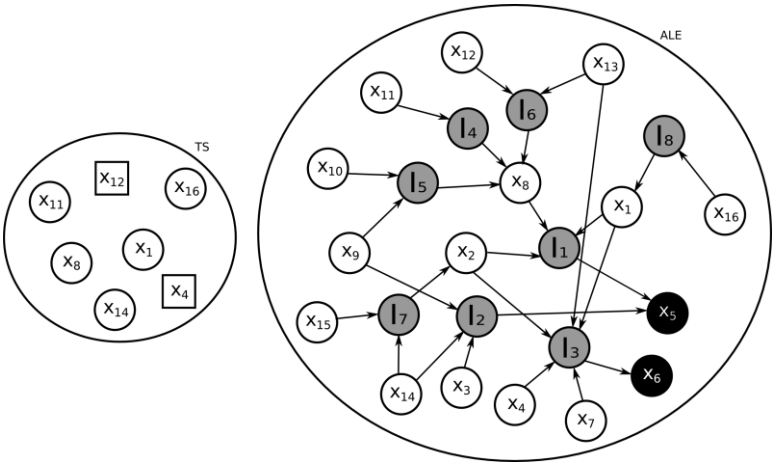


Fig. 4. TS for detection of the short-term memory symptom

In order to make the prediction of the LM as correct as possible, special TSs must be constructed. We will consider the prediction as correct if the LM shows at least one of the two previously mentioned symptoms.

For the purpose of detecting the short-term memory symptom, the TSs will be filled dynamically – if the learner states that the concept is known and the concept is related with the smallest number of ICs, then the ICs as well will be included in the TS.

Suitable example is shown on Fig. 4. The key concepts are x_1 and x_6 because they can be learned by the smallest number of ICs possible – 1, and most of them (in this example both) must be included in the TS in order to make the test relevant. If they are positively confirmed1 by the learner, their ICs are included in the TS as well, until one of them has negative confirmation. For the root concepts x_{14} and x_4 there are no proper conclusions to be made besides the fact that the learner is aware of the first one, and has not previously learned the second one, because there are not any BCs in the TS whose knowledge is provided by them in the ALE.

The dynamic filling of the TS occurs also when it comes to detecting the symptoms of incorrect meanings. Obviously the difference from the previous symptom is the more clear impact of the confirmation results. The best choices for initial concepts to be included in the TS are those with more IC sets and/or ICs. When asked about the confirmation, if positive, then some of the concept's ICs will be included as well until it is confirmed they are truly known by the learner. Otherwise, if the concept is not known by the learner, then the filling continues with other initial concepts.

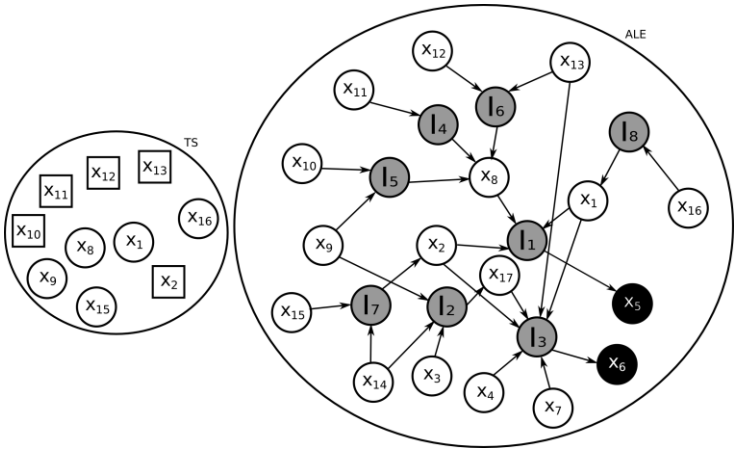


Fig. 5. TS for detection of the incorrect meanings symptom

The example on Fig. 5 shows the key concepts x_1, x_2 and x_8 as the ones that will make the decision whether the learner shows signs of dyslexia. The root concept x_{15} is irrelevant in the final conclusion because its concept x_2 is not positively confirmed by the learner. Let x_1 be the first to fill the TS. Since it is confirmed positively, the next to go will not be x_2 but x_1's IC x_{16}. Because x_{16} is confirmed positively and there is no other concept for it in the ALE, we conclude that the learner has learned the concept

1 The circles denote the positive confirmation and the squares negative

x_1 by previously learning x_{16}. Thus there is no dyslexic problem with the learner here, so far.

After the negative confirmation of x_2 the next in the queue is the concept x_8. It is positively confirmed and we start to check each of its ICs by IC set (first x_9 and x_{10}, then x_{11} and finally x_{12} and x_{13}). Because only x_9 is confirmed positively, the conclusion is that there is no way for the learner to be aware of the concept x_8 by having learned only x_9 from all of its ICs. The case would be different if only x_{11} was positively confirmed but in this situation besides x_9, the IC x_{10} must be learned along in order to complete the knowledge passed from the BC I_5 to the concept x_8.

5 Instantiation for the LM of Dyslexic Learner

Having simulated or detected the LM of the dyslexic learner, the final step before the start of the learning process is the creation of the IALE. The purpose of IALE, as in every learner model scenario, is to provide the required concepts prior to the learning of the terminal concepts.

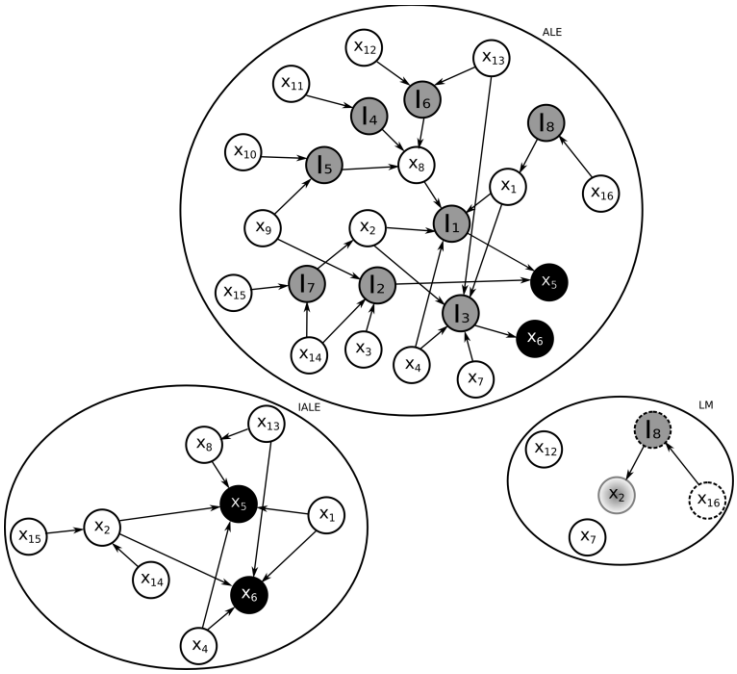


Fig. 6. IALE with smallest possible set of concepts for dyslexic learner

However there are two possible interpretations of the best IALE for dyslexic learners, as a consequence of the interpretations of the short-term memory symptom:

- IALE contains the smallest possible set of ALE's concepts required to reach the terminal concepts;

- IALE contains only the concepts representing the smallest set to complete a required IC set in order to reach the terminals;

The first interpretation is generic and such instantiation can be applied for every learner model as shown in Fig. 14. It can be noticed that a total of 8 concepts are required in order to learn the terminals.

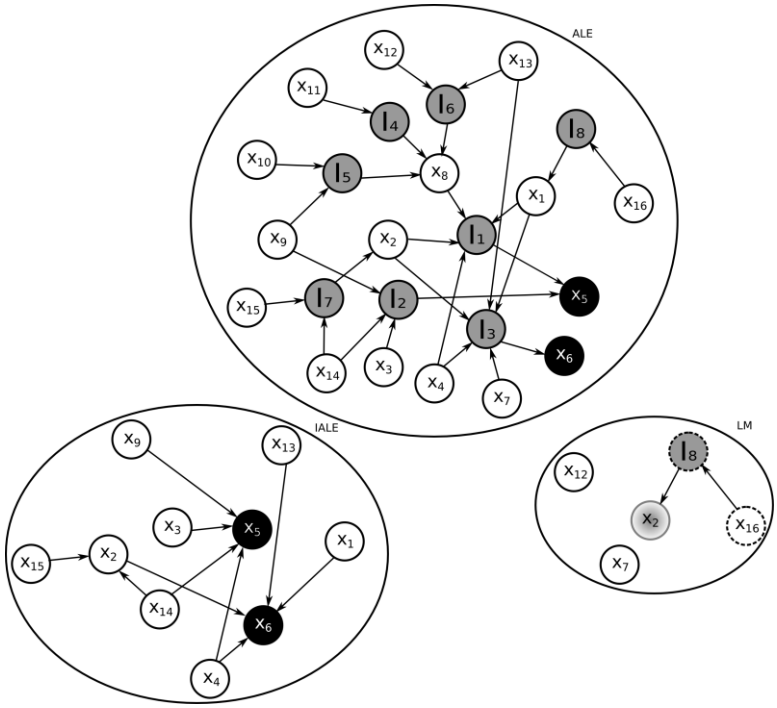


Fig. 7. IALE with smallest number of required ICs for dyslexic learner

The second interpretation is adapted for usage in case of dyslexic learners only. It takes advantage of the short-term memory explanation when it is always better to learn each concept by the smallest number of ICs possible even if it results in a greater total number of concepts required to learn the terminals. To easily construct the IALE we prefer to move backwards starting from the terminal concepts and then choose the IC set which needs smaller number of ICs in order to be complete. Fig. 15 shows an example of such IALE where the BC I_2 is chosen over I_1. Although this improves the chances of learning the concept x_5 with only 3 ICs, it increases the total number of concepts needed to learn both terminals (8 instead of 7).

6 Conclusion

In this paper we managed to simulate some of the behaviors of an individual suffering from dyslexia. In order to achieve the communication with the LE, we succeeded in simulating a learner model reflecting two dyslexia symptoms. Our previous research and conclusions [7] of the PL, was used to make a bridge with the learning process in reality, expressing the short-term memory and incorrect meanings symptoms as if they were features of the PL's components. In addition we showed the possible interactions between the ALE and the LM of a dyslexic learner, explaining the detection of the dyslexia symptoms and the generation of IALE for an LM detected as dyslexic. However, we did not cover the final learning process and the order of the concepts to be learned by the dyslexic individual. The process is the same as with every other learner model and the delivery order strategies are applied as well.

When comparing the two IALE it is obvious that the second representation is faster and easier to implement. The reason for this is the avoidance of the calculation of every possible combination of IC sets in order to reach the terminals. And the increasing of the total number of learned concepts is often insignificant. Therefore our conclusion will prefer using the second type of IALE when the dyslexia learner is using our AAL system and the first type in general cases of learners.

With our research, development and upgrade of the AAL model, we want to show the possibilities laying behind the knowledge representation of learners with disabilities considering the information processing. The existing adaptive assistants for dyslexia learners like AGENT-DYSL [6], [9] try to solve only the surface consequences of dyslexia by using adaptive text annotation techniques or making a choice between different multimedia presentations, but are unable to manipulate with the knowledge units on lower level like we do with our learner model in PL. Also our research focuses on detecting and working with dyslexic learners in real time, unlike the long term examination and following of the development of the learners during certain time periods, as given in [11] or workaround the syndrome by learning games [13]. The practical results of our learning system will follow after our research in PL is completed, because of the difficulties in dividing the learning material as optimal as possible. The main obstacle still remains in the form of detecting a dyslexic state of the student, which has its roots in the identification problems explained in [10].

References

1. Graorkoski, V., Gusev, M.: Model of adaptive e-learning. Master thesis. Institute of Informatics, PMF, University “Sv. Kiril i Metodij”, Skopje (2010)
2. Graorkoski, V., Madevska-Bogdanova, A., Gusev, M.: Beyond the basics of the adaptive learning. ICT-ACT conference, Skopje (2011)
3. Graorkoski, V., Madevska-Bogdanova, A.: Usage of RDF blank nodes in partial learning. CIIT, Bitola (2012)
4. Lucid Research Ltd.: Understanding dyslexia. www.Lucid-Research.com (2006)
5. S. Garg, C., S. Patel, J., Jyoti Sen, D., J. Vyas, P., N. Barot, H., J. Shah, M., H. Shah, D.: Dyslexia: The Developmental Reading Disorder. Department of Pharmaceutical Chemistry

and Quality assurance, Shri Sarvajanik Pharmacy College, Gujarat Technological University, Arvind Baug, Mehsana-384001, Gujarat, India (2011)
6. Schmidt, A., Schneider, M.: Adaptive Reading Assistance for Dyslexic Students: Closing the Loop. FZI Research Center for Information Technologies Haid-und-Neu-Str. 10-14, 76131 Karlsruhe (2007)
7. Graorkoski, V., Madevska-Bogdanova, A.: Learner Models and Instantiation in PL. Technical report at FINKI, University "Sv. Kiril i Metodij", Skopje (2012)
8. Graorkoski, V., Madevska-Bogdanova, A.: Usage of machine learning methods in adaptive learning environment. Technical report at Institute of Informatics, University "Sv. Kiril i Metodij", Skopje (2009)
9. Tzouveli, P., Schmidt, A., Schneider, M., Symvonis, A., Kollias, S.: Adaptive Reading Assistance for the Inclusion of Learners with Dyslexia: The AGENT-DYSL approach. School of Electrical & Computer Engineering, National Technical University of Athens
, Greece (2008)
10. L. Huitt, K.: Teaching Dyslexic Students. Journal of Learning Disabilities (1999)
11. I. Nicolson, R., J. Fawcett, A.: Dyslexia, Learning and the Brain. Massachusetts Institute of Technology (2008)
12. Sangineto, E., Capuano, N., Gaeta, M., Micarelli, A.: Adaptive course generation through learning styles representation. DI, Department of Informatics, University of "La Sapienza", Rome (2007)
13. Smythe, I., Giulivi, S.: A Model of Dyslexia-Friendly Language-Learning Computer Game. University of Wales - United Kingdom, DFA-SUPSI - Switzerland (2011)
14. Reid, G.: Learning Styles: Enhancing Learning. Learning Styles and dyslexia, Paul Chapman publishers (2005)

Design optimization of distribution transformers based on Differential Evolution Algorithms

Rasim Salkoski[1], Ivan Chorbev[2]

[1] University for Information Science and Technology, Building at ARM, 6000 Ohrid, R. of Macedonia
rasim.salkoski@uist.edu.mk
[2] Faculty of computer science and engineering, University of Ss Cyril and Methodius, Rugjer Boshkovich 16, P.O. Box 393, 1000 Skopje,R. of Macedonia
ivan.chorbev@finki.ukim.mk

Abstract. Genetic algorithms and their variants have been extensively used for solving combinatorial optimization problems. One area of great importance that can benefit from the effectiveness of such algorithms is electric energy distribution. Transformers deserve extensive treatment in the field of research and production, due to the fact that the electric energy undergoes several transformations on its way from generators to the consumers. In that regard, special interest is dedicated to the minimization of production and exploitation costs of a transformer. In the paper the combinatorial optimization algorithm based on Differential Evolution is described and applied to the problem of minimizing the cost of the active part of wound core distribution transformers. Constraints imposed both by international specifications and customer needs are taken into account. The Objective Function that is optimized is a minimization dependent on multiple input variables. Constraints are normalized and modeled as inequalities.

Keywords: Combinatorial optimization, Transformer design optimization methodology, Differential Evolution algorithm, Optimization methods, distribution transformer, Wound core type transformer.

1 Introduction

With the very rapid development of computers, transformer designers are freed from the cumbersome routine calculations. Within a matter of minutes or even seconds, computers can generate a number of different transformer designs (by changing current density, flux density, core dimensions, type of magnetic material and so on) and eventually come up with an optimum design. Because of the software design approach and the ease of making multiple iterations of the same design layout, it is easy to optimize the transformer to use a minimal set of expensive materials. There are several packages which are covering the branch of transformer optimization like: Non-linear optimization program (**TOPT)** , Transformer Tap Optimization software analysis module (**ETAP)**, Transformer Design Optimization (**TDO**) software

S. Markovski, M. Gusev (Editors): ICT Innovations 2012, Web Proceedings, ISSN 1857-7288

package for transformer design optimization and economic evaluation analysis and etc. The difficulty in resolving the optimum balance between the transformer cost and its performance is becoming even more complicated nowadays, as the main transformer`s materials (copper or aluminum for transformer windings and steel for magnetic circuit) are stock exchange commodities and their prices vary daily.

Techniques that include mathematical models containing analytical formulas, based on design constants and approximations for the calculation of the transformer parameters are often the base of the design process used by transformer manufacturers. Genetic algorithms and their variants have been extensively used for solving combinatorial optimization problems. One area of great importance that can benefit from the effectiveness of such algorithms is electric energy distribution. The work in this paper introduces the use of an evolutionary algorithm, named Differential Evolution (DE) in conjunction with the penalty function approach to minimize the transformer active part cost while meeting international standards and customer needs. A simple additive penalty function approach is used in order to convert the constrained problem into an unconstrained problem. Due to this conversion, the solution falling outside the feasible region is penalized and the solving process is guided to fall into the feasible solution space after a few generations. The method of penalty function approach is very sensitive when the penalty parameters are large. Penalty functions tend to be very sensitive near the boundary of the feasible domain and that result in a local optimal solution or an infeasible solution. It is always necessary to have careful selection of the penalty parameters for the proper convergence to a feasible optimal solution.

Moreover, the proposed method finds the global optimum transformer design by minimizing the active part cost while simultaneously satisfying all the constraints imposed by international standards and transformer user needs, instead of focusing on the optimization of only one parameter of transformer performance (e.g., no-load losses or short-circuit impedance). Using the proposed technique, a user-friendly DE computer program is developed that combines transformer design with analysis and optimization tools, useful for design optimization. The method is applied to the design of distribution transformers of several ratings and loss categories and the results are compared with a heuristic transformer design optimization methodology, resulting in significant cost savings.

2 Related work

In this paper the Penalty Function method is implemented to handle the constraint using the Differential Evolution (DE) algorithm. Other authors have proposed different approaches to solve constrained optimization with DE-based algorithms.

B.V.Babu and M. Mathew Leenus Jehan in [9] have applied Differential Evolution with a Penalty Function Method and Weighing Factor Method for finding a Pareto optimum set for the different problems. DE is found to be robust and faster in optimization. DE managed to give the exact optimum value within less generations compared to a simple Genetic Algorithm.

Mezura-Montes and Coello Coello in [12] present a Differential-Evolution based approach to solve constrained optimization problems. Three selection criteria based on feasibility are used to deal with the constraints of the problem and also a diversity mechanism is added to maintain infeasible solutions located in promising areas of the search space. The conventional DE algorithm highly depends on the chosen trial vector generation strategy and associated parameter values used. DE researchers have suggested many empirical guides for choosing trial vector generation.

Storn and Price [7] suggested that a reasonable value for NP should be between 5D and 10D, and a good initial choice of F was 0.5. The effective range of F values was suggested between 0.4 and 1. The first reasonable attempt of choosing CR value can be 0.1. However, because the large CR value can speed up convergence, the value of 0.9 for CR may also be a good initial choice if the problem is near unimodal or fast convergence is desired. Moreover, if the population converges prematurely, either F or NP can be increased.

Recently, Rönkkönen in [15] suggested using F values between [0.4,0.95] with 0.9 being a good initial choice. The CR values should lie in [0,0.2] when the function is separable while in [0.9,1] when the function's parameters are dependent. However, when solving a real engineering problem, the characteristics of the problem are usually unknown. Hence, it is difficult to choose the appropriate CR value in advance.

Zaharie proposed a parameter adaptation for DE (ADE) based on the idea of controlling the population diversity, and created a multipopulation approach [16]. Following the same ideas, Zaharie and Petcu designed an adaptive Pareto DE algorithm for multiobjective optimization and analyzed its parallel version [17].

The researchers have developed some techniques to avoid manual tuning of the control parameters. For example, Das et al. [18] linearly reduced the scaling factor F with increasing generation count from a maximum to a minimum value, or randomly varied F in the range (0.5,1). They also have employed a uniform distribution between 0.5 and 1.5 (with a mean value of 1) to obtain a new hybrid DE variant [19].

3 The Differential Evolution (DE) algorithm

Differential Evolution (DE) algorithm is a population-based stochastic method for global optimization developed by Rainer Storn and Kenneth Price [6],[7] for optimization problems over continuous domains. The original version of DE with constituents can be defined as follows ([8], [14]):

1) The population

$$\begin{aligned} &P_{\boldsymbol{x},g} = \left(\boldsymbol{x}_{i,g}\right), \quad i = 0,1,.....,NP, \;\; g = 0,1,.....,g_{max}, \\ &\boldsymbol{x}_{i,g} = \left(x_{j,i,g}\right), \; j = 0,1,.....,D-1. \end{aligned} \tag{1}$$

where *NP* is the number of population vectors, g defines the generation counter, and *D* the number of parameters.

2) The initialization of the population through

$$x_{j,i,0} = rand_j\left[0,1\right) \cdot \left(b_{j,U} - b_{j,L}\right) + b_{j,L} \; . \tag{2}$$

The *D*-dimensional initialization vectors, $\boldsymbol{b}_L$ and $\boldsymbol{b}_U$ indicate the lower and upper bounds of the parameter vectors $\boldsymbol{x}_{i,j}$. The random number generator, $rand_j[0,1)$,

returns a uniformly distributed random number from within the range [0,1), i.e., $0 \leq rand_j[0,1) < 1$. Indication that a new random value is generated for each parameter is denoted by the subscript j.

3) The perturbation of a base vector $\boldsymbol{y}_{i,g}$ by using a difference vector mutation

$$\boldsymbol{v}_{i,g} = \boldsymbol{y}_{i,g} + F \cdot \left(\boldsymbol{x}_{r1,g} - \boldsymbol{x}_{r2,g} \right). \quad (3)$$

to generate mutation vector $v_{i,g}$. The difference vector indices, r_1 and r_2, are randomly selected once per base vector. Setting $y_{i,g} = x_{r0,g}$ defines what is often called classic DE where the base vector is also a randomly chosen population vector. The random indexes r_0, r_1, and r_2 should be mutually exclusive.

4) Diversity enhancement

The classic variant of diversity enhancement is crossover which mixes parameters of the mutation vector $\boldsymbol{v}$i,g and the so-called ***target vector*** $\boldsymbol{x}$i,g in order to generate the ***trial vector*** $\boldsymbol{u}$i,g. The most common form of crossover is uniform and is defined as

$$\boldsymbol{u}_{i,g} = u_{j,i,g} = \begin{cases} v_{j,i,g} \;\; if \left(rand_j \left[0,1 \right) \leq CR \right) \\ x_{j,i,g} \;\; otherwise \end{cases} \quad (4)$$

In order to prevent the case $\boldsymbol{u}_{i,g} = \boldsymbol{x}_{i,g}$ at least one component is taken from the mutation vector $\boldsymbol{v}$i,g, a detail that is not expressed in Eq. (4).

5) Selection

DE uses simple one-to-one survivor selection where the trial vector $u_{i,g}$ competes against the target vector $x_{i,g}$. The vector with the lowest objective function value survives into the next generation $g + 1$.

$$\boldsymbol{x}_{i,g+1} = \begin{cases} \boldsymbol{u}_{i,g} \;\; if \; f\left(\boldsymbol{u}_{i,g} \right) \leq f\left(\boldsymbol{x}_{i,g} \right) \\ \boldsymbol{x}_{i,g} \;\; otherwise. \end{cases} \quad (5)$$

Along with the DE algorithm came a notation (5) to classify the various DE-variants. The notation is defined by DE/*x*/*y*/*z* where *x* denotes the base vector, *y* denotes the number of difference vectors used, and *z* representing the crossover method. For example, DE/rand/1/bin is the shorthand notation for Eq. (1) through Eq. (5) with $y_{i,g} = x_{r0,g}$. DE/best/1/bin is the same except for $y_{i,g} = x_{best,g}$. In this case $x_{best,g}$ represents the vector with the lowest objective function value evaluated so far. With today's extensions of DE the shorthand notation DE/*x*/*y*/*z* is not sufficient any more, but a more appropriate notation has not been defined yet.

Price and Storn [6] gave the working principle of DE with single strategy [7]. They suggested ten different strategies for DE. Different strategies can be adopted in the DE algorithm depending upon the type of problem to which DE is applied. The strategies can vary based on the vector to be perturbed, number of difference vectors considered for perturbation, and finally the type of crossover used. The following are the ten different working strategies: 1. DE/best/1/exp, 2. DE/rand/1/exp, 3. DE/rand-to-best/1/exp, 4. DE/best/2/exp, 5. DE/rand/2/exp, 6. DE/best/1/bin, 7. DE/rand/1/bin, 8. DE/rand-to-best/1/bin, 9. DE/best/2/bin, 10. DE/rand/2/bin.

As it is explained the general convention used above is DE/*x*/*y*/*z*. DE stands for Differential Evolution, *x* represents a string denoting the vector to be perturbed, *y* is the number of difference vectors considered for perturbation of *x*, and *z* stands for the type of crossover being used (exp: exponential; bin: binomial). Hence the perturbation can be either in the best vector of the previous generation or in any randomly chosen

vector. Similarly for perturbation either single or two vector differences can be used. For perturbation with a single vector difference, out of the three distinct randomly chosen vectors, the weighted vector differential of any two vectors is added to the third one. In exponential crossover, the crossover is performed on the D variables in one loop until it is within the CR bound. The first time a randomly picked number between 0 and 1 goes beyond the CR value, no crossover is performed and the remaining D variables are left intact. In binomial crossover, the crossover is performed on each of the D variables whenever a randomly picked number between 0 and 1 is within the CR value. So for high values of CR, the exponential and binomial crossover methods yield similar results [1].

A strategy that works out to be the best for a given problem may not work well when applied to a different problem. Also, the strategy and the key parameters to be adopted for a problem are to be determined by trial and error. However, strategy-7 (DE/rand/1/bin) appears to be the most successful and the most widely used strategy. In all, three factors control evolution under DE, the population size NP, the weight applied to the random differential F and the crossover constant CR. More details regarding DE are available in [7], [8], [9] and [14].

4 Mathematical modeling of power objects and optimization

A mathematical description of a global constrained minimization problem requires us to apply an appropriate model which has limited number of parameters (design variables). Any kind of optimization problem can be formalized to find the appropriate set of design variables in the multidimensional parameter space, which can optimize the main objective function. In the mathematical notation the optimization problem can generally be represented as a pair (S, f) , where $S \subseteq R^n$ is a bounded set on R^n and $f : S \rightarrow R$ is an n-dimensional real-valued function. The problem is to find a point $\boldsymbol{x}_{min} \in S$ such that $f(\boldsymbol{x}_{min})$ is a global minimum on S . More specifically, it is required to find an $\boldsymbol{x}_{min} \in S$ such that

$$\forall \boldsymbol{x} \in S : f(\boldsymbol{x}_{min}) \leq f(\boldsymbol{x}) \quad (6)$$

$$g_i(\boldsymbol{x}) \leq 0,\ i = 1, 2, \bullet\bullet\bullet, q \quad (7)$$

$$h_j(\boldsymbol{x}) = 0,\ j = q + 1, \bullet\bullet\bullet, m \quad (8)$$

where $\boldsymbol{x}$ is the vector of unknown quantities $\boldsymbol{x} = [x_1, x_2, \bullet\bullet\bullet, x_n]^T$, $g_i(\boldsymbol{x})$ and $h_j(\boldsymbol{x})$ are the restriction constraints, which can be represented mathematically as equations and/or inequations, m and q are integer numbers [13]. Generally, for each variable x_i it satisfies a constrained boundary

$$l_i \leq x_i \leq u_i,\ i = 1, 2, \bullet\bullet\bullet, n \quad (9).$$

In order to find the global optimum design of a distribution transformer, DE in conjunction with the penalty function approach technique is used. The goal of the proposed optimization method is to find a set of integer variables linked to a set of continuous variables that minimize the objective function (active part cost) and meet the restrictions imposed on the transformer design. Under these definitions, a DE algorithm in conjunction with the penalty function approach is focused on the minimization of the cost of the transformer's active part:

$$\min_{x} \sum_{j=1}^{3} c_j \cdot f_j(\boldsymbol{x}) \tag{10}$$

where c_1 is the primary winding unit cost (€/kg), f_1 is the primary winding weight (kg), c_2 is the secondary winding unit cost (€/kg), f_2 is the secondary winding weight (kg), c_3 is the magnetic material unit cost (€/kg), f_3 is the magnetic material weight (kg), and $\boldsymbol{x}$ is the vector of the five design variables, namely the width of secondary winding (a), the diameter of core leg (D), the core window height (b), the current density of secondary winding (g) and the magnetic flux density (B).

The minimization of the cost of the transformer is subject to the constraints:

$S - S_N \leq 0; P_{CU} - P_{CUN} \leq 0; P_{FE} - P_{FEN} \leq 0; U_K - U_{KN} \leq 0$

where: S is designed transformer rating (kVA), S_N is transformer nominal rating (kVA), P_{FE} is designed no-load losses (W), P_{CU} is designed load losses (W), U_K is designed short-circuit impedance (%), P_{FEN} is guaranteed no-load losses (W), P_{CUN} is guaranteed load losses (W) and U_{KN} is guaranteed short-circuit impedance (%).

It should be noted that functions f_1, f_2, f_3, appearing in the objective function (10) are composite functions of the design variables $\boldsymbol{x}$, e.g., $f_1 = f_1(g_1(h_1(\boldsymbol{x})))$ the transformer design optimization problem is a hard problem in terms of both modeling and solving.

The single objective Differential Evolution optimization algorithm with penalty function approach has been applied. The program has two input files, "Limits.txt" and "ParameterLimits.txt" and generates two output files, "ReportDE.html" and "Convergence.txt". Accordingly, the objective function for the model is:

$$f(x_2, x_3, x_5) = \left(3.9655 \cdot 10^4 \cdot x_5 + 2.40546 \cdot 10^5 \cdot x_3 + 2.987 \cdot 10^3\right) \cdot x_2^2 + 1.8924 \cdot x_2^3 + \left(6.96522 \cdot 10^5 \cdot x_2 + 1.42442 \cdot 10^6 \cdot x_3 + 1.3478 \cdot 10^4\right) \cdot x_3 \cdot x_5 \tag{11}$$

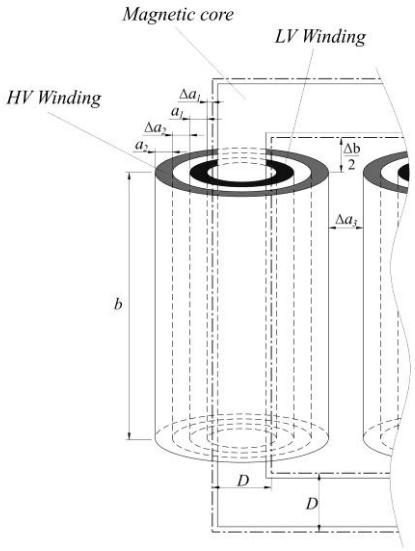


Fig.1 Active part of distribution transformer – main dimensions

The inequality constraints should be modified to the less or equal format, g(x) ≤ 0. If the problem is an unconstrained optimization problem, the user need not enter anything in the space specified for the constraints coding. The constraints of the analyzed mathematical model are entered as follows: Constraint 12 match to transformer nominal rating, Constraint 13 match to guaranteed load losses, Constraint 14 match to guaranteed no-load losses and Constraint 15 guaranteed short-circuit impedance. Constants in front of decision variables have been taken from the Fig.1 and reference [11].

$$317.82 \cdot x_1 \cdot x_2^2 \cdot x_3 \cdot x_4 \cdot x_5 \cdot 10^6 - 50 \cdot 10^3 \leq 0 \tag{12}$$

$$\left(3.638 \cdot 10^{-7} \cdot x_2 + 8.113 \cdot 10^{-7} \cdot x_3 + 7.51 \cdot 10^{-9}\right) \cdot x_3 \cdot x_4^2 \cdot x_5 - 1050 \leq 0 \tag{13}$$

$$\left(-0.4237 \cdot x_1^2 + 1.2712 \cdot x_1 - 0.0241\right) \cdot \left(\left(3.9655 \cdot 10^4 \cdot x_5 + 2.405 \cdot 10^5 \cdot x_3 + 2.987 \cdot 10^3\right) \cdot x_2^2 + 1.892 \cdot x_2^3\right) \cdot 0.4 - 190 \leq 0 \tag{14}$$

$$\left(0.008 \cdot x_2 + 0.0186 \cdot x_2 \cdot x_3 + 0.032 \cdot x_3 + 1.7744 \cdot x_3^2 + 1.6 \cdot 10^{-4}\right) \cdot 317.82 \cdot 0.0186 \cdot x_3 \cdot x_4 / x_1 \cdot x_2^2 - 4.1 \leq 0 \tag{15}$$

These values are multiplied by a penalty co-efficient (entered by the user of the program), which is then added to the objective function to continue the process of optimization. This process is often termed as penalty function approach.

4.1 Experimental results

After inserting the objective function and constraints, the user needs to prepare the two input files ("Limits.txt" and "ParameterLimits.txt"). In the "Limits.txt" input file, the lower and upper bound for each decision variable separated by a tab, is entered. The number of decision variables for the analyzed mathematical model is five.

Lower and Upper bound of decision variables in the "Limits.txt" input file for the analyzed mathematical model is as follows: 1.6 up to 1.8 for the magnetic flux density (*B*) in Tesla, .1 up to .125 for the diameter of core leg (*D*) in m, .015 up to .020 for the width of secondary winding (*a*) in m, 2.4 up to 3.0 for the current density of secondary winding (*g*) in A/mm^2 and .210 up to .230 for the core window height (*b*) in m.

The "ParameterLimits.txt" input file requires the number of decision variables (in the first row), maximum number of generations (in the second row), minimum and maximum number of population (*NP)*, crossover constant (*CR*), weighting factor (*F*) along with their step length for sensitivity analysis in the third, fourth and fifth rows respectively.

The input file for the analyzed mathematical model is as follows: Number of decision variables is 5, Maximum number of generations is 30, Minimum, maximum and step length for *NP* 20,20,10, Minimum, maximum and step length for *CR* 0.8, 0.9, 0.1 and Minimum, maximum and step length for *F* 0.5, 0.6, 0.1 .

This completes the preparation of inputs files. To generate the optimal set of solutions for the optimization problem, the DE program is compiled and executed. The program can be compiled and execution using any standard C compiler for the Windows environment.

When the program is run for different combinations of *NP, CR* and *F*, the optimal set of parameters is determined based on two factors i.e., minimum objective function value and lower CPU time requirement. In any given situation, if minimum objective function values are the same for any given combination(s), the next criteria that is chosen for selecting optimal combination is lower CPU time requirement. In this program, these two factors are considered for choosing optimal set of parameters. The output figures in Table 1 are given for the analyzed mathematical model generated after the successful completion of the program's execution [1]. The optimal value of objective function and decision variables for the optimization problem is also recorded.

Table 1. Output figures with time to complete strategies on each of them

Strat. No.	Strategy	NP	CR	F	Optimal Value	Constraint Violation	NFE	Time Taken(ms)
1	DE/rand/1/bin	20	0.80	0.50	4.976027E+002	0.0000E+000	835	30
2	DE/best/1/bin	20	0.80	0.50	4.973189E+002	0.0000E+000	1429	40
3	DE/best/2/bin	20	0.80	0.50	4.973945E+002	0.0000E+000	681	11
4	DE/rand/2/bin	20	0.80	0.60	4.976027E+002	0.0000E+000	791	20
5	DE/rand-to-best/1/bin	20	0.80	0.50	4.973220E+002	0.0000E+000	1869	40
6	DE/rand/1/exp	20	0.80	0.50	4.975862E+002	0.0000E+000	989	20
7	DE/best/1/exp	20	0.80	0.60	4.973355E+002	0.0000E+000	747	20
8	DE/best/2/exp	20	0.90	0.50	4.975857E+002	0.0000E+000	769	10
9	DE/rand/2/exp	20	0.90	0.50	4.976027E+002	0.0000E+000	703	20
10	DE/rand-to-best/1/exp	20	0.90	0.60	4.973240E+002	0.0000E+000	1319	30

Best Strategy is Strategy DE/rand/2/bin, Minimum constraint violation (CV) : 0.0000E+000, Minimum objective value with min CV: 4.976027E+002 and Minimum time taken : 20

Table 2. Output table results of the analyzed mathematical model

Parameter	Value
X_1	1.630365
X_2	0.100062
X_3	0.015006
X_4	2.968991
X_5	0.210373

The parameters X_1, X_2, X_3, X_4, X_5 match respectively to the magnetic flux density (*B*), the diameter of core leg (*D*), the width of secondary winding (*a*), the current density of secondary winding (*g*) and the core window height (*b*).

Table 3. Comparative results of the analyzed mathematical model with produced object by the specified optimized method

	The mag. flux density (T)	LV wind. Curr. density (A/mm^2)	Diam. of Core Leg (mm)	Width of LV wind. (mm)	The Core window (mm)	The cost of the transf. active part
DE Algorithm with penalty function approach	1.630	2.969	100	15	210	497
Lagrange multipliers with Newton Raphson approach [11]	1.642	2.388	117	20	221	703

5 Conclusion

This paper presents an efficient implementation of a Single Objective Optimization Program using the Differential Evolution algorithm with a penalty function approach, applied to a power object. Our penalty function approach integrates established techniques in existing EA's in a single unique algorithm. Our approach was tested on a three phase distribution transformer and the results indicate that this approach can be used to solve a range of SOOs with linear/nonlinear equality/inequality constraints, as well as continuous/discontinuous search spaces. Moreover, this approach is easy to implement and its computational cost is relatively low.

The use of the DE computer program is applied to the analyzed mathematical model. In the first methodology (Table 3) the single objective DE optimization showed that single optimum could be obtained fast even when constraints in the penalty function method are complex and compared with the second methodology in the same table, the cost materials for the active part of the reviewed object are lower.

References

1 Arunachalam, Vasan,: Optimization Using Differential Evolution. Water Resources Research Report. Book 22. Department of Civil and Environmental Engineering, The University of Western Ontario, Publication Date 7-2008.

2. A. Zamuda, J. Brest, B. Bošković, V. Žumer.: Differential Evolution with Self-adaptation and Local Search for Constrained Multiobjective Optimization. IEEE Congress on Evolutionary Computation (CEC), pp. 195-202 (2009)

3. J. Brest, A. Zamuda, B. Bošković, V. Žumer.: Dynamic Optimization using Self-Adaptive Differential Evolution. IEEE Congress on Evolutionary Computation (CEC) 2009, pp. 415-422. Trondheim, Norveška (2009).
4. Pang-Kai Liu, Feng-Sheng Wang.: Hybrid differential evolution including geometric mean mutation for optimization of biochemical systems, Journal of the Taiwan Institute of Chemical Engineers 41 (2010) 65–72, Department of Chemical Engineering, National Chung Cheng University, Chia-yi 62102, Taiwan (2010).
5. Differential Evolution Homepage, http://www.icsi.berkeley.edu/~storn/code.html
6. Onwubolu, G. C., and Babu, B. V.: New Optimization Techniques in Engineering,Springer-Verlag, Germany (2004).
7. Price V. Kenneth., Storn M. Rainer.: Differential evolution - A simple evolution strategy for fast optimization. Dr. Dobb's Journal, 22, 18-24 and 78. (1997).
8. Price, V. Kenneth., Storn, M. Rainer., and Lampinen, A. Jouni.: Differential evolution: A practical approach to global optimization. Springer-Verlag Berlin,Heidelberg (2005).
9. B.V.Babu, M. Mathew Leenus Jehan .: Differential Evolution for Multi-ObjectiveOptimization. Chemical Enginering Department B.I.T.S. Pilani, India (2005).
10. Elefterios I. Amoiralis, Pavlos S. Georgilakis, Marina A. Tsili .: Design optimization of distribution transformers based on mixed integer programming methodology. Technical University of Athens, Greece (2008).
11. Rasim Salkoski.: Selection of an optimal variant of 3-phase transformers with round and rectangular section of the magnetic core from aspect of minimum production costs. Master Thesis, Electrotechnical University in Skopje (2000).
12. Mezura-Montes.: E. Laboratorio NI Avanzada, Rébsamen 80, Centro, Xalapa, Veracruz 91090, Mexico, Velazquez-Reyes, J., Coello Coello, C.A.: Modified Differential Evolution for Constrained Optimization , pp 25 – 32, Conference Publications, Evolutionary Computation, CEC 2006 (2006).
13. Wenyin Gong, Changmin Chen, Zhihua Cai.: Simple Diversity Rules and Improved Differential Evolution for Constrained Global Optimization School of Computer Science China University of Geosciences, Wuhan 430074, P. R. China (2006).
14. Uday K. Chakraborty (Ed.).: Advances in Differential Evolution, Mathematics & Computer Science Department, University of Missouri, St.Louis, USA, Springer-Verlag Berlin Heidelberg (2008).
15. J. Rönkkönen, S., Kukkonen, and K. V. Price.: Real-parameter optimization with differential evolution. Proc. IEEE Congr. Evolut.Comput., Sep. 2005, pp. 506–513, Edinburgh, Scotland (2005).
16. D. Zaharie.: Control of population diversity and adaptation in differential evolution algorithms. Proc. Mendel 9th Int. Conf. Soft Comput., R. Matousek and P. Osmera, Eds., Brno, Czech Republic, pp. 41–46, Brno, Czech Republic (2003)
17. U. K. Chakraborty, S. Das, A. Konar.: Differential evolution with local neighborhood. Proc. Congr. Evolut. Comput., pp. 2042-2049,Vancouver, BC,Canada (2006).

Mobility Sensitive Admission Control Algorithm for WiMAX-WLAN Vertical Handovers

Kire Jakimoski[1,*], and Toni Janevski[2]

[1]Faculty of Information and Communication Technology,
FON University, Skopje, Republic of Macedonia
kire.jakimoski@fon.edu.mk
[2]Faculty of Electrical Engineering and Information Technologies,
Ss. Cyril and Methodius University, Skopje, Republic of Macedonia
tonij@feit.ukim.edu.mk

Abstract. The admission control in 4G and 5G heterogeneous mobile and wireless networks will be more complex because it will need to deal with many different networks and decide to admit not only new calls and horizontal handovers, but also vertical handover calls. In this paper we propose a mobile sensitive algorithm for admission control in heterogeneous networks for seamless vertical handovers between WiMAX and WLAN. The results show that with the implementation of the proposed admission control algorithm we improve the Quality of Service of the users while moving from WiMAX into WLAN networks.

Keywords: Admission Control, Heterogeneous, VoIP, WiMAX, WLAN.

1 Introduction

Nowadays we have different mobile and wireless networks deployed, and their integration and maximum efficiency is the main subject in the future development in the field of mobile and wireless networks. 4G standardization is already finished in 3GPP radio access with LTE advanced and in non-3GPP radio access with Mobile WiMAX 2.0 (IEEE 802.16m). As the concept of heterogeneous networks already began to implement as an idea in the 4G approach, it is for sure that in 5G approaches [1] mobile users will also have the possibility to access different RATs (radio access technologies) during their sessions without any interruption in the communication process.

The concept of heterogeneous networks deployed in the NGN (Next Generation Networks) requires more control functionalities in the core networks. As each radio access technology has its own radio resource management and own admission control procedure in the radio resource management, when we have different radio interfaces in the new terminals with opportunity to have seamless connection to different RATs, we need more complex admission control during the vertical handovers between dif-

S. Markovski, M. Gusev (Editors): ICT Innovations 2012, Web Proceedings, ISSN 1857-7288

ferent technologies. Moreover, the seamless network handover and continuous service provisioning to users require admission control mechanisms that will satisfy the experience of the users [2].

There are many recent research papers such as [3-6] that deal with the issues of admission control in wireless networks. There are also recent studies about admission control in heterogeneous networks. In [7] the authors propose a new adaptive admission control algorithm for 4G heterogeneous networks that checks if the session is real time or non-real time, new or handover session. In [8] authors also give the design of a new admission control algorithm suitable for 4G networks that considers network load, user's QoS requirements, user's context and link quality.

But, none of the aforementioned and so far published contributions for admission control in heterogeneous networks, including IEEE 802.21 standard, have taken into consideration the effect of the mobile node terminal speed of users to decide upon vertical handover to WLAN networks from WiMAX, UMTS or LTE networks. WLAN network is more sensitive to higher speeds of the mobile terminals comparing with WiMAX or LTE (UMTS), hence it is very important to consider the speed when deciding whether to admit or reject the handoff calls to WLAN network. This fact raises the need for design of admission control algorithm that will be sensitive to the speed of the mobile terminal nodes when vertical handovers are processed from WiMAX to WLAN networks. Information for the mobile node's speed can be detected from the Doppler spread in the received signal envelope [9]. For this purpose in this paper we propose admission control algorithm that will be dependent upon the velocity of the users and will consider velocity threshold regarding the handoff to WLAN network.

The structure of this paper is as follows. Section 2 explains the vertical handover process between WiMAX and WLAN in heterogeneous networks. Section 3 describes the proposed admission control algorithm. Then, in Section 4 are presented results from the performance evaluation of the proposed solution. Finally, Section 4 concludes the paper.

2 Vertical Handover Process between WiMAX and WLAN

The order of events that occurs between the mobile node and network in the process of vertical handover from WiMAX to WLAN coverage is explained in this section. We assume that a specific mobile node starts to move in a WiMAX cell and in its trajectory WLAN network is detected. When this happens WLAN interface from the mobile node detects beacons from 802.11 and triggers the event "Link Detected". MIH (Media Independent Handover) Agent that is in the mobile node is receiving this event and because it is better interface it gives command to the WLAN interface of the mobile node to connect to the WLAN access point.

After this WLAN interface from the mobile node and the WLAN access point exchange frames with "Association Request" and "Response" in order to make a link between the mobile node and the WLAN cell. WLAN interface triggers "Link Up" event after it receives the "Association Response". This event is received from the

MIH Agent in mobile node and after that it commands to the MIPv6 agent of the mobile node to request ND (Neighbor Discovery) Agent in order to send an RS (Router Solicitation).

The access point from the WLAN network receives the RS, so it detects that it is a new neighbor. It reacts on that with sending a RA (Router Advertisement) that includes the router lifetime, prefix valid lifetime, network prefix and advertisement interval. The WLAN interface of the mobile node receives the RA and reconfigures its address in dependence on the received prefix. The MIH Agent of the mobile node is notified about this.

The MIPv6 Agent from the mobile node gives command to the WLAN interface to send "Redirect" message to the CN (Correspondent Node) for the purpose of informing the CN about the new location of the mobile node. The MIPv6 Agent of the CN receives then the "Redirect" message and sends after that an Ack (Acknowledge) message that is received by the WLAN interface of the mobile node. It then notifies the MIH Agent of the mobile node.

Now the MIH Agent from the mobile node has the confirmation that CN knows the new address of the mobile node and redirects the receiving of the traffic from the WiMAX interface to the WLAN interface. Hence, the traffic now uses the link between the WLAN interface from the mobile node and the AP.

The MIH Agent from the mobile node gives command to the WLAN interface to send MIH Capability Request to the access point (AP). MIH Capability Response is responded from the AP including the MIHF (Media Independent Handover Function) identification. Consequently, MIH Capability Response is received from the MIH Agent with the identification of the new remote MIHF identification.

When the mobile node is approaching the boundary of the WLAN coverage the WLAN interface is triggering the event "Link Going Down" that is based on the received power of the beacon frames. The probability that the WLAN link is going down starts to increase because of the speed of the mobile node. When it gets a predefined value (usually 90%) and because the WiMAX interface of the mobile node is still active, MIPv6 Agent of the mobile node gives command to the WiMAX interface to send the message "Redirect" to the CN. In this way CN is informed about the new location of the mobile node. MIH Agent of the mobile node also gives a command to the WLAN interface to execute "Link Scan" for searching another WLAN network.

If a "Probe Response" is received only in the channel where the mobile node is currently, MIH Agent from the mobile node is assured that this is the only accessible WLAN network. Then, MIPv6 Agent of the CN receives the message "Redirect" and sends the message "Ack" that is received by the WiMAX interface of the mobile node. MIH Agent is informed about this.

The MIH Agent of the mobile node is having the confirmation that the CN is informed about the new address of the mobile node and is redirecting the reception of the traffic from WLAN interface to the WiMAX interface. The traffic now uses the link from the WiMAX interface of the mobile node and the base station. In the same time WLAN interface of mobile node is triggering the event "Link Down", hence the mobile node is disconnected from the WLAN network.

3 Admission Control Algorithm

The admission control algorithm determines whether the incoming requests to WLAN network will be rejected or accepted. If the speed of the mobile node while approaching the WLAN network from WiMAX network is above the acceptable threshold for the WLAN network the incoming request for handover will be rejected. If the speed of the mobile node is below the threshold, the user will be admitted to the system and the vertical handover procedure will start. Fig. 1 presents the procedure of the algorithm.

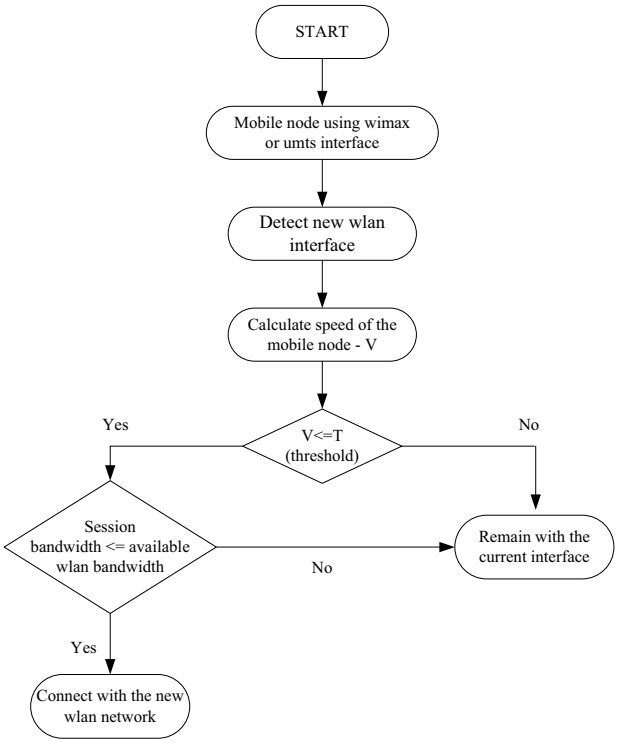


Fig. 1. Admission control algorithm procedure between WiMAX and WLAN.

To achieve better channel utilization of the heterogeneous networks while still satisfying the QoS requirements of the users we designed the following admission control algorithm for handoff calls between WiMAX and WLAN network. Let V be the speed and let T be the threshold speed of the mobile node terminal. Then, the algorithm will be as follows:

```
- Wait for vertical handover request arrival to WLAN
- If a vertical handover request arrives
- Calculate speed of the mobile node
```

```
- If V <=T and if Session bandwidth <= available WLAN
bandwidth
- Admit the request call to WLAN
- Else
- Reject the vertical handover request.
```

In the above admission control algorithm users that move with speed that is above the acceptable velocity threshold will not be allowed to perform handover to WLAN from WiMAX. Hence, such Mobility Sensitive (MS) admission control algorithm can be implemented in any proposed admission control architecture for heterogeneous networks in 4G and 5G standards.

We have implemented and tested our designed MS admission control algorithm in the IEEE 802.21 standard for Media Independent Handover (MIH) [10, 11]. We have implemented the designed MS admission control algorithm into the handover module in network simulator (ns) by enhancing it with the information about the mobile terminal speed, that is now included when deciding whether to admit or reject the vertical handover session from other networks to WLAN.

4 Performance Evaluation

In our simulations, for the purpose of testing the above proposed admission control algorithm, we use a two-tier heterogeneous wireless network structure that is composed of a WiMAX cell overlaying a WLAN hotspot.

Table 1. Types of Mobile Nodes.

Mobile terminals	Velocity of the MN [kmph]	Type of Movement	Type of traffic
1, 2	3.6	Only in WiMAX coverage	MPEG-4
3, 4	3.6	Only in WiMAX coverage	FTP
5, 6	3.6	Only in WiMAX coverage	Telnet
7-10	5	Starting in WiMAX, crossing through WLAN, ending in WiMAX	VoIP G.711
11-14	10	Starting in WiMAX, crossing through WLAN, ending in WiMAX	VoIP G.711
15-18	15	Starting in WiMAX, crossing through WLAN, ending in WiMAX	VoIP G.711
19-22	20	Starting in WiMAX, crossing through WLAN, ending in WiMAX	VoIP G.711
23-26	25	Starting in WiMAX, crossing through WLAN, ending in WiMAX	VoIP G.711
27-30	30	Starting in WiMAX, crossing through WLAN, ending in WiMAX	VoIP G.711

The simulation study was conducted by simulating 30 mobile terminal nodes attached to the WiMAX network. 6 mobile terminals are moving only in WiMAX coverage and they are using MPEG-4, FTP and Telnet traffic. The other 24 terminals are moving across WLAN network and they are performing vertical handovers between WiMAX and wireless LAN. Detailed characteristics of the mobile node terminals are explained in Table 1. The network topology and types of the trajectories of the mobile nodes are shown in Fig. 2 and Fig. 3. Using this model each node moves along a straight line from some starting point to end point. We use 4 types of traffic in the simulations as examples of conversational, streaming, web and background traffic. They are VoIP G.711, MPEG-4, Telnet and FTP sessions. VoIP G.711 traffic is simulated with a packet size of 160 bytes at application layer and inter-arrival packet time of 20 ms.

The propagation model is TwoRayGround, considering both the direct path and a ground reflection path. The total simulation time is 200 seconds.

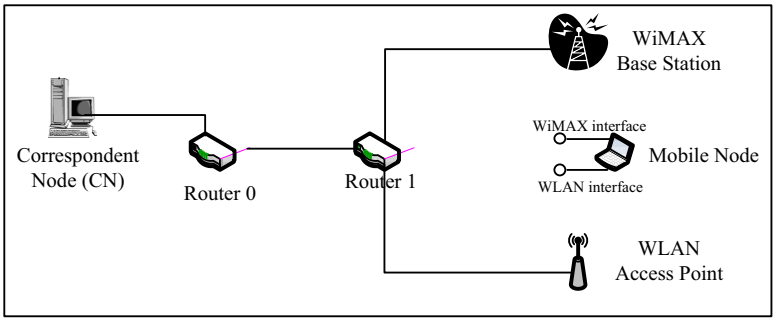


Fig. 2. Heterogeneous network topology.

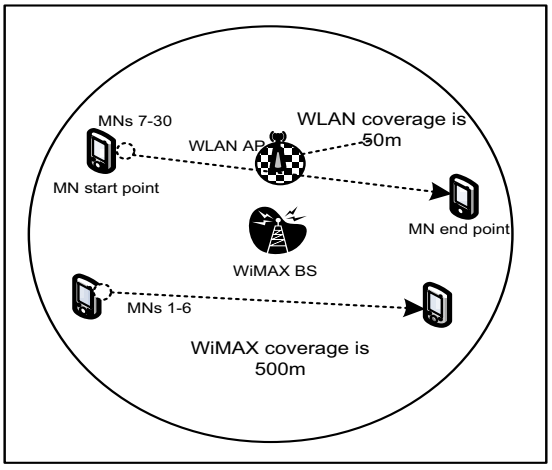


Fig. 3. Heterogeneous network scenario with 30 mobile terminals.

7

Firstly, the simulations are done without applying our designed algorithm. After analysis of the results we repeated the same simulation scenario with previously applied MS admission control algorithm. The compared results of the QoS performances of average throughput, packet loss and vertical handover latency are shown in Fig. 4, Fig. 5 and Table 2.

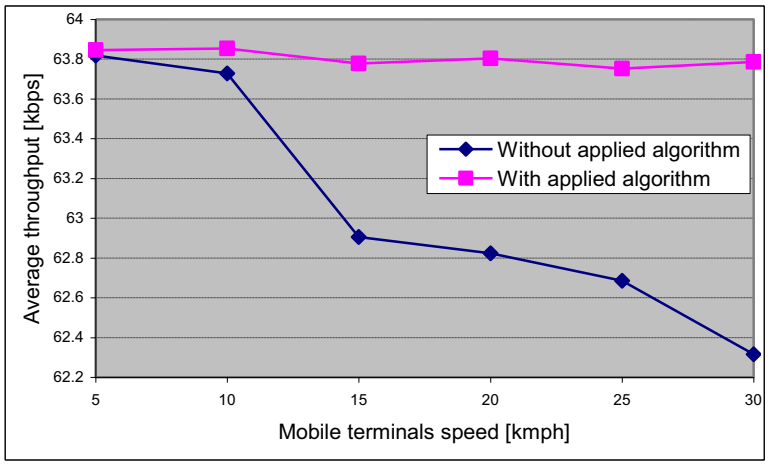


Fig. 4. Average throughput of the VoIP users with and without MS admission control algorithm.

Analyzing the results using the 802.21 simulator from NIST 24 vertical handovers are triggered between WiMAX cell and WLAN hotspot. After implementation our admission control algorithm this number of vertical handovers is decreased to only 8. It occurred only for the mobile terminal nodes that are moving with speed equal or below the threshold speed. We set the threshold speed on 10 kmph. Decision for setting this threshold speed was made because when mobile users move with higher speed than 10 kmph it is wasting of the resources and time to process vertical handover to small cells like a WLAN. In this case mobile users will be covered in a short time period. Furthermore, WLAN network is more sensitive to higher speeds of the mobile terminals. This is confirmed in Fig. 4 where average throughput of the users decreases when they move with speed above 10 kmph.

The performance study implementing our algorithm was done considering the metrics vertical handover latency expressed in milliseconds, packet loss expressed in % and average throughput expressed in kbps. The handover performance study implementing our algorithm was done considering the metric vertical handover latency. Vertical handover latency is defined as the difference between the time when the mobile node is last able to send or receive an IP packet by way of the previous WiMAX network, and the time when the mobile node is able to send or receive an IP packet through the new WLAN network.

Fig. 4 shows average throughput of the mobile users that use VoIP G.711 traffic (7^{th} to 30^{th} mobile terminal) while they are crossing WLAN network. After applying

the algorithm the users that move with speed above 10 kmph are staying in WiMAX network when they are detecting the WLAN coverage. It is obvious from the graph that if we don't apply our algorithm average throughput of the users that move with speed above 10 kmph in WLAN network degrades. Implementing the speed sensitive algorithm improves the results for average throughput of the users moving above 10 kmph. We have to put an accent on the fact that improved results for average throughput as a result of the applied algorithm for speeds above 10 kmph in the Fig. 4 are when mobile terminal is using WiMAX network. Without our applied algorithm when mobile terminals move in WLAN network with speeds from 15 to 30 kmph (from 15th to 30th user) results for average throughput are decreasing. This happens because WLAN technology is sensitive when users move with higher speeds. This fact and the improved results for average throughput justify the benefit of applying our algorithm.

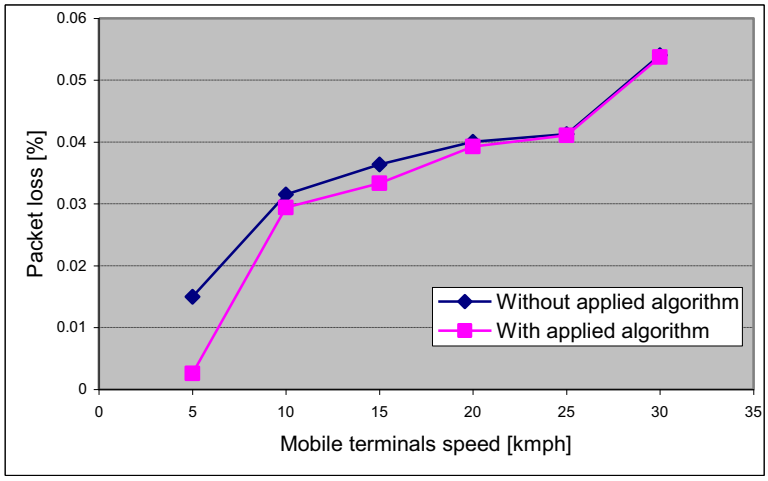


Fig. 5. Packet loss of the VoIP users with and without MS admission control algorithm in the WiMAX network.

As we can see from the analyzed results, we analyze the benefit of applied algorithm on different aspects of the simulated scenario. Packet loss in Fig. 5 is analyzed for the VoIP users only when they are attached to WiMAX network. The averaged results for all speeds of the mobile terminals are presented before and after the implementation of the mobile sensitive algorithm. The comparison results show that implemented algorithm improved the packet loss of the VoIP traffic while mobile terminals are using WiMAX network. Because the number of vertical handover processes in the specific scenario is decreased from 24 to 8 after implementing the algorithm, vertical handover duration for the users moving with 5 and 10 kmph is decreased. This leads to better packet loss performances of the users moving with 5 and 10 kmph, while they are still attached to WiMAX base station during the vertical handover procedure. Difference of the packet loss improvement with applied algorithm compared without applying the algorithm is lower at higher speeds in Fig. 5. This happens due to the fact

that the influence of the speed on packet loss is higher at higher speeds and the effect of the improvement of the packet loss results as a consequence of decreasing the number of vertical handovers is then lower.

Table 2. Average vertical handover latency for VoIP users.

Averaged results for 24 VoIP users	Without our applied algorithm	With applied algorithm
Vertical handover latency between WiMAX and WLAN [ms]	323.67	298.57

Table 2 compares the average values for vertical handover latency without and with our added MS admission control algorithm during the vertical handover process. The values are averaged for all 24 mobile terminals that use VoIP traffic before implementing our algorithm and after implementation of the algorithm the values are averaged only for the 8 mobile terminals that are moving with 5 and 10 kmph. The other 16 vertical handovers hasn't been triggered between WiMAX and WLAN after adding our algorithm, so we cannot compare them. The results in Table 2 show that implemented admission control not only decreased unnecessary vertical handovers of the users, but also decreased the vertical handover latency, which is very important for delay sensitive traffic like VoIP.

This happens due to the fact that more users in the same time must perform vertical handover process from WiMAX to WLAN. Because of this vertical handover latency for some of them will increase because of the congestion. When some users start to perform vertical handover other users are already in the process of vertical handover, and they must wait for more time to finish the procedure of disconnecting from WiMAX and connecting to WLAN coverage. When fewer users perform vertical handovers, probability of this kind of problems decreases.

5 Conclusion

The handover target selection in future heterogeneous wireless and mobile networks should be done through combination of mobility and resource management mechanisms such as admission control. That is why in the media independent handover services in [10] in the set of parameters required for performing admission control and resource reservation for the MN at the target network information about the MN speed should be added. After adding the parameter of MN speed in the RequestedResourceSet of parameters in [10], our admission control algorithm can be integrated with IEEE 802.21 technologies.

We have successfully simulated the effect of the proposed admission control algorithm for 4G and 5G heterogeneous networks using IEEE 802.21 standard on WiMAX (802.16) and WLAN (802.11) access technologies. We applied admission control algorithm that reacts on mobile node terminal speed and admits or rejects the vertical handover triggers to WLAN network from WiMAX (as in our simulation case), UMTS or some other wireless technology.

The contribution of this paper is the design of the admission control algorithm that can be easily implemented in the resource management for heterogeneous networks. Furthermore, we practically applied the algorithm in the already existing NIST simulation tool for 802.21 and practically tested with 24 VoIP users that are crossing from WiMAX to WLAN network. We proved that implementing this speed sensitive algorithm we avoided unnecessary vertical handovers to WLAN network and improved the QoS of the mobile node terminals that are using VoIP traffic expressed in vertical handover latency, throughput and packet loss which are very important metric for real time sessions.

References

1. T. Janevski, *5G Mobile Phone Concept*, IEEE CCNC 2009, Las Vegas, USA, 10-13 January 2009. DOI: 10.1109/CCNC.2009.4784727.
2. C.E. Rothenberg, A. Roos, *A Review of Policy-Based Resource and Admission Control Functions in Evolving Access and Next Generation Networks*, J Netw Syst Manage, Springer Science+Business Media, LLC 2008.
3. C.-F Wu, L.-T. Lee, H.-Y. Chang, D.-F. Tao, *A Novel Call Admission Control Policy Using Mobility Prediction and Throttle Mechanism for Supporting QoS in Wireless Cellular Networks*. Journal of Control Science and Engineering, Volume 2011, Article ID 190643, 11 pages, doi:10.1155/2011/190643.
4. J.M. Bauset, E.B. Mor, *Insensitive Call Admission Control for Wireless Multiservice Networks*, IEEE Communications Letters, 2011.
5. S. Alwakeel, A. Prasetijo, *A Policy-Based Admission Control Scheme for Voice over IP Networks*, Journal of Computer Science 5 (11): 817-821, 2009, ISSN 1549-3636.
6. R.M. Prasad, P.S. Kumar, *An Efficient Connection Admission Control Mechanism for IEEE 802.16 Networks*, American Journal of Scientific Research, ISSN 1450-223X Issue 27(2011), pp. 120-127.
7. C.S. Rao, K.C.K. Reddy, D.S. Rao, *QoS Based Adaptive Admission Controller for Next Generation Wireless Networks*, International Journal of Computer Theory and Engineering, Vol. 3, No. 5, October 2011.
8. E.Z. Tragos, G. Tsiropoulos, G.T. Karetsos, S.A. Kyriazakos, *Admission Control for QoS Support in Heterogeneous 4G Wireless Networks.* IEEE Network, May/June 2008.
9. S. Mohanty: *VEPSD: A Novel Velocity Estimation Algorithm for Next-Generation Wireless Systems*, IEEE Trans. Wireless commun., vol. 4, 2005, pp.2655-60.
10. IEEE Std 802.21-2008, *IEEE Standard for Local and Metropolitan Area Networks, Part 21: Media Independent Handover Services*, IEEE, January 2009.
11. The network simulator NS-2.29 NIST add-on, IEEE 802.21, NIST, January 2007.

Improving the Wholesales Trough Using the Data Mining Techniques

Goran Vitanov , Igor Stojanovic, Zoran Zdravev

University Goce Delcev – Shtip

goran.vitanov@gmail.com, {igor.stojanovik, zoran.zdravev}@ugd.edu.mk

Abstract. This paper describes the practical use of data mining techniques in wholesales though concrete steps of developing the distribution network. We are choosing a number of techniques for solving the problem through analyzing the specifics of the market situation, defining the business problem, and setting up its model. The cluster analysis enables us to group the data according to similarities of the market segmentation, product categories, regions and groups by turnover, while with the decision trees we are separating the big data collection into consecutive groups. Using these techniques we gain knowledge and have the opportunity to predict the future trends with high probability and on this basis to make a more precise and trustworthy business decision.

Keywords: Data Mining, Cluster, Decision trees, Wholesales.

1 Introduction

The data mining techniques are used in order to foresee a potential future result in other words to predict future possibilities and trends. There are many different automatic analysis techniques of large amount of data with multiple variables, for example: clustering, decision trees, analysis of basket sales, regression models, neural networks, genetic algorithms, hypothesis testing, and many more. In the predicting models the data is gathered, the statistic model is formulated, the prediction is completed, and the model is checked when additional information is available. The analysis combines the knowledge from business and the statistic analytical techniques in order for the user to find information from the existing data. The resulting information enables the companies to better understand the buyers, the sellers, the distributers, etc. Furthermore, this enables the user reach strategic decisions, for example where they can invest, whether to increase or decrease a portfolio of products, or whether to move onto new markets. These analyses not only show what needs to be done, but when and where.

S. Markovski, M. Gusev (Editors): ICT Innovations 2012, Web Proceedings, ISSN 1857-7288

2 Description of the market condition

Increased competition, increased product variety, constant changing of the market segmentation, the complexity of covering certain territories through engaging new commercialists, price increases, the buyers becoming more sophisticated, and the need for using distribution centers are only a portion of the complex dynamics, which is present in the field of distribution. The constant changes in the market environment from the aspect of retail segmentation, the demographic movements, and the increased choices of the buyers, are causing the distribution companies to face further challenges. To be able to survive in such environment the need for analytical, directed, and predicting approach rises. This would result in better understanding the challenges with aim to get more organized, to increase the market coverage, and to increase the profit. Today, in order to reply to the existing challenges there are multiple tools for data mining and prediction, which through the years become more sophisticated and their usage further simplified. Unlike the past, nowadays with the development of the hardware and software these technologies are increasingly becoming an integral part of every large company. This enables the company to bring a newer component in the decision system. The management today can identify the variables that have the most influence over the performance of the business and on this basis to yield better decisions.

An example company that is working on distribution of products from multiple manufacturers in Macedonian market is analyzed. The sale of certain categories of products is changed significantly and the need for more detailed analysis, from the aspect of the changes in market segmentation, is imposed. Based on the database results it has been realized that the number of buyers is changing, which brings the requirement the data to be analyzed in detail. A research is required to understand what is the reason why there parameters are changing, with the goal to increase the profit. The current trend shows that this research should be continued.

3 Information System

The company has six distributed locations with applications connected to central database. The distributed applications are covering the commercial business as (invoicing, orders, purchases, account receivable and account payable, warehouse stocks and etc.). In central location is information system for accounting and management.

Two years ago is developed data warehouse system to support reporting, controlling and planning. Separating and regrouping the information allows seeing a trend, exceptions, patterns, and relationships. This enables to analyze multidimensional data from multiple perspectives using consolidation, drill-down, slicing and dicing. The ETL (extraction, transportation, transformation and loading) solution is developed in order to load the data warehouse.

The operating system is Microsoft Server 2008 and the database system is Microsoft SQL Server 2008. For data mining the company is using Microsoft SQL Server Analysis Services, Data Mining Client for Excel and SPSS.

4 Setting up the model in order to solve the problem

Given the complexity of the problem more analysis is needed in order to get a clear picture of the market condition. The initiating meeting of a data mining project will often involve the data miner, the business expert and the data analyst. It is critical that all three come together to initiate the project. Their different understandings of the problem to be tackled all need to come together to deliver a common pathway for the data mining project. In particular, the data miner needs to understand the business perspective and understand what data is available that relates to the problem and how to get that data, and identify what data processing is required prior to modeling. To be more exact, one should answer the following questions:

- What would the market segmentation be based on the product category?
- What is the connection between a manufacturer and a market segment?
- What is the trend between a regional and a market segment?

Answering these questions lowers the risk when making the decision and it affects the future profitability of the company. In order to complete this, the following techniques of data mining are used as a part of the analysis: Association discovery, Cluster analysis and Decision trees.

4.1 Association discover

In the case of a given collective data out of which every entry contains certain number of elements, the goal of the association would be to find the connection among the elements in the set. The association rule in the data mining can be defined as: let I={i1, i2, ..., in} be a set of n binary attributes which are called: items. Let D={t1, t2, ..., tm} be a set of transactions called data base. Every transaction in D has a unique transaction ID and contains subset of items in I. The rule is established with the implication of the form: X=>Y where X, Y .. I and X .. Y = 0. The set of items X and Y are called predecessor (left side) and successor (right side) of the rule.

Using the association rule we arrive to the following results - Table 1:

Table 1. Number of found items using the association rule.

Support	Size	Itemset
2181	1	Category = TT
1861	1	Grupa Promet >10000
1778	1	Region = South
1704	1	Region = North
1662	1	Category = LKA
1661	1	Region = Исток
1495	1	Category = NKA
1372	1	Grupa Promet = 5001-10000
1369	1	Region = West

1362	1	Grupa Promet = 2001-5000
1023	1	Grupa Promet = 1000-2000
894	1	Grupa Promet = 1000<
780	1	Category = Wholesaler
602	1	Category Producer = КОНЗЕРВИРАН ЗЕЛЕНЧУК
587	1	Category Producer = КОНЗЕРВИ
581	2	Region = South, Category = TT
580	2	Region = Исток, Category = TT
579	1	Category Producer = МЛЕЧНИ ПРОИЗВОДИ
522	2	Region = North, Category = TT
498	2	Region = West, Category = TT
494	1	Category Producer = СУВОМЕСНАТО
492	2	Category = LKA, Region = South
485	1	Category Producer = ГРИЦКИ
483	2	Region = West, Grupa Promet >10000

The system recognizes 2181 buyers in the category TT, 1861 accounts in the "Group of operations" > 10000, 1778 sales in the region "south" etc. This data with further application of the association rule yields the following results:

Table 2. : Results from the association rule usage.

Probability	Importance	Rule
54%	0.22	Group turnover = 1000<, Region = East -> Category = TT
51%	0.19	Group turnover = 1000<, Region = West -> Category = TT
51%	0.19	Group turnover = 1000-2000, Region = West -> Category = TT
46%	0.17	Group turnover = 1000< -> Category = TT
44%	0.13	Group turnover = 1000-2000, Region = East -> Category = TT
43%	0.11	Group turnover = 1000-2000, Region = South -> Category = TT
42%	0.12	Group turnover = 1000-2000 -> Category = TT
42%	0.1	Group turnover = 1000<, Region = North -> Category = TT
41%	0.09	Group turnover = 1000<, Region = South -> Category = TT
40%	0.08	Group turnover = 2001-5000, Region = West -> Category = TT

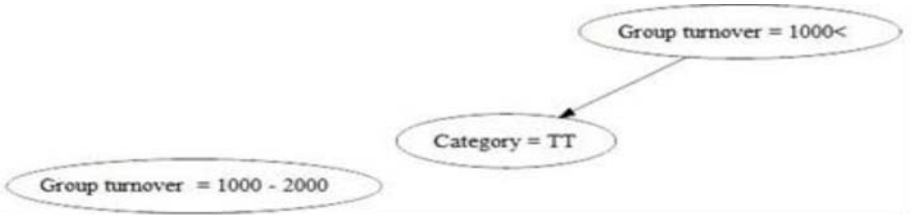


Fig. 1. : Connection strength between categories.

Based on these results it could be concluded that the operations under 1000 denars in all of the regions is connected with the TT channel of sale. The same rule can be

applied for the operations from 1000 to 2000 denars. This means that we have not deduced a separate rule for sales in certain product category, but that the size of the invoice is important. Furthermore, the TT channel during procurement prefers to be more careful of the account size rather than the product portfolio. This is also described in Figure 1.

4.2 Cluster analysis

The goal of the cluster analysis or clustering is to combine a set of objects in a group called "cluster", that way the objects in one cluster have the most similarities among themselves, rather than among other objects in a different cluster. The cluster itself is not a separate algorithm but it is a group of algorithms that enable grouping of objects according to their common attributes. Other terms used to describe clusters are: automatic classification and numerical taxonomy. In this paper the tool "Detect Categories", which is included in "Microsoft Excel", was used to detect clustering of the objects. After using this technique the following results are acquired:

Table 3. Categories created after using "Detect Categories" tool.

Category Name	Row Count
ТТ-Сувомеснато_Запад	2451
Југ-ЛКА	2600
Исток-Средства за садови	2012
НКА-Конзервиран зеленчук	2237

Category	Column	Value	Relative Importance
ТТ-Сувомеснато_Запад	Kategorija	ТТ	41
ТТ-Сувомеснато_Запад	Kategorija Proizvoditel	СУВОМЕСНАТО	35
ТТ-Сувомеснато_Запад	Region	Запад	23
ТТ-Сувомеснато_Запад	Kategorija Proizvoditel	ПИЈАЛОЦИ	10
ТТ-Сувомеснато_Запад	Kategorija Proizvoditel	КЕЧАП	10
ТТ-Сувомеснато_Запад	Kategorija Proizvoditel	ЛЕБ	7
ТТ-Сувомеснато_Запад	Kategorija Proizvoditel	ФЛИПС	7
ТТ-Сувомеснато_Запад	Region	Север	6
ТТ-Сувомеснато_Запад	Kategorija Proizvoditel	АЛКОХОЛНИ ПИЈАЛОЦИ	2
ТТ-Сувомеснато_Запад	Kategorija Proizvoditel	МАСТИКИ	1
Југ-ЛКА	Region	Југ	77
Југ-ЛКА	Kategorija	LKA	74
Југ-ЛКА	Kategorija Proizvoditel	ДЕЗОДЕРАНСИ	18
Југ-ЛКА	Kategorija Proizvoditel	ГРИЦКИ	5
Југ-ЛКА	Kategorija Proizvoditel	ЧИПС	2
Југ-ЛКА	Kategorija Proizvoditel	МЛЕЧНИ ПРОИЗВОДИ	2

Југ-ЛКА	Kategorija Proizvoditel	БОЈА ЗА КОСА	1
Југ-ЛКА	Kategorija Proizvoditel	ДЕТЕРГЕНТИ	1
Југ-ЛКА	Kategorija Proizvoditel	КОНЗЕРВИ	1
Исток-Средства за садови	Region	Исток	100
Исток-Средства за садови	Kategorija Proizvoditel	СРЕДСТВА ЗА САДОВИ	13
Исток-Средства за садови	Kategorija	ТТ	2
Исток-Средства за садови	Kategorija Proizvoditel	ТОАЛЕТНА ХАРТИЈА	1
НКА-Конзервиран зеленчук	Kategorija	NKA	100
НКА-Конзервиран зеленчук	Kategorija Proizvoditel	КОНЗЕР. ЗЕЛЕНЧУК	13
НКА-Конзервиран зеленчук	Kategorija Proizvoditel	ЧИПС	2
НКА-Конзервиран зеленчук	Kategorija Proizvoditel	ДЕТЕРГЕНТИ	2
НКА-Конзервиран зеленчук	Kategorija	Wholesaler	1
НКА-Конзервиран зеленчук	Kategorija Proizvoditel	КОНЗЕРВИ	1
НКА-Конзервиран зеленчук	Region	Север	1

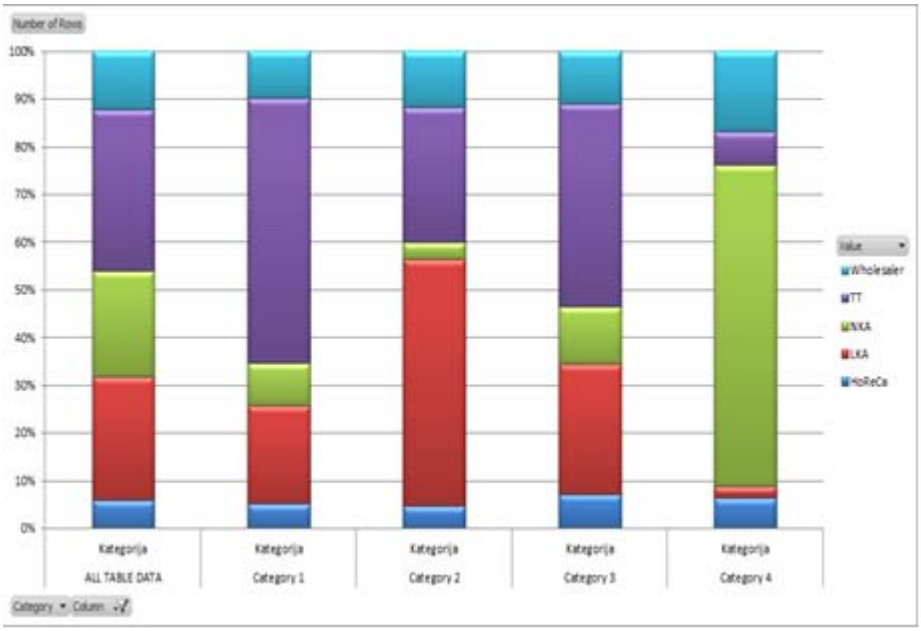


Fig. 2. : Created Clusters

Four categories are formed, which give us a precise picture of the market. Hence the TT channel is important for sales of cold cuts, especially in the western region. The western region is also important for sales of ketchup and beverages. That means if

changes of the distribution network are planned it is important to consider the sales of these groups of products in this region. The southern region is a region in which the LKA buyers are more important. From this we conclude that the investments need to be bigger in this channel from the viewpoint of the marketing activities and the strengthening of the sales force. The east region is important for sale of soaps and detergent and therefore there is a need to support this type of products. The HKA cluster and sales of canned foods gives us direction for increasing the activities of these buyers connected with the sales of canned products. This is an expected trend because NKA is mainly concentrated in the larger towns and the buyers are mainly employed people. By applying the technique Microsoft Cluster on the same data, 9 clusters are created, which are shown on Figure 2 . From the shown data it can be seen that cluster 1 represents the southern region that no groups of products stand out, that in the sales channels TT LKA, is dominant, that the participation of NKA is the smallest cluster 2 represents the north region of NKA sales without important participation of any product, etc.

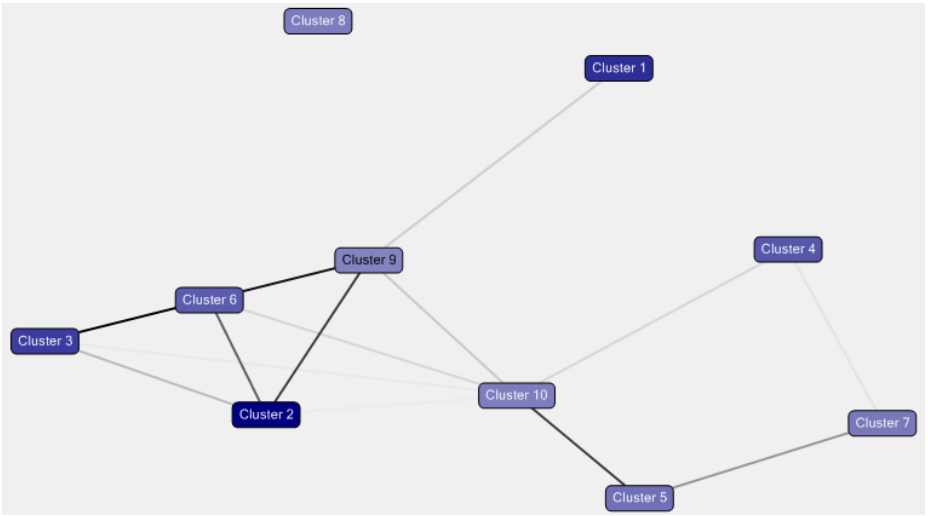


Fig. 3. : Connections among the created clusters

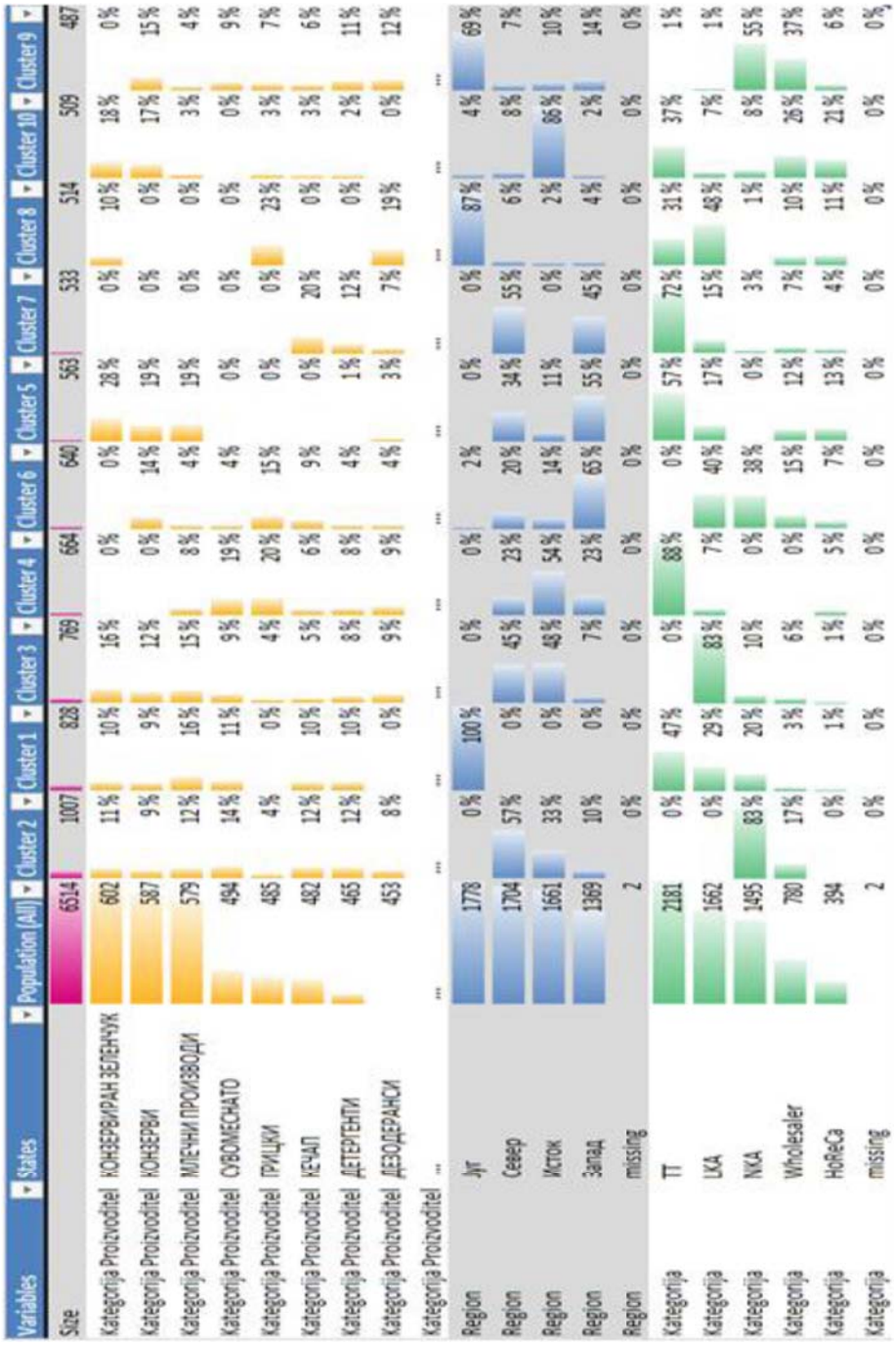


Fig. 4. : Clusters created with Microsoft Cluster Technique

4.3 Decision trees

The decision tree is a powerful and popular technique that can be used for classification and prediction. The attractiveness of the methods based on decision tree is mostly based on the fact that they are rules. These rules can be represented with SQL expressions in order for information from a specific category to be obtained. Because the decision trees combine the research of information and modeling, they are a powerful first step in the process of modeling even if the final model is obtained with some other technique. The decision tree is a structure that can be used to divide a large collection of data into a number of smaller sets of data by using simple decision rules. With each new division, the members of the obtained sets become more similar with each other. The model of the decision tree consists of sets and rules for division of a heterogeneous population into a number of smaller homogenous group with respect to some given variable. In our case, the SPSS tool is used and analysis is done on a category of products as a dependent variable, the sales channel (category) as an independent variable and value as a dependent variable that is influenced by everything.

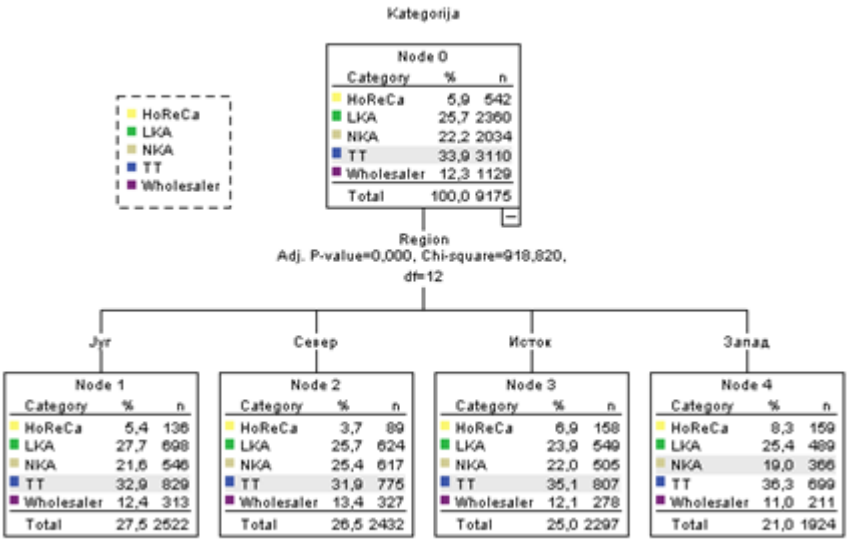


Fig. 5. : Decision tree in SPSS

Through the decision tree it can be seen that there is a regular division of the sales by region. The category buyers are differently arranged by region, where in the north and south region there are changes that benefit NKA and LKA. Based on the derived trend, it can be assumed that the same situation will be present in the other region in the future. This conclusion is based on the rapid development of LKA buyers, who are both in the segment of organized trade and NKA.

5 Conclusion

The giant competition, which in some respect is caused by the globalization, forces every company to start using methods for data mining with the goal to become competitive and to respond correctly to the market challenges. The modern management in no means can let decision making without prior analysis of the result obtained with data mining. To enable this, it is important to create a system that can respond to the current flows. The system will contain hardware, software and hiring computer scientists that will build and constantly improve it. It is important to surpass the classical systems of information in the form of printed reports of the type cards or search of buyers, top buyers, top products and so on. The aim is to implement a system which with the use of the techniques of data mining will enable companies to obtain knowledge and opportunities of prediction of the flows in the future with good probability.

6 Reference

1. Q. M. Mark Whitehorn and M. Keith Burns, "Microsoft Corporation," 07 2008. [Online]. Available: http://technet.microsoft.com/en-us/library/cc719165(v=sql.100).aspx. [Accessed 27 03 2012].
2. G. S. L. Michael J.A. Berry, Data Mining Techniques For Marketing, Sales, and Customer Relationship Management, Wiley Publishing, Inc., Indianapolis, Indiana, 2004.
3. "The Transition of Data into Wisdom," 20 03 2012. [Online]. Available: http://www.information-management.com/news/2784-1.html.
4. D. C. A. M. Gene Bellinger, "Data, Information, Knowledge, and Wisdom," [Online]. Available: http://www.systems-thinking.org/dikw/dikw.htm. [Accessed 16 03 2012].
5. Dijcks, Jean Pierre, "Oracle: Big Data for the Enterprise," Oracle, (2012).
6. OLTP vs. OLAP," [Online]. Available: http://datawarehouse4u.info/OLTP-vs-OLAP.html. [Accessed 12 03 201].
7. Brian Larson, Delivering business intelegence with Microsoft SQL Server 2008, Copyright © 2009 by The McGraw-Hill Companies.
8. Javier Torrenteras and Carlos Martinez , SSAS Cube Exploration: Digging Through the Details with Drillthrough..
9. Ian H. Witten, Eibe Frank, Mark A. Hall, Data Mining Practical Machine Learning Tools and Techniques Third Edition, Copyright © 2011 Elsevier Inc.
10. Art Tennick, Practical DMX Queries for Microsoft® SQL Server® Analysis Services 2008, Copyright © 2011 by The McGraw-Hill Companies.

Performance Analysis of a Connection Fault-Tolerant Model for Distributed Transaction Processing in Mobile Computing Environment

Tome Dimovski and Pece Mitrevski

St. Clement Ohridski University, Faculty of Technical Sciences, Ivo Lola Ribar bb
7000 Bitola, Republic of Macedonia
{tome.dimovski,pece.mitrevski}@uklo.edu.mk

Abstract. Mobile embedded systems increasingly use transactions for applications like mobile inventory, mobile commerce or commercial applications. Yet, many issues are challenging and need to be resolved before enabling mobile devices to take part in distributed computing. Mobile environment limitations make it harder to design appropriate and efficient commit protocols. There are a handful of protocols for transaction execution in distributed mobile environment, but almost all consider a limited number of communication models. In this paper, we evaluate the performance of a Connection Fault-Tolerant Model, comparing the results in several deferent scenarios, as well as its contribution to the overall mobile transaction commit rate. Our simulation-based performance analysis determines the impact of (i) ad-hoc communication between mobile hosts and (ii) the implementation of a decision algorithm, to the mobile transaction commit rate. We also determine the connection timeout values that contribute the most for a high ad-hoc communication impact on mobile transaction commit rate.

Keywords: Distributed mobile transactions, ad-hoc communication, Decision Algorithm.

1 Introduction

The increasing emergence of mobile devices contributes to rapid progress in wireless technologies. Mobile devices interacting with fixed devices can support applications such as e-mail, mobile commerce (m-commerce), mobile inventory, etc. But there are many issues that are challenging and need to be resolved before enabling mobile devices to take part in distributed computing. For distributed systems, a transaction is a set of operations that fulfill the following condition: either all operations are permanently performed, or none of them are visible to other operations (known as the *atomicity* property). In the execution of transactions the key issue is the protocol that ensures atomicity.

The mobile environment is comprised of mobile devices with limited resources like processing, storage, energy capacity and continuously varying properties of the

S. Markovski, M. Gusev (Editors): ICT Innovations 2012, Web Proceedings, ISSN 1857-7288

wireless channel. Wireless communication induces much lower bandwidth, higher latency, error rates and much higher costs. This increases the time needed for mobile hosts to execute transactions and can even lead to execution failure. *Mobile hosts (MHs)* are highly vulnerable devices because they are easily damaged or lost. MHs naturally show frequent and random network disconnections. These limitations and characteristics of the mobile environment make it harder to design appropriate and efficient commit protocols. A protocol that aborts the transaction, each time the MH disconnects from the network, is not suitable for mobile environments because it is part of the normal mode of operation. In other words, disconnections need to be tolerated by the protocol.

The *two-phase commit (2PC)* protocol [1] that allows the involved parties to agree on a common decision to commit or abort the transaction even in the presence of failures is the most commonly used protocol for fixed networks but is unsuitable for mobile environments [2]. There are several other protocols for transaction execution in distributed mobile environment [3-13], but almost all consider limited number of communication models. A Connection Fault-Tolerant Model (CFT) that shows resilience to connection failures of the mobile devices has been presented in [14]. It differs from other infrastructure based protocols [3-5], [7], in the fact that beside the standard communication between mobile hosts and the fixed network, it supports (i) ad-hoc communication between mobile hosts and (ii) introduces a decision algorithm that is responsible for decision making on behalf of a mobile host in a special case, when neither standard, nor ad-hoc communication is possible.

The main contribution of this paper is the simulation-based performance analysis of the CFT model for mobile distributed transaction processing. We evaluate the performance of the CFT model, comparing the results in several deferent scenarios, as well as its contribution to the overall mobile transaction commit rate. We also determine the optimum connection timeout values that contribute the most for the high ad-hoc communication impact on transaction commit rate.

The paper is organized as follows. Section 2 gives a survey of related work. In Section 3 we present the model of the mobile environment, and in Section 4 the transaction model. Description of the Connection Fault-Tolerant Model is given in Section 5, whereas in Section 6 we present simulation results and their analysis. Section 7 discusses conclusions.

2 Related work

Transaction Commit on Timeout (TCOT) protocol [3] is based on a timeout approach for Mobile Database Systems, which can be universally used to reach a final transaction termination decision in any message oriented system. This protocol limits the amount of communication between the participants in the execution of the protocol. It decreases the number of wireless messages during execution, and does not consider mobile hosts as active participants in the execution of transactions.

The basic idea of the *Two-Phase Commit Protocol for Mobile Wireless Environment (M-2PC)* [4] is to adapt the 2PC protocol for mobile systems with distributed transactions. Mobile hosts are active participants in the execution of a transaction and they send confirmation (that the work is done) to the agent or to the

fixed device, in order to save energy. This model requires simultaneous connection of all mobile participants at the beginning of a transaction. This protocol does not provide adequate management of mobility and failures caused by network disconnection, nor does it provide a mechanism to control the competitiveness of distributed transactions.

Fault-Tolerant Pre-Phase Transaction Commit (FT-PPTC) protocol [5] provides mechanisms for dealing with disturbances in the system in mobile environment. The protocol supports heterogeneous mobile databases. FT-PPTC implements distributed transaction in two phases: pre-phase, one which is covering the mobile hosts and the main phase which refers to the fixed part of the network. Mobile hosts are active participants in the execution of a transaction. No mechanisms are developed for competition in mobile distributed transactions. FT-PPTC does not provide adequate management of mobility, as well, because when mobile hosts are disconnected from the fixed network for a long time, they can block resources on the fixed participants. This, in turn, leads to an increased number of mobile transaction aborts.

In the *concurrency control without locking* approach [6], a concurrency control mechanism in mobile environment is proposed, by introducing the concept of *Absolute Validity Interval (AVI)* which is a time period in which the data item is said to be valid. It does not use the traditional locking mechanism. The new mechanism provides reading same available data item by multiple mobile hosts. If the data item is updated by any mobile host, the mobile hosts which have already read the same value must be invalidated. It reduces the waiting time for execution of a transaction and resources are not unnecessary locked.

3 Model of the mobile environment

In this paper we consider a system model for the mobile distributed environment consisting of a set of *mobile hosts (MHs)* and a set of *fixed hosts (FHs),* presented in Fig. 1. The model has two main parts: the fixed part of the network and a mobile part of the network. Communication between the two is conveyed via *Mobile Support Stations (MSS)*, which are connected to the fixed part of the network via wired links. MHs can cross the border between two different geographical areas covered by different MSSs.

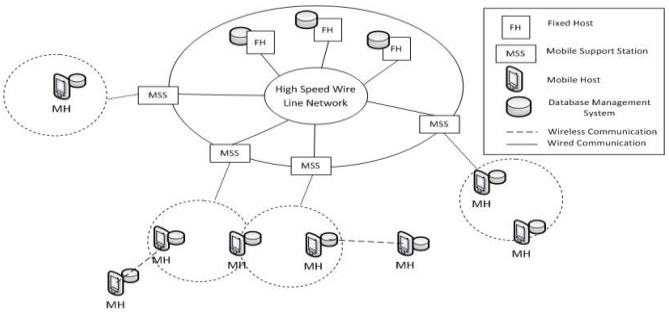


Fig. 1. Communication model in the mobile environment

In the considered system model, first, MHs can communicate with the FHs through a MSS via wireless channels only when they are located within the MSS coverage area. Second, the MHs can ad-hoc communicate with neighboring MHs via wireless channels. When MHs enter a geographical area that is out of coverage of any MSS, in order to access database servers in the fixed network, they may connect through a neighboring MH which is in the covering area of any MSS.

In brief, we consider a mobile distributed environment where (i) MHs can communicate with each other (and/or with FHs) through MSS and, in addition, (ii) MHs can ad-hoc communicate with neighboring MHs in order to reach the fixed part of the network. We assume that database servers are installed on each FH, and each MH has a mobile database server installed.

4 Transaction model

A distributed transaction where at least one MH participates is called a *Mobile Transaction (MT).* We identify a MH where a transaction is issued as a *Home-MH (H-MH).* Participating MHs and FHs in the execution of a mobile transaction are called *participant MHs (Part-MH)* and *participant FHs (Part-FH).*

We assume the existence of a *Coordinator (CO)* which is responsible for coordinating the execution of the corresponding transaction. The CO is responsible for storing information concerning the state of the transaction execution. Based on the information collected from the participants of the transaction, the CO takes the decision to commit or abort the transaction and informs all the participants about its decision. The CO should be executed on a fixed host or hosts. This means that logs will be kept more safely.

5 Connection Fault-Tolerant Model

5.1 Overview

Some of the most frequent failures in mobile environments are communication failures. When MHs are in motion, they may exit the geographical area that is covered by some MSS and the resources of the fixed participants may potentially be blocked for an undefined period of time. If MHs do not reestablish connection with any MSS the transaction is aborted.

To minimize the number of mobile transaction aborts by tolerating failures caused by network disconnections and reduce the resource blocking times of fixed participants, we propose a CFT model for distributed transaction processing in mobile computing environment. The CFT model ensures the atomicity property.

The CFT model supports two communication protocols and a decision algorithm, as well:

1. The first protocol is a *Standard communication protocol* when MHs can directly connect to the fixed part of the network through MSSs.

2. The second protocol is an *Ad-hoc communication protocol* when MHs cannot directly connect to the fixed part of the network through any MSS. With this protocol, MHs can ad-hoc communicate with neighboring MHs in order to reach the fixed part of the network.

In the Standard communication protocol, similar as in [15], to minimize the use of the wireless communication and conservation of the resources of MHs, to each MH we assign a *Mobile Host Agent (MH-Ag)* that we add to the fixed network. We assume that in the execution of a transaction MH-Ag is representing the MH in the fixed network and it acts as an intermediator between the MH and the transaction CO. All communications between MH and CO go through the MH-Ag. The MH-Ag is responsible for storing all the information related to the states of all MTs involving the MH. In the fixed network, a server or servers can be designated, where MH-Ag is created for each participating MH.

The second (Ad-hoc) communication protocol is when MHs cannot directly connect to the fixed network, or MH-Ag cannot directly communicate with its MH through any MSS. In that case, they try to connect via ad-hoc communication with any neighboring MH which is in the covering area of any MSS. To allow this, we assign a MH-*Relay Agent (MH-RAg)* to each a MH. The *MH-RAg* is responsible for ensuring relay wireless link between neighboring MHs. This means that MH which is out of the coverage area can connect to its MH-Ag in the fixed network via *MH-RAg* of the neighboring MH which is in coverage area of any MSS.

In the CFT model we define an additional function of a MH-Ag that we call a *Decision Algorithm (DAlg)*. DAlg is used during the execution of a transaction when MH-Ag cannot directly or ad-hoc communicates with its MH for a defined period of time. DAlg's task is to check if *Transaction Processing Fragment (TPF)* function is **WRITE** (insert/update/delete) or **READ**. If TPF function is **WRITE,** DAlg saves the TPF in a FIFO (First-In First-Out) queue list and makes a decision for the MH to send "Yes" vote to the transaction CO. When the connection between MH and the corresponding MH-Ag is reestablished, MH-Ag's first task is to send all saved TPFs to the corresponding MH. If TPF function is **READ**, DAlg will wait for connection reestablishment between MH and the corresponding MH-Ag, for a defined period of time. If the connection is not reestablished in the specified time period, DAlg makes a decision for the MH to send "No" vote to the transaction coordinator.

5.2 Connection Fault-Tolerant Model operation

In this section we make a short review of the operation of the CFT model. Fig. 2 illustrates the execution of a mobile transaction for the proposed model.

If H-MH is connected to the fixed network through some MSS, it initiates a mobile transaction by sending TPF to the transaction CO through its corresponding MH-Ag which acts as an intermediator between CO and MH.

Transaction CO computes the *Execution timeout (Et)*, which is a time limit for all participants to complete the execution of the TPFs and send a vote to CO. After that, CO sends Et and TPFs to all Part-FHs and MH-Ags that represent the Part-MHs in the fixed network, asks them to PREPARE to commit the transaction, and enters the wait state. Subsequently, every MH-Ag initiates *Connection timeout (Ct)*, which is a time

limit for MH-Ag to establish connection with its MH, and try to send Et and TPFs to its MH through standard or ad-hoc communication protocol. If MH-Ag cannot establish connection with its MH before Ct expires, it activates the DAlg.

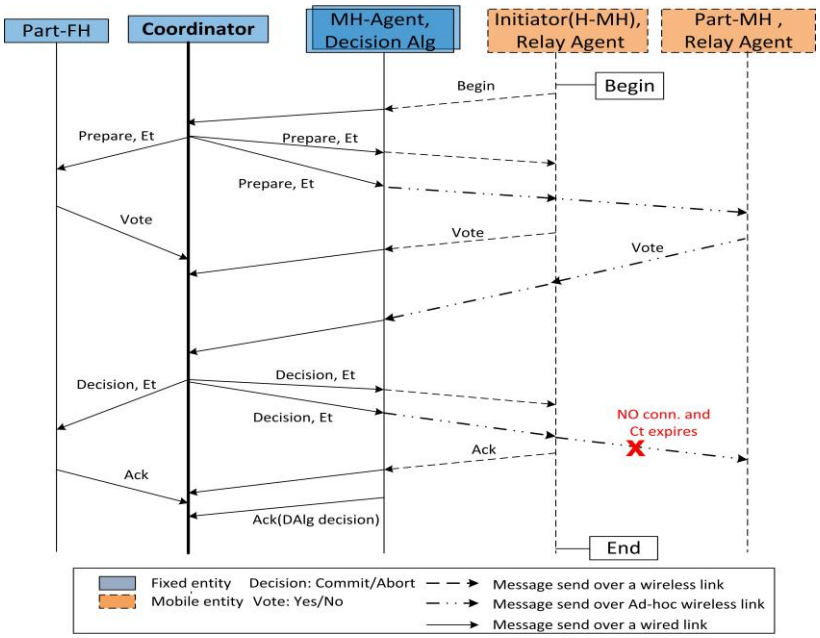


Fig. 2. Execution of a mobile transaction in mobile environment

DAlg checks if the *TPF* function is **WRITE** or **READ**. In the case of **WRITE,** DAlg saves the TPF in a FIFO queue list and makes a decision for the MH to send "Yes" VOTE to the transaction CO. As already mentioned before, if the TPF function is **READ**, DAlg makes a decision for the MH to send "No" VOTE to the transaction CO. If MH-Ag establishes connection with its MH before Ct expires, it sends TPF to its MH and resets Ct.

When the participants receive the PREPARE message, they check if they could commit the transaction. If so, and if MH establishes connection with MH-Ag before Ct expires, MH sends "Yes" VOTE to CO through its corresponding MH-Ag via standard or ad-hoc communication protocol. If Ct expires, MH-Ag activates DAlg.

After the CO has received VOTE from every participant, it decides whether to COMMIT or ABORT the transaction. If, for any reason, even one of the participants votes "No" or Et expires, the CO decides to ABORT the transaction and sends "Abort" message to all participants. Otherwise, if all the received votes are "Yes" and Et is not expired, the CO decides to COMMIT the transaction and sends "Commit" message, with reset Et to all participants. The participants need to ACKNOWLEDGE the CO's decision before the reset Et expires.

6 Simulation results and analysis

In our simulation-based performance analysis of the Connection Fault-Tolerant Model, we focus on mobile transaction commit rate performance metric. For the simulation analysis, we used SimPy [16], a process-based discrete-event simulation package based on standard Python programming language [17]. A simulation run is set to simulate 10 hours. Transactions are generated with exponentially distributed interarrival times, with an average of 30 seconds. We assume that all transactions are of similar length, but experience different connection conditions. The number and the nature of Part-MHs and Part-FHs are randomly selected in order to model arbitrary heterogeneity. Table 1 summarizes our simulation parameters.

Table 1. Simulation settings

Parameter	Value
Number of Part-MHs	3-5
Fragment execution time (Part-MH)	[0.3-0.7]s
Fragment execution time (Part-FH)	[0.1-0.3]s
Transmission delay (wireless link)	[0.2-1.0]s
Transmission delay (wireless ad-hoc link)	[0.4-2.0]s
Transmission delay (wired link)	[0.01-0.03]s
Disconnection Rate	[0 – 95]%
Ad-hoc support	[10 – 90]%
Distributed transaction WRITE function	[10 – 90]%

Hence, disconnection rate is defined as the ratio of the time when the participating MH is disconnected from the fixed network, against the total simulation time. Ad-hoc support is the ratio of the time when ad-hoc communication is available between MHs, against the total simulation time. It is hard to quantify the level of ad-hoc support between MHs in mobile distributed environment. In some parts of the wireless network ad-hoc support can be lower compared to other. For that reason, in our simulation we define three groups that represent different parts of the wireless network with different levels of ad-hoc support (there is a fundamental relationship between node density and delay in wireless ad-hoc networks with unreliable links). Every MH in the wireless network is a member of one of the defined groups.

As we mentioned before in Section 5, the CFT model supports two communication protocols – *Standard* and *Ad-hoc*, and introduces *Decision Algorithm* in order to minimize the number of mobile transaction aborts by tolerating failures caused by network disconnections. Considering several different scenarios with- or without support of the DAlg, our simulation-based performance analysis determines the CFT model contribution to the overall mobile transaction commit rate.

For the maximum ad-hoc communication contribution to the mobile transaction commit rate in the CFT model, we need to determine the values of the Connection timeout, which will allow ad-hoc communication before activating the DAlg. Fig. 3 shows the mobile transaction commit rate against different connection timeouts and different ad-hoc support values. Execution timeout is set to UNLIMITED. We assume that the functions of all mobile transactions are READ, which means that if DALg activates, the transaction will be aborted. From the chart in Fig. 3 can be concluded

that for any level of ad-hoc support, commit rate increment is more evident for increments of connection timeout up to 2.4*s*. After that point, the commit rate only slightly increases when the connection timeout time rises. It means that connection timeout value of 2.4*s* is the *optimum value* for maximum ad-hoc communication contribution on mobile transaction commit rate in the CFT model.

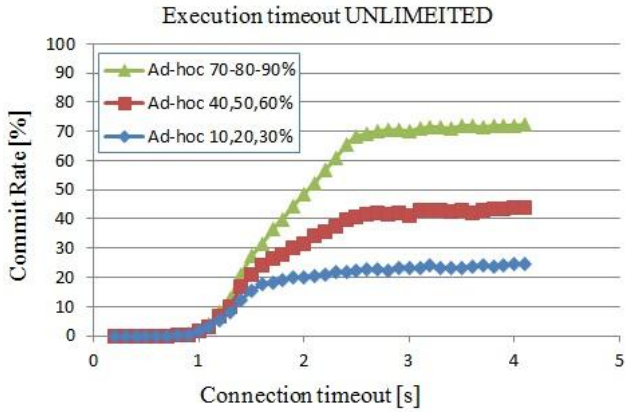


Fig. 3. Impact of Connection timeout on commit rate

To determine the ad-hoc communication contribution and DAlg contribution over transaction commit rate in the CFT model, we performed numerous simulations whose results are shown in Figs. 4-7. Figs. 4 and 6 show the mobile transaction commit rate against different disconnection rates and different percentages of mobile transactions whose function is WRITE. It is evident that higher level of ad-hoc support leads to higher commit rate. It is also evident that the mobile transaction commit rate is higher if the number of mobile transactions, whose function is WRITE, is higher.

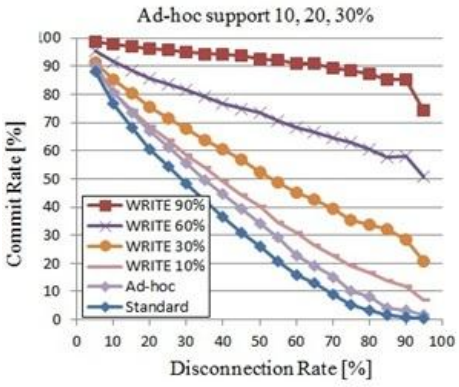


Fig. 4. CFT Model contribution over commit rate

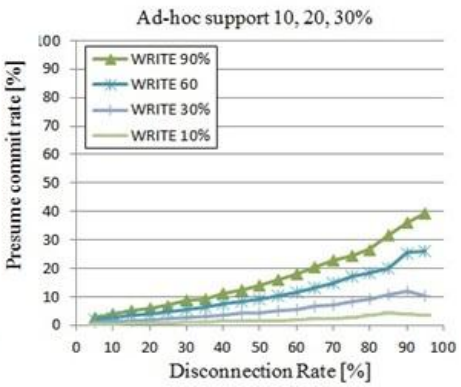


Fig. 5. DAlg contribution over commit rate

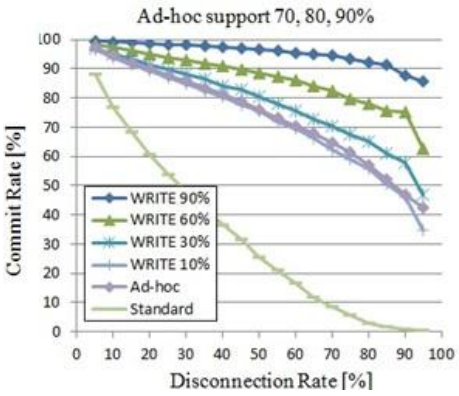


Fig. 6. CFT Model contribution over commit rate

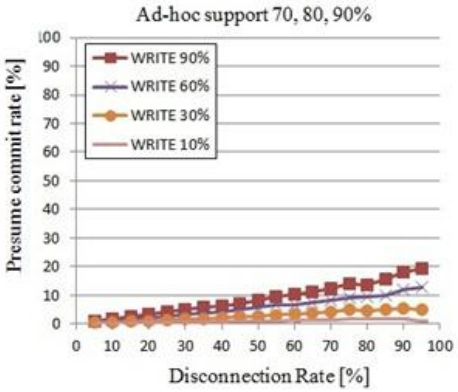


Fig. 7. DAlg contribution over commit rate

Figs. 5 and 7 show mobile transaction "presumed commit" rate against different disconnection rates and different percentages of mobile transactions whose function is WRITE ("presumed commit" rate is the percentage of committed mobile transactions by DAlg). From the charts it can be inferred that if the level of ad-hoc support is lower (Fig. 5), the impact of DAlg in the CFT Model on the mobile transaction commit rate is higher. The other way around, if the level of ad-hoc support is higher (Fig. 7), the impact of DAlg on the mobile transaction commit rate is lower.

7 Conclusion

In this paper we made a detailed review of the operation of a Connection Fault-Tolerant Model for mobile transaction processing. Our simulation-based performance analysis determines the contribution of (i) the existence of ad-hoc communication and (ii) the implementation of a decision algorithm to the mobile transaction commit rate. We also determined the connection timeout values that contribute the most for a high ad-hoc communication impact on transaction commit rate (before DAlg's activation). Simulation results show that the contribution of ad-hoc communication and DAlg on the transaction commit rate is counter proportional: if the level of ad-hoc support is lower, the impact of DAlg in the CFT Model on the mobile transaction commit rate is higher and vice versa. Conjointly, they both increase the transaction commit rate.

In our future work we plan to expand the parameters that the decision algorithm uses for decision making on behalf of the MH, as well as to employ the class of Deterministic and Stochastic Petri Nets (DSPNs) for modeling and analysis of the proposed model (deterministic transitions – to capture the timeout mechanisms, and exponential transitions – to capture the interarrival times of transactions, as well as the stochastic nature of random network disconnections), in order to evaluate a range of availability, reliability, performance and performability measures.

References

1. Gray, J.: Notes on Data Base Operating Systems. In: Operating Systems, An Advanced Course, pp. 393–481 (1978)
2. Santos, N., Ferreira, P.: Making Distributed Transactions Resilient to Intermittent Network Connections. In: Proc. of the 2006 International Symposium on World of Wireless, Mobile and Multimedia Networks, pp. 598 – 602. IEEE Computer Society, Washington (2006)
3. Kumar, V.: A Timeout-Based Mobile Transaction Commitment Protocol. In: Proc. of the East-European Conference on Advances in Databases and Information Systems, pp. 339–345 (2000)
4. Nouali, N., Doucet, A., Drias, H.: A Two-Phase Commit Protocol for Mobile Wireless Environment. In: Proc. of the 16th Australasian Database Conference, pp. 135–143 (2005)
5. Ayari, B., Khelil, A., Suri, N.: FT-PPTC: An efficient and fault-tolerant commit protocol for mobile environments. In: Proc. of SRDS, pp. 96–105 (2006)
6. Moiz, S.A., Nizamudin, M.K.: Concurrency Control without Locking in Mobile Environments. In: Proc. of the First International Conference on Emerging Trends in Engineering and Technology, pp. 1336-1339. Nagpur, Maharashtra (2008)
7. Ayari, B., Khelil, A.,Suri, N.: On the design of perturbation-resilient atomic commit protocols for mobile transactions. J. ACM Transactions on Computer Systems, 29 (2011)
8. Miraclin Joyce Pamila, J.C., Thanushkodi, K.: Framework for transaction management in mobile computing environment. J. ICGST-CNIR. 9, 19--24 (2009)
9. Ravimaran, S., Maluk Mohamed, M. A.: An improved kangaroo transaction model using surrogate objects for distributed mobile system. In: 10th ACM International Workshop on Data Engineering for Wireless and Mobile Access. ACM (2011)
10. Hu, S., Muthusamy, V., Li, G., Jacobsen, H.: Transactional Mobility in Distributed Content-Based Publish/Subscribe Systems. In: 29th IEEE International Conference on Distributed Computing Systems. IEEE (2009)
11. Li, G., Yang, B., Chen, J.: Efficient optimistic concurrency control for mobile real-time transactions in a wireless data broadcast environment. In: Proc. of the 11th IEEE International Conference on Embedded and Real-Time Computing Systems and Applications, pp. 443 – 446. IEEE Computer Society Press, Washington (2005)
12. Salman, M., Lakshmi , R.: Single Lock Manager Approach for Achieving Concurrency in Mobile Environments. In: Proc. of 14th IEEE International Conference on High Performance Computing, Springer (2007)
13. Hien, N., Mads, N.: A transaction framework for mobile data sharing services. In: Proc. of the 2nd International conference on Mobile Ubiquitous Computing, Systems, Services, and Technologies, pp.109—115. IEEE (2008)
14. Dimovski, T., Mitrevski, P.: Connection Fault-Tolerant Model for distributed transaction processing in mobile computing environment. In: Proc. of the 33rd International Conference on Information Technology Interfaces, pp.145—150. IEEE Conference Publications (2011)
15. Xiang, L., Yue-long, Z., Song-qiao, C., Xiao-li, Y.: Scheduling Transactions in mobile distributed real-time database systems. Journal of Central South University of Technology, pp. 545-551 (2008)
16. SimPy simulation package, http://simpy.sourceforge.net
17. Python Programing Language, http://www.python.org

E-Lab: Web Based System for Automatic Assessment of Programming Problems

Tomche Delev[1] and Dejan Gjorgjevikj[1]

Faculty of Computer Science and Engineering
{tomche.delev,dejan.gjorgjevikj}@finki.ukim.mk

Abstract. *E-Lab* is a system developed at Faculty of Computer Science and Engineering for solving and auto-grading programming problems from introduction to programming courses. The main goal is to simplify and improve the organization and the process of solving programming problems from large group of students in dedicated computer labs using centralized server. All the work from the students is done in a web browser using a web-based code editor and everything is stored, compiled and executed on the server. The system keeps records of all problem attempts from identified students which are used as attendance records. All the problems and solutions are under version control system (Git). The platform supports different types of problems in several programming languages (C, C++, Java) and it's designed to be easily extended.

Keywords: Online submission, programming languages, automated assessment

1 Introduction

Programming is one of the essential practical skills taught at introduction level courses in computer science curriculums. Mastering this skill can improve students' chances in finding fair job and developing successful career. The rewarding career and the constant raising of the market for programmers, makes the computer science programs very popular among high school students. This, results in larger number of students at the introductory classes with several hundred students enrolled.

Programming is not an easy skill to develop. By some studies [11], mastering this skill requires up to ten years. As easy as it seems, teaching basic programming raises challenges to the academic stuff and good organisation with the right tools is required to tackle these challenges. Similarly to other practical skills, good strategy for learning programming involves great amount of time actually doing it. For introduction level courses involving some kind of programming, this translates to solving a lot of basic algorithmic examples. To incorporate this practice in the courses, the students organized in smaller groups are required to spend considerable amount of time solving this kind of programming problems organized in problem sets by the topic and held in dedicated computer labs.

S. Markovski, M. Gusev (Editors): ICT Innovations 2012, Web Proceedings, ISSN 1857-7288

Several hundred students working on a set of problems every week, produces thousands attempted solutions in form of source code that should be examined and graded. In the current environment students work on PC workstations using simple text editor or some kind of IDE for the programming language they use. They save, compile and execute the solutions on their local machines. After they finish, no records of their work is stored on server repository, so there is no possibility for the instructors to examine and grade their solutions afterwards. The time limit for each group of students is in the range of 90 to 120 minutes, and the instructors usually have only up to 30 minutes to examine, test and grade all the students of the group assigned to them (the group they are tutoring) which is usually around 20, each with solutions for several problems. These settings makes almost impossible for the instructors to quality assess the students work, or if possible makes it a very dubious task.

The nature of the programming problems in great part of the introduction programming courses is algorithmic. This makes it possible to develop a fairly simple platform for creating problems, test cases and system for automatic assessment and grading the solutions. Most algorithmic problems can be designed to take some input from the standard input in some prescribed format, apply some simple algorithm, and finally produce an output in some prescribed format and print it on the standard output. Having this kind of problems we can take the executable of the program, feed the test input and then compare and verify for correctness on the test output. This process which is widely used in many competitive programming systems should emphasize the importance in programming to have a working solution, instead of only writing a code that some times doesn't compile.

The E-Lab system is developed with several goals. If the first goal was to improve the organization and implementation of the programming exercises, other important goals are the motivation of the students and the continuous feedback they will have using this platform. With E-Lab we want to shift the role of the instructors from teachers and graders to motivators, which are shown to give better effects in teaching programming [7].

2 Related Work

Systems that automatically assess programming assignments have been designed and used for more than forty years. In [5] authors review a number of the influential systems for automatic test-based assessment of programming assignments. These systems are broadly categorized according to age in three generations.

The first generation or early assessment systems were those originating from the time when programming was done using punch cards and the evaluation was done by executing programs and manually evaluating the output. Some of these early systems had specially designed programs to compare the output of the execution to some predefined output.

The second generation or the tool oriented assessment systems are developed using pre-existing tool sets and utilities supplied with the operating system or

programming environment. One notable example of these systems is the BOSS system originated at the University of Warwick in the UK [8] which, in his last development cycles, has become an assessment management system. Other example is the Scheme-Robo project [9] which has been supplemented by a graphical user interface and an algorithm-animation component.

The third-generation assessment systems are characterized by using the latest developments in web technology and adopt advanced testing approaches. Previously mentioned system BOSS has evolved in this generation. CourseMarker, developed at Nottingham University [6] and RoboProf deployed at Dublin City University [4] are examples of this last generation of automatic assessment systems.

3 The E-Lab Philosophy

We developed E-Lab with the idea that we should build it using the latest web technologies and state of the art tools that have been proven to work over the years. The result is a possible forth generation system, where we integrate latest technologies to produce modern, extendable, scalable and easy to use platform. We achieve this using the experience over the years observing students working on programming problems in introduction level courses.

3.1 Integrated problem view

Most of the time available to students trying to solve the problems in the dedicated computer labs is (or should be) spent in three equally important phases. In the first phase students should carefully read and understand the problems, the second phase should be part in which they can refer to the related course material actually attempting to solve the problem resulting in coding the solution in some programming language, and in the third and final phase they should get the feedback for the correctness of their solution.

According to this observation, the platform was designed in such a manner, so that on a single screen students can work and accomplish all the phases involved in solving the problems. As can be seen on figure 1, on that single screen we have the problem text to be read, the web-based code editor to write the solution and the actions pane, so they can run their solution and get instant feedback in the output area. With this design we try to implement the extreme apprenticeship method [10] which is based on a set of values and practices that emphasize learning by doing, together with continuous feedback as the most efficient means for learning programming.

3.2 Authentication

All users of the system are authenticated using the Central Authentication System (CAS) which is used by all services at the faculty. With this mechanism we can identify students and their solutions, and later use this identity to export

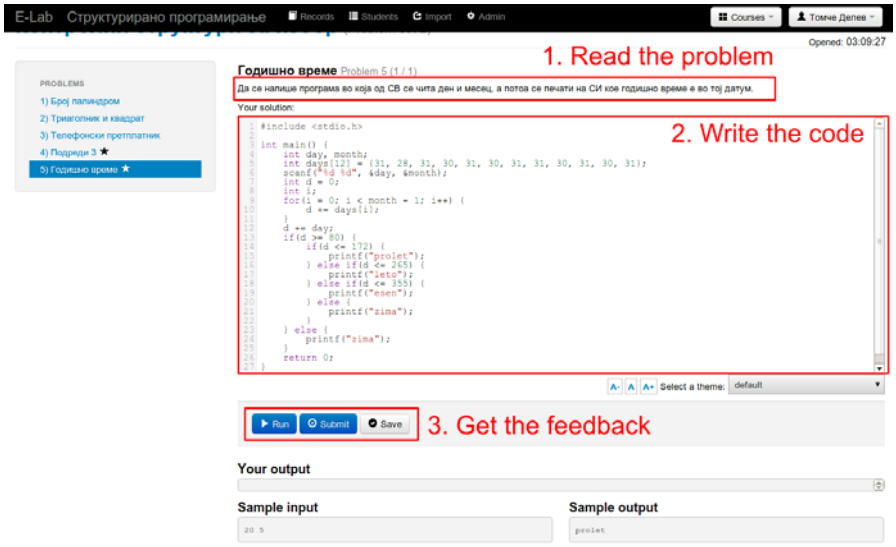


Fig. 1. The student screen trying to solve a problem.

attendance and score records or check for plagiarism, malicious code or other abusive usages.

3.3 Problems design

The central entities in the programming exercises are the programming problems. Each problem is designed in two phases. In the first phase we define the problem text, name and in some problems provide starter code. This information define only the basis of the problem, so in the second phase we need to provide sample input and output for the problem as an example, and at least one test case, also in form of input and output data, so the solutions of the problem can be tested. For each problem, we can also add contextual help or hints that can be helpful for students to solve the problem.

3.4 Automatic assessment

Having limited resources in time and the identified difficulties that tutors and instructors have trying to assess all of the student solutions to the given problems makes the automatic assessment top priority in the platform. Since the platform covers only introduction level courses in programming and algorithms, most of the problems can be designed so they can be assessed by simple black-box testing methodology. For each problem assignment the author provides a reference solution, and using this solution the system automatically generates test cases. Each test case consists of simple input and output text files. When the system

tests the solution, if it's compiled successfully, the executable is fed with the input file and to be correct it should print out the same output as the contents in the generated output file. One of the test cases is a sample and is visible for the students, so they can better understand the problem. In our implementation, each problem should have at least one test case and up to ten test cases. This shouldn't be taken as a general rule, be our choice to limit the number of test cases, was to be able to provide instant feedback. If we have more test cases, then if their execution always results in time out (the program doesn't end in a limited time), the user will need to wait this time out period times the number of test cases.

4 Architecture

The overview of the system architecture can be seen on figure 2, showing its primary components. The data repository is the most interesting component of the system. We propose a specific way of storing the problems and all the work from the students by using a combination of database and file system. All the relational data and metadata of the problems such as the name, the problem set it belongs, the text, are stored in relational database. The other part containing the starter code, reference solution, help contents in mark-up text and all the test cases in form of input and output text files are stored on the file system. And finally, all of the students' solutions are stored solely on the file system in organized directory structure.

4.1 Problems and solutions repository

Almost all of the problems information and solutions are in form of simple text or source code files. Very practical way of storing this kind of data is using some kind of version control system. With this system we get features such as management of the changes of the documents and full revision tracking capabilities. The choice of Git, which is very fast distributed revision control system, gives the system the reliability of the distributed repositories that doesn't depend on single server.

4.2 The client-server

The system is a form of standard client-server web architecture. This architecture allows the client, which is standard and web browser available on all platforms, to run on virtually every PC in our computer lab environment. This lowers the costs of maintenance of the computer labs, because no specific software such as separate client software, compilers, IDEs or text editors should be installed and maintained.

The web server is composed of two separate servers. The front-end web server is a fast web server that serves as a fast proxy and load balancer to the application server. The application server is Java server that uses a scalable RESTfull architecture. The web application on this server follows the MVC architectural

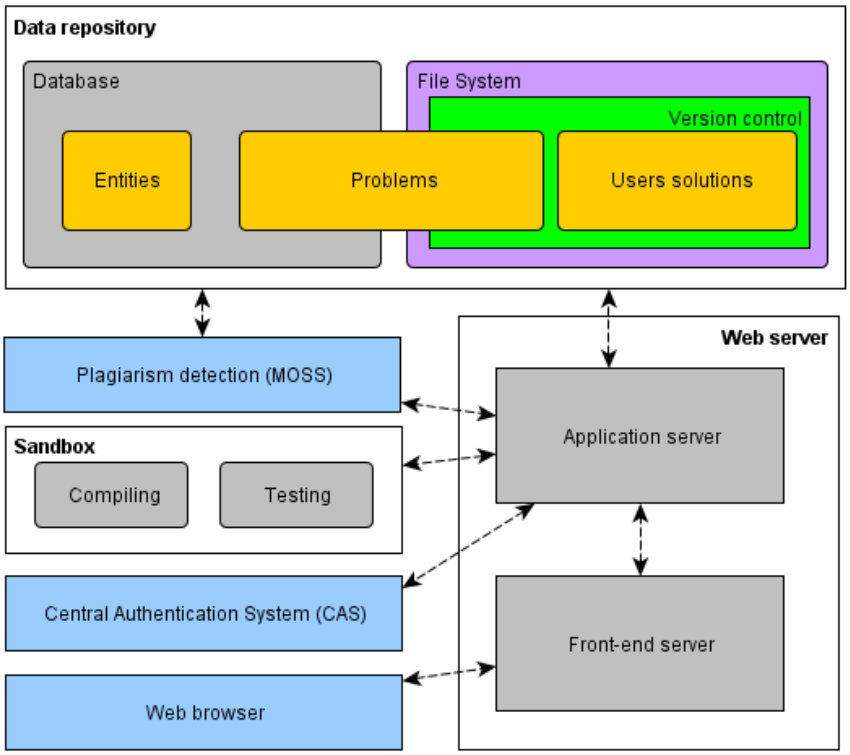


Fig. 2. The E-Lab architecture.

pattern applied to the web architecture. The authentication of the users is done on a central authentication server using HTTPS.

4.3 Asynchronous jobs

In our architecture as in most web based architectures, the web application server is intended to work with very short requests. It uses a fixed thread pool to process requests queued by the HTTP connector. To get optimum results, the thread pool should be as small as possible. The typical optimum value for the default pool size is the number of processors + 1.

That means that if a request is very long, such as waiting the execution of a program that times out (for example 3 seconds), will block the thread pool and penalize the application responsiveness. Of course, we could add more threads to the pool, but it will result in wasted resources, and anyway the pool size will never be infinite.

In the example when users submitting solutions that should be tested on 3 test cases and each of these solutions times out (3 seconds), then the request will last at least 9 seconds (3 test cases x 3 seconds each). When 10 users simultaneously try to submit their solutions, the server will need at least 10 execution threads. This number is feasible, but if we want to have scalable system that supports hundreds or more users submitting solutions, we need different approach.

In these cases, our web framework allows us to temporarily suspend the request. The HTTP request will stay connected, but the request execution will be popped out of the thread pool and tried again later. We execute our long lasting operations such as compiling, saving and executing in an asynchronous way. We use for execution, something called asynchronous job, and while these jobs are executing, the HTTP request is suspended and waits for the result to be available. When the jobs are done with the execution the HTTP request resumes and returns the result to the user.

4.4 Sandboxed execution

The system allows students to write, run and execute any kind of program code that will be executed on a remote server. This can harm the server in many undesirable ways. The malicious code can contain unprivileged read and write access, can create fork bombs, allocate all the available memory or simply consume all the processing power the server has. To control or prevent these security issues all the execution is done in a sandbox environment. In this sandbox each execution is limited by processing time and memory, and also constrained in the number of processes it can fork.

5 Detecting Plagiarism

Source code plagiarism is a serious problem and we must make it very clear to students that the automated system will not be tolerating any kind of source code

plagiarism. The large number of submissions makes it very difficult to manually check for evidences of plagiarism in all possible combinations of solutions. We must use some automatic system for plagiarism detection. Automatic plagiarism detection has been the subject of many studies [2], [3] and there are many systems available online.

In the E-Lab system we incorporate one of these systems trying to prevent and detect plagiarism cases. We use the MOSS system developed by Alex Aiken at UC Berkeley [1]. The Measure Of Software Similarity system makes it possible to objectively and automatically check all problem solutions for evidence of plagiarism or simple copying. MOSS works with programming languages like C, C++, Java, Python and many others. The strategy in our system is to present it very clear to the students, that their solutions will be checked for plagiarism against all solutions submitted by other students. Some of the introduction level problems have very short and simple solutions. We exclude these submissions from plagiarism detection, because the nature of these problems makes it very difficult to write conceptually different solutions.

6 Conclusion

With the development and introduction of the E-Lab system we try to address many organizational aspects of the lab exercises from introduction level programming courses at our faculty. We try to simplify and improve the process of creating and managing simple programming problems. The system is focused on the student and his work and the role of the instructors is to motivate and help students to write working solutions for most of the problems.

The implementation of the central and reliable data repository should also bring many advantages. It contains all students solutions and other important information such as the time when problems were solved or time needed to solve. All the solutions are version controlled, so we can track and analyze the stages in solving and fixing bugs from beginner programmer perspective. All these records, provides us with valuable information from the learning process of the students. From this data, very easy we can extract information such as students attendance records and final scores.

With this system, we are not trying to solve all the organizational and educational problems or entirely exclude the human factor. E-Lab is developed to help with these problems and create modern environment that will motivate and support students work in programming.

S. Markovski, M. Gusev (Editors): ICT Innovations 2012, Web Proceedings, ISSN 1857-7288

References

[1] Aiken, A., et al.: Moss: A system for detecting software plagiarism. University of California–Berkeley. See www. cs. berkeley. edu/aiken/moss. html (2005)
[2] Baker, B.: On finding duplication and near-duplication in large software systems. In: Reverse Engineering, 1995., Proceedings of 2nd Working Conference on. pp. 86–95. IEEE (1995)
[3] Clough, P.: Plagiarism in natural and programming languages: an overview of current tools and technologies
[4] Daly, C., Horgan, J.: An automated learning system for java programming. Education, IEEE Transactions on 47(1), 10–17 (2004)
[5] Douce, C., Livingstone, D., Orwell, J.: Automatic test-based assessment of programming: A review. Journal on Educational Resources in Computing (JERIC) 5(3), 4 (2005)
[6] Higgins, C., Hegazy, T., Symeonidis, P., Tsintsifas, A.: The coursemarker cba system: Improvements over ceilidh. Education and Information Technologies 8(3), 287–304 (2003)
[7] Jenkins, T.: Teaching programming–a journey from teacher to motivator (2001)
[8] Joy, M., Griffiths, N., Boyatt, R.: The boss online submission and assessment system. Journal on Educational Resources in Computing (JERIC) 5(3), 2 (2005)
[9] Saikkonen, R., Malmi, L., Korhonen, A.: Fully automatic assessment of programming exercises. In: ACM SIGCSE Bulletin. vol. 33, pp. 133–136. ACM (2001)
[10] Vihavainen, A., Paksula, M., Luukkainen, M.: Extreme apprenticeship method in teaching programming for beginners. In: Proceedings of the 42nd ACM technical symposium on Computer science education. pp. 93–98. ACM (2011)
[11] Winslow, L.: Programming pedagogy-a psychological overview. ACM SIGCSE Bulletin 28(3), 17–22 (1996)

S. Markovski, M. Gusev (Editors): ICT Innovations 2012, Web Proceedings, ISSN 1857-7288

Toward to the Development of an Integrated Spatial Data Infrastructure in Armenia

H. Astsatryan[1], W. Narsisian[1], V. Ghazaryan[1], A. Saribekyan[1], Sh. Asmaryan[2], V. Muradyan[2], Y. Guigoz[3,4], G. Giuliani[3,4,5] , N. Ray[3,4,5]

[1] Institute for Informatics and Automation Problems of the National Academy of Sciences of the Republic of Armenia, 1, P. Sevak str., Yerevan, 0014, Armenia

hrach@sci.am, vner75@ipia.sci.am, vardangh@gmail.com, albertsaribekyan@rambler.ru

[2] Center for Ecological-Noosphere Studies of the National Academy of Sciences of the Republic of Armenia, 68, Kh. Abovtan str., Yerevan, 0025, Armenia

ashuk@ecocentre.am, muradyan-asx@rambler.ru

[3] Institute for Environmental Sciences, University of Geneva, 7 route de Drize, CH 1227 Carouge / GE, Switzerland

[4]United Nations Environment Programe, Global Resource Information Database, 11 chemin des Anémones, CH 1219 Châtelaine/GE, Switzerland

[5]Forel Institute, University of Geneva, 10 route de Suisse, CP 416, CH-1290 Versoix, Switzerland

{nicolas.ray, gregory.giuliani}@unige.ch, yaniss.guigoz@unepgrid.ch

Abstract. Armenia as a developing country has a very severe need to integrate and collaborate with the international Geospatial community in order to build a reliable and efficient geospatial infrastructure. Moreover, the availability of such an infrastructure could foster the availability of geospatial data, could facilitate the development of a data sharing policy at national level, and could help to efficiently provide Armenian data to the local and international communities. The goal of the article is to introduce the proposed integrated Geospatial infrastructure that is composed of a Spatial Data Infrastructure, tools, services, geocomputational facilities and benchmarking results of some services. The infrastructure will be discussed in the context of hydrological modeling in Armenia.

Keywords: SDI, geospatial data, interoperability, distributed computing, Lake Sevan, SWAT, hydrological model.

S. Markovski, M. Gusev (Editors): ICT Innovations 2012, Web Proceedings, ISSN 1857-7288

1 Introduction

Geographic information systems (GIS) provide facilities to handle spatially referenced data (commonly known as geospatial data) and information. GIS allow one to integrate and handle datasets from different distributed sources. Even when a GIS is available, additional technology is needed to facilitate data sharing and processing across various users and producers. A Spatial Data Infrastructure (SDI) provides such a framework. In this context the deployment of a SDI at national level can facilitate harmonious production, management and usage of geospatial data.

Armenia as a developing country has an essential need to join regional and international geospatial communities to collaborate in building a reliable and efficient infrastructure. Moreover, the availability of such an infrastructure could foster the availability of geospatial data and facilitate the development of a data sharing policy at national level. This would help to efficiently share Armenian data with the local and international communities.

The issue is essential for Armenia, because the country was one of the most industrialized republics of the Soviet Union. Its large industrial enterprises were integrated into a single industrial conglomerate. Large-scale industrial activities (e.g., mining, chemical and electrical industry, machine construction) lead to severe impacts on the environment. Till now the economic policy shifted towards a strong support to industrial development, greatly ignoring ecological interests. Geographically, Armenia lays the highest position in the South Caucasus and is a place of origination of major water arteries: rivers Kura and Araks. All countries of the region sharing borders with Armenia use Kura-Araks catchments and share emerging environmental problems. Consequently, information regarding the ecological state of shared environmental compartment of catchments is necessary for development of short-term and long-term plans for economic and social development of all countries of the Southern Caucasus region, especially considering planned large projects for construction of oil and gas pipelines, construction of a Europe–Caucuses–Asia transport corridor, and other projects. Environment research, awareness and conservation in Armenia are vital for the country's future and for the greater area, which is a direct neighbor and partner of Europe. Thus one of the responses to threats is to build research capacity, the first stage of which is the deployment of the SDI, which will support decision-making at local and regional level.

This paper introduces the technical concepts of an integrated SDI platform for Armenia, which consists of a harmonization approach for tools and services, geocomputational facilities (see fig. 1), and benchmarking results of some services.

S. Markovski, M. Gusev (Editors): ICT Innovations 2012, Web Proceedings, ISSN 1857-7288

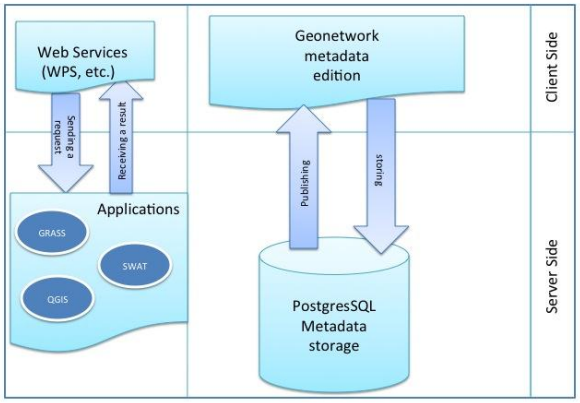


Fig. 1. Structure of Suggested SDI in Armenia

2 Integrated Spatial Data Infrastructure in Armenia

2.1 Geospatial data

There is an obvious acknowledgment of the effective use of geospatial data and information, and standardization is essential in this context. Standardization allows data from one source to be easily used with those from another source to create richer and more useful data. The infrastructure that best permits standardization is the SDI that can be defined as "the relevant base collection of technologies, policies and institutional arrangements that facilitate the availability of and access to spatial data" (Nebert, 2001). The SDI can be complemented with other management tools in order to better handle and process the data. For this purpose the OpenGeo Suite [1] community edition used in conjunction with GeoNetwork allows building a SDI based on free and open source software. These web server applications permit to store data in a PostgreSQL/PostGIS database, to publish data and metadata using interoperability OGC and ISO standards (e.g., Web Map Service, Web Feature Service, Web Coverage Service), to document and catalog data and services in a metadata management system (e.g. GeoNetwork), to disseminate data in various formats, and finally to build webGIS applications. OpenGeo Suite incorporates wide database and raster format support and is designed for strong interoperability. Moreover, it publishes data from any major spatial data source using open standards.

2.2 Tools and Services

The integrated SDI developed for Armenia is composed of several tools that facilitate spatial analysis. We discuss below each of those tools.

GRASS GIS [2-3] is a free and open source GIS software used for geospatial data management and analysis, image processing, graphics/maps production, spatial modeling, and visualization. GRASS is currently used in academic and commercial settings around the world, as well as by many governmental agencies and environmental consulting companies. GRASS is written in a fully modular way. The latest stable release provides more than 400 modules for data management and analysis. GRASS can be run fully automated on distributed computing infrastructures, such as on high performance computing clusters (HPC). However, it is not an easy job to port GRASS modules directly to HPC environment, because we should have good balancing of both data and task distribution, and effective solution to distribute data and tasks among single or multiple clusters environments. Series of experiments to predict and investigate the scalability of GRASS modules on the available computational platform have been carried out.

For example, the module of the Normalized Difference Vegetation Index (NDVI) [4] has been implemented (see fig. 2) for the region including Lake Sevan [5]. Lake Sevan is the unique large water body of Armenia and has a crucial meaning not only in the water balance of the whole South Caucasus, but also in the northern regions of neighbor countries. It is the main strategic supply source of drinking water for Armenia and neighboring countries.

Fig. 2. GRASS NDVI output for Lake Sevan

The benchmarking of NDVI (see fig. 3) has been carried using four computational nodes (each 8 cores) of the Armenian National Grid Initiative (ArmNGI) [6].

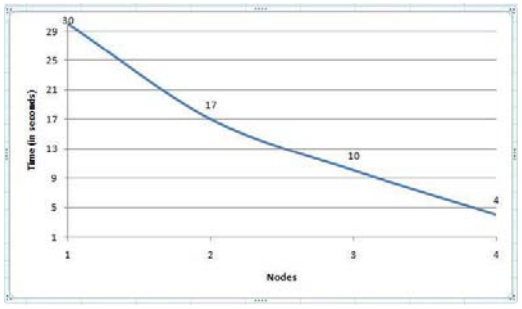


Fig. 3. NDVI benchmarking using four nodes

The experiments show good scalability of the NDVI module and justify the use of distributed computational resources for geoprocessing. For example in this particular case, the running time is about six times less in case of serial using. Some services to stakeholders via PyWPS [7] and WPS GRASS BRIDGE [8] will be later provided, which will use HPC resources in case of complex requests, such as in case of high-resolution satellite image processing.

SWAT (Soil and Water Assessment tool) is another tool [9] that is available in the infrastructure. SWAT is a river basin scale model developed to quantify the impact of land management practices in large, complex watersheds. The importance of the implementation of this tool in Armenia is crucial because there is no exact or accurate estimation of any land use impact on any watershed in Armenia. Additionally, we have to mention the importance of getting the HRU (Hydrological Response Units) properties of each sub-basin in the watershed. HRUs are the smaller spatial units describing watershed properties (soil type, landuse, etc). A calibrated SWAT model can then be used to obtain outputs on various important aspects of the watershed such as: water quality, water quantity, sediment concentration and more. Series of experiments have been carried out using the model based on ArcGIS (version 9.3.1) [10] with ArcSWAT [11]. As an initial implementation, corresponding HRU for Argichi river basin (part of Lake Sevan catchment basin) have been delineated and obtained (see fig. 4). The HRU analyses full report includes land use, soils and slope distribution, and final HRU distribution.

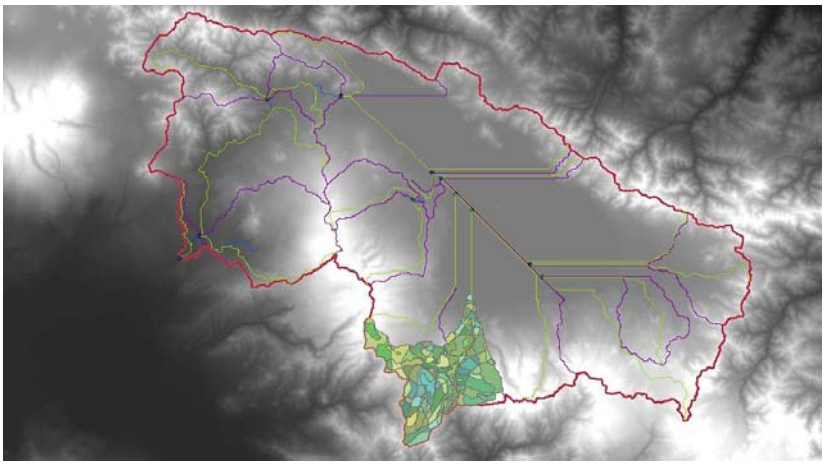


Fig. 4. Delineation of HRU for Argichi river basin (red - mask border, green – sub-basin border, violet - HRU border)

Finally, as services it is planned to implement the standards promoted by the Open Geospatial Consortium [12]. Some experiments have been carried out using WPS (Web Processing Service), which provides rules for standardizing how inputs and outputs (requests and responses) for geospatial processing services (such as polygon overlay) are encoded in the web service. The WPS standard also defines how a client can request the execution of a process, and how the output from the process is handled.

2.3 Computational and Storage Resources

The heterogeneous computational (more than 500 cores) and storage resources offered by ArmNGI and located in the leading research (National Academy of Sciences, Yerevan Physics Institute) and educational (Yerevan State University, State Engineering University) organizations of Armenia, was used for the distributed resources for SDI. For instance, spatial satellite image processing requires a large amount of computation time due to its complex and large processing criteria. In the future the integrated SDI could benefit from other VOs outside the ArmNGI, such as the common geocomputation infrastructure of the ENVIROGRIDS [13] community.

3 Conclusion

The suggested infrastructure is compatible with OGC international standards, and can help sharing Armenian data to International initiatives such as GEOSS

(Global Earth Observation System of Systems) and INSPIRE (Infrastructure for Spatial Information in the European Community). The infrastructure is scalable over a distributed infrastructure. The various tools to complement the SDI are now integrated for strengthening research capacities and promoting innovative ways of approaching challenges related to the ever increasing flow of environmental data in Armenia. These new capacities will facilitate data exchange, sharing and updating, and will improve the accessibility to Armenian environmental data. Due to new facilities, Armenian researchers will be empowered with new highly-competitive technical skills by enhancing their national, regional and international networking.

Acknowledgements

This work was supported by the Swiss National Science Foundation (grant n° 137325) through the project SCOPES ARPEGEO ("Deploying ARmenian distributed Processing capacities for Environmental GEOspatial data" [14]).

References

1. OpenGeo, http://opengeo.org
2. Neteler, N., Mitasova, H., Open Source GIS: A GRASS GIS Approach. Third edition. 420 pages, Springer, New York (ISBN-10: 038735767X; ISBN-13: 978-0387357676), 2007.
3. GRASS Development Team, 2011. Geographic Resources Analysis Support System (GRASS) Software. Open Source Geospatial Foundation Project. http://grass.osgeo.org
4. Carroll, M.L., C.M. DiMiceli, Sohlberg, R.A., Townshend, J.R.G, 250m MODIS Normalized Difference Vegetation Index, 250ndvi28920033435, Collection 4, University of Maryland, College Park, Maryland, Day 289, 2003
5. Lake Sevan Action Program. Main report.1999. Ministry of Narture Prote tion and The World Bank: Yerevan and Washington DC. 47pp.
6. H. Astsatryan, V. Sahakyan, Yu. Shoukourian, Brief Introduction of Armenian National Grid Initiative, Book of Abstracts of 4th International Conference "Distributed Computing and Grid-technologies in Science and Education" (Grid'2010), pp. 30-31, June 28 - July 3, 2010, Dubna, Russia.
7. Python Web Processing Service, http://pywps.wald.intevation.org/
8. WPS GRASS BRIDGE, http://code.google.com/p/wps-grass-bridge.

9. Neitsch, S. L., Arnold, J. G., Kiniry, J. R., & Williams, J. R. (2001). Soil and Water Assessment tool (SWAT) user's manual version 2000. Grassland Soil and Water Research Laboratory. Temple, TX: ARS.
10. Environmental Systems Research Institute. http://www.esri.com
11. ArcGIS-ArcView extension and graphical user input interface for SWAT, http://swatmodel.tamu.edu/software/arcswat/
12. Open Geospatial Consortium, http://www.opengeospatial.org
13. Building Capacity for a Black Sea Catchment Observation and Assessment System supporting Sustainable Development, http://envirogrids.net/
14. Deploying ARmenian distributed Processing capacities for Environmental GEOspatial data, http://arpegeo.sci.am

Increasing the Decoding Speed of Random Codes Based on Quasigroups

Aleksandra Popovska-Mitrovikj, Smile Markovski, and Verica Bakeva

University "Ss Cyril and Methodius" - Skopje,
Faculty of Computer Science and Engineering
P.O. Box 393, Republic of Macedonia
{aleksandra.popovska.mitrovikj,smile.markovski,
verica.bakeva}@finki.ukim.mk

Abstract. Error-correcting codes based on quasigroup transformations are defined elsewhere. The speed of the decoding process is one of the biggest problems for these codes. In order to improve the decoding speed, we have defined a new algorithm of decoding. Now, we use two transformations of the redundant message by using different parameters, and the candidates for the decoded messages are obtained by using intersection of the corresponding sequences. In such a way the decoding is at least 4 times faster. Also, some other improvement of the standard decoding process are considered where improving of the packet-error probability (PER) and the bit-error probability (BER) are obtained.

Keywords: quasigroup, quasigroup transformation, random code, error-correcting code, packet-error probability, bit-error probability

1 Introduction

Random codes based on quasigroups (RCBQ) are proposed in [1]. RCBQ have several parameters and in [2] we have investigated the influence of the code parameters to the code performances. In this paper we define some modifications of the process of decoding in order to improve the performances of these codes. Since the decoding speed is one of the biggest problem for these random codes we propose a new method of decoding such that the modified decoding process is 4.5 times faster than the original one. Also, we consider some other modifications of the decoding algorithm for decreasing the packet-error probability (PER) and the bit-error probability (BER).

The paper is organized as follows. In Section 2 we give the definition of quasigroup transformations and description of the RCBQ, i.e., the algorithms of coding and decoding. In Section 3 we propose a new algorithm of decoding and modification on the decoding rule in order to eliminate the unsuccessful decoding with more candidates. In this section we present the experimental results obtained with these modifications. In Section 4, we consider other modification of the standard decoding process which improve the packet-error probability

S. Markovski, M. Gusev (Editors): ICT Innovations 2012, Web Proceedings, ISSN 1857-7288

(PER) and the bit-error probability (BER). Finally, some conclusions are given in Section 6.

2 Preliminary Definitions and Description of RCBQ

Since the RCBQ are designed using quasigroup string transformations we give breafly some definitions of quasigroups transformations and description of the coding and decoding algorithm of these codes.

2.1 Quasigroups and Quasigroups String Transformation

A quasigroup $(Q, *)$ is a groupoid, i.e., a set Q with a binary operation $* : Q^2 \to Q$, such that for all $u, v \in Q$, there exist unique $x, y \in Q$, satisfying the equalities $u * x = v$ and $y * u = v$. In the sequel we assume that the set Q is a finite set. The main body of the multiplication table of a quasigroup is a Latin square over the set Q. Given a quasigroup $(Q, *)$ a new operation $\backslash$, called a parastrophe, can be derived from the operation $*$ as follows:

$$x * y = z \Leftrightarrow y = x \backslash z. \tag{1}$$

Then the algebra $(Q, *, \backslash)$ satisfies the identities

$$x \backslash (x * y) = y \quad \text{and} \quad x * (x \backslash y) = y, \tag{2}$$

and $(Q, \backslash)$ is also a quasigroup.

Quasigroup string transformations are defined on a finite set Q endowed with a quasigroup operation $*$, and they are mappings from Q^+ to Q^+, where Q^+ is the set of all nonempty words on Q. Here, we use two types of quasigroup transformations as explained below. Let $l \in Q$ be a fixed element, called a leader. For every $a_i, b_i \in Q$, $e-$ and $d-$transformations are defined as follows:

$$\begin{aligned} e_l(a_1 a_2 \dots a_n) = b_1 b_2 \dots b_n &\Leftrightarrow b_{i+1} = b_i * a_{i+1}, \\ d_l(a_1 a_2 \dots a_n) = b_1 b_2 \dots b_n &\Leftrightarrow b_{i+1} = a_i \backslash a_{i+1}, \end{aligned} \tag{3}$$

for each $i = 0, 1, \dots, n-1$, where $b_0 = a_0 = l$. By using the identities (2), we have that $d_l(e_l(a_1 a_2 \dots a_n)) = a_1 a_2 \dots a_n$ and $e_l(d_l(a_1 a_2 \dots a_n)) = a_1 a_2 \dots a_n$. This means that e_l and d_l are permutations on Q^n, mutually inverse. In the code design compositions of e_l and d_l are used.

Theorem 1. [3] *Consider an arbitrary string $\alpha = a_1 a_2 \dots a_n$ where $a_i \in Q_i$, and let β be obtained after k applications of an $e-$transformation on α. If n is an enough large integer then, for each $1 \le t \le k$, the distribution of substrings of β of length t is uniform.* (We note that for $t > k$ the distribution of substrings of β of length t is not uniform.)

Code design uses the alphabet $Q = \{0, \dots, 9, a, b, c, d, e, f\}$ of nibbles (4-bit letters), and a quasigroup operation $*$ on Q together with its parastrophe $\backslash$.

2.2 Description of RCBQ

Let $M = m_1 m_2 ... m_r$ be a block of N_{block} bits, where m_i is a nibble (4-bit letter); hence, $N_{block} = 4r$. We first add redundancy as zero bits and produce block $L = L^{(1)}L^{(2)}...L^{(s)} = L_1 L_2 ... L_m$ of N bits, where $L^{(i)}$ are 4-nibble words, L_i are nibbles, so $m = 4s$, $N = 16s$. After erasing the redundant zeros from each $L^{(i)}$, the message L will produce the original message M. On this way we obtain an (N_{block}, N) code with rate $R = N_{block}/N$. The codeword is produced from L after applying the encryption algorithm given in Figure 1. For that aim, previously, a key $k = k_1 k_2 \dots k_n$ of length n nibbles should be chosen. The obtained codeword of M is $C = C_1 C_2 ... C_m$, where C_i are nibbles.

After transmission through a noise channel (for our experiments we use binary symmetric channel), the codeword C will be received as message $D = D^{(1)}D^{(2)}...D^{(s)} = D_1 D_2 ... D_m$, where $D^{(i)}$ are blocks of 4 nibbles and D_j are nibbles. The decoding process consists of four steps: (i) procedure for generating the sets with predefined Hamming distance, (ii) inverse coding algorithm, (iii) procedure for generating decoding candidate sets and (iv) decoding rule.

Encryption	**Decryption**
Input: Key $k = k_1 k_2 \dots k_n$ and $L = L_1 L_2 \dots L_m$ **Output**: codeword $C = C_1 C_2 ... C_m$	**Input**: The pair $(a_1 a_2 \dots a_s, k_1 k_2 \dots k_n)$ **Output**: The pair $(c_1 c_2 \dots c_s, K_1 K_2 \dots K_n)$
For $j = 1$ to m $X \leftarrow L_j$; $T \leftarrow 0$; For $i = 1$ to n $X \leftarrow k_i * X$; $T \leftarrow T \oplus X$; $k_i \leftarrow X$; $k_n \leftarrow T$ **Output**: $C_j \leftarrow X$	For $i = 1$ to n $K_i \leftarrow k_i$; For $j = 0$ to $s - 1$ $X, T \leftarrow a_{j+1}$; $temp \leftarrow K_n$; For $i = n$ to 2 $X \leftarrow temp \setminus X$; $T \leftarrow T \oplus X$; $temp \leftarrow K_{i-1}$; $K_{i-1} \leftarrow X$; $X \leftarrow temp \setminus X$; $K_n \leftarrow T$; $c_{j+1} \leftarrow X$; **Output**: $(c_1 c_2 \dots c_s, K_1 K_2 \dots K_n)$

Fig. 1. Algorithm for encryption and decryption

The probability that $\leq t$ bits in $D^{(i)}$ are not correctly transmitted is $P(p;t) = \sum_{k=0}^{t} \binom{16}{k} p^k (1-p)^{16-k}$, where p is probability of bit-error in a binary symmetric channel. Let B_{max} be an integer such that $1 - P(p; B_{max}) \leq q_B$ and $H_i = \{\alpha | \alpha \in Q^4, \ \ H(D^{(i)}, \alpha) \leq B_{max}\}$, for $i = 1, 2, \dots, s$, where $H(D^{(i)}, \alpha)$ is the Hamming distance between $D^{(i)}$ and α.

The decoding candidate sets S_0, S_1, S_2,..., S_s are defined iteratively. Let $S_0 = (k_1 \dots k_n; \lambda)$, where λ is the empty sequence. Let S_{i-1} be defined for $i \geq 1$.

Then S_i is the set of all pairs $(\delta, w_1 w_2 \dots w_{16i})$ obtained by using the sets S_{i-1} and H_i as follows. (Here, w_j are bits). For each $(\beta, w_1 w_2 \dots w_{16(i-1)}) \in S_{i-1}$ and each element $\alpha \in H$, we apply the inverse coding algorithm (i.e. algorithm for decryption given in Figure 1) with input (α, β). If the output is the pair (γ, δ) and if both sequences γ and $L^{(i)}$ have the redundant nibbles in the same positions, then the pair $(\delta, w_1 w_2 \dots w_{16(i-1)} c_1 c_2 \dots c_{16}) \equiv (\delta, w_1 w_2 \dots w_{16i})$ is an element of S_i.

The decoding of the received codeword D is given by the following rule: If the set S_s contains only one element $(d_1 \dots d_n, w_1 \dots w_{16s})$ then $L = w_1 \dots w_{16s}$. In this case, we say that we have a *successful decoding*. In the case when the set S_s contains more than one element, we say that the decoding of D is unsuccessful (of type *more-candidate-error*). In the case when $S_j = \emptyset$ for some $j \in \{1, \dots, s\}$, the process will be stopped (*null-error* appears). We conclude that for some $m \leq j$, $D^{(m)}$ contains more than B_{max} errors, resulting with $C_m \notin H$.

Theorem 2. [1] *The packet-error probability of these codes is* $q = 1 - (1 - q_B)^s$.

2.3 Calculating PER and BER

In experiments with RCBQ in [2] we calculate the values of PER and BER on the following way. If the decoding process completed successfully (the last set S_s of candidates for decoding has only one element), the decoded message is compared with the input message. If they differ at least one bit, then we say that an uncorrected-error appears. Then we compute the number of incorrectly decoded bits as Hamming distance between the input and the decoded message.

We also calculate the number of incorrectly decoded bits when the decoding process finish with *more-candidate-error* or *null-error*. Then, that number is calculated as follows.

When *null-error* appears, i.e., $S_i = \emptyset$, we take all the elements from the set S_{i-1} and we find their maximal common prefix substring. If this substring has k bits and the length of the sent message is m bits ($k \leqslant m$), then we compare this substring with the first k bits of the sent message. If they differ in s bits, then the number of incorrectly decoded bits is $m - k + s$.

If a *more-candidates-error* appears we take all the elements from the set S_s and we find their maximal common prefix substring. The number of incorrectly decoded bits is computed as previous.

The total number of incorrectly decoded bits is the sum of all of the previously mentioned numbers of incorrectly decoded bits.

We compute the probability of packet-error as

$$\text{PER} = \#(\textit{incorrectly decoded packets})/\#(\textit{all packets})$$

and the probability of bit-error as

$$\text{BER} = \#(\textit{incorrectly decoded bits in all packets}) \ / \ \#(\textit{bits in all packets}).$$

3 A New Algorithm of Decoding

From our previous experiments with the RCBQ we concluded that the speed of the decoding process is one of the biggest problem for these random codes. Depending on the chosen pattern for redundant zero nibbles, the decoding process is slow in some iterations since the number of elements in the sets S is very large. If we distribute the redundant zeros more uniformly, than the number of elements in the sets S are not going to be very large, but then, many more-candidate-errors will appear. Therefore, in the patterns we need to put more redundant zeros at the end. In fact, finding good equilibrium for placing the redundant zeros is an open problem and by experiments several enough satisfactory patterns are discovered.

In order to improve the decoding speed, in this paper we propose a new algorithm of decoding, called *cut-decoding algorithm*. In the new method, instead of using a (k, n) code with rate R, we use together two $(k, n/2)$ codes with rate $2R$, that decode a same message of k bits. First, we apply the coding algorithm, given in Figure 1, on the same redundant message L twice using different parameters (different keys or quasigroups). In this way we obtain the codeword of the message as concatenation of the two codewords of $n/2$ bits. After transmitting through a binary symmetric channel, we divide the outgoing message $D = D^{(1)}D^{(2)}...D^{(s)}$, where $D^{(i)}$ are blocks of 4 nibbles, in two messages $D^{(1)} = D^{(1)}D^{(2)}...D^{(s/2)}$ and $D^{(2)} = D^{(s/2+1)}D^{(s/2+2)}...D^{(s)}$ with equal lengths and we decode them parallel with the corresponding parameters. In the new method of decoding we make modification in the part (iii) of the decoding process, i.e., in the procedure for generating decoding candidate sets. In the new algorithm we generate these sets on the following way.

Step 1. Let $S_0^{(1)} = (k_1^{(1)} \dots k_n^{(1)}; \lambda)$ and $S_0^{(2)} = (k_1^{(2)} \dots k_n^{(2)}; \lambda)$ where λ is the empty sequence, $k_1^{(1)} \dots k_n^{(1)}$ and $k_1^{(2)} \dots k_n^{(2)}$ be the initials keys used for obtaining the two codewords.

Step 2. Let $S_{i-1}^{(1)}$ and $S_{i-1}^{(2)}$ be defined for $i \geq 1$.

Step 3. Let two decoding candidate sets $S_i^{(1)}$ and $S_i^{(2)}$ be obtained in the both decoding processes, on the same way as in the standard RCBQ.

Step 4. Let $V_1 = \{w_1w_2 \dots w_{16i} | (\delta, w_1w_2 \dots w_{16i}) \in S_i^{(1)}\}$, $V_2 = \{w_1w_2 \dots w_{16i} | (\delta, w_1w_2 \dots w_{16i}) \in S_i^{(2)}\}$ and $V = V_1 \cap V_2$.

Step 5. For each $(\delta, w_1w_2 \dots w_{16i}) \in S_i^{(1)}$, if $w_1w_2 \dots w_{16i} \notin V$ then $S_i^{(1)} \leftarrow S_i^{(1)} \backslash \{(\delta, w_1w_2 \dots w_{16i})\}$. Also, for each $(\delta, w_1w_2 \dots w_{16i}) \in S_i^{(2)}$, if $w_1w_2 \dots w_{16i} \notin V$ then $S_i^{(2)} \leftarrow S_i^{(2)} \setminus \{(\delta, w_1w_2 \dots w_{16i})\}$.
(* Actually, we eliminate from $S_i^{(1)}$ all elements whose second part does not matches with the second part of an element in the $S_i^{(2)}$, and vice versa. In the next iteration the both processes use the corresponding reduced sets $S_i^{(1)}$ and $S_i^{(2)}$. *)

Step 6. If $i < s$ then increase i and go back to Step3.

If, after the last iteration, the reduced sets $S_s^{(1)}$ and $S_s^{(2)}$ have only one element with same second component $w_1 \dots w_{rs}$, then $L = w_1 \dots w_{rs}$. In this case, we say that we have successful decoding. If, after the last iteration, the reduced sets $S_s^{(1)}$ and $S_s^{(2)}$ have more than one element we have *more-candidate-error*. If we obtain $S_i^{(1)} = \emptyset,\ S_i^{(2)} \neq \emptyset$ or $S_i^{(2)} = \emptyset,\ S_i^{(1)} \neq \emptyset$ in some iteration then the decoding of the message continues only with the nonempty set $S_i^{(2)}$ or $S_i^{(1)}$, by using the standard RCBQ decoding process. In the case when $S_i^{(1)} = S_i^{(2)} = \emptyset$ in some iteration, then the process will be stopped (*null-error* appears).

In experiments with the new method of decoding, we noticed a significant reduction in the number of elements in the sets S and we achieve big improvement in the speed of the decoding process. Namely, the new method of decoding is 4.5 times faster. In the Table 1 we give one example of the average number of elements in the sets $S_i^{(1)}$ and $S_i^{(2)}$, in each iteration, before and after the reduction given in Step 4 of the new method of decoding. As it is expected, the proposed reduction in Step 4 significantly decreases the number of elements in the decoding candidate sets, that can be seen from Table 1.

Table 1. Average number of elements in the decoding candidate sets before and after reduction

No. iteration	No. elements in $S_i^{(1)}$	No. elements in $S_i^{(2)}$	No. elements in reduced sets $S_i^{(1)}$, $S_i^{(2)}$
1	10.4	9.9	1.6
2	244.2	244.2	18.2
3	178.5	181.9	8.1
4	79.7	80.5	4.4
5	693.5	695.1	34.8
6	343.9	339.5	14.2
7	140.6	139.3	6.2
8	60.8	61.5	3.4
9	1.1	1.0	1.0

The problem in the new method of decoding is that for obtaining code with rate R we need a pattern for code with twice larger rate. But, it is hard to make good pattern for larger rates, since the number of redundant zeros in these patterns is smaller. Therefore, with the new proposed method of decoding we obtain worse results in the number of unsuccessful decoding of type *more-candidate-error*, but the number of unsuccessful decoding with *null-error* is smaller.

To resolve the problem of greater number of *more-candidate-errors* we propose one heuristic in the decoding rule for elimination of this type of errors. Namely, from the experiments with the RCBQ we can see that when the decoding process ends with more elements in the last set S_s, almost always in this set is the correct message. So, in these cases we can randomly select a message from the set S_s in the last iteration and it can be taken as the decoded message. If the selected message is the correct one, then the bit-error is 0, so the BER will

also be reduced. In the experiments we have made with this modification we got that in around half of the cases, when the last set S_s contains more than one element, the correct message is selected.

3.1 Experimental Results

We presented in [2] the experimental results obtained for (72,288) codes with rate R=1/4 for different patterns of redundancy, when binary symmetric channel is used. The best result for this codes was for the pattern: 11001100100000000110010001000000011001100100000000110010 0010000000000000, key $k = 0123456789$ and a quasigroup given in the same paper. For making a comparison, we have made experiments for (72,288) codes with the new algorithm of decoding, and we have considered 17 different patterns for (72, 144) codes of rate 1/2.

Experiments with different keys. First, we made experiments using only different keys in the two process of coding and decoding and the same quasigroup given in [2]. The best results we obtained for the pattern 110011101100110011101100110011000000 and keys $k_1 = 01234$, $k_2 = 56789$ (we also tried with keys of length 10 nibbles but the results were similar). The obtained results for PER and BER with the new cut-decoding algorithm and with the standard method of decoding, for different values of bit-error probability p of binary symmetric channel and $B_{max} = 4$, are given in Table 2 and presented in Figure 2 and Figure 3.

Table 2. Experimental results for packet-error and bit-error probability

p	PER_{old}	PER_{new}
0.02	0.00125	0.00171
0.03	0.00171	0.00257
0.04	0.00594	0.00514
0.05	0.01594	0.01714
0.06	0.03594	0.02657
0.07	0.06656	0.06171
0.08	0.11313	0.10800
0.09	0.18875	0.15886

p	BER_{old}	BER_{new}
0.02	0.00076	0.00011
0.03	0.00089	0.00083
0.04	0.00343	0.00302
0.05	0.00928	0.01134
0.06	0.02239	0.01943
0.07	0.04065	0.04243
0.08	0.06485	0.07512
0.09	0.11357	0.11621

We conclude from the results that we have approximately the same results for the values of PER and BER with both decoding algorithms. Moreover, we note that for $p > 0.05$ we have slightly better values of PER with the new algorithm of decoding.

Experiments with different keys and different quasigroups. We have made experiments with the new method of decoding using (72,144) codes with different keys and different quasigroups. The results obtained using the quasigroup given in [2] and the quasigroup, with high coefficient of period growth, given at the end of the paper [4], are similar with the previous results obtained using same quasigroup in the both process. Nevertheless, if we use the cyclic group of order 16 in one of the codes then the results for PER and BER are much worse.

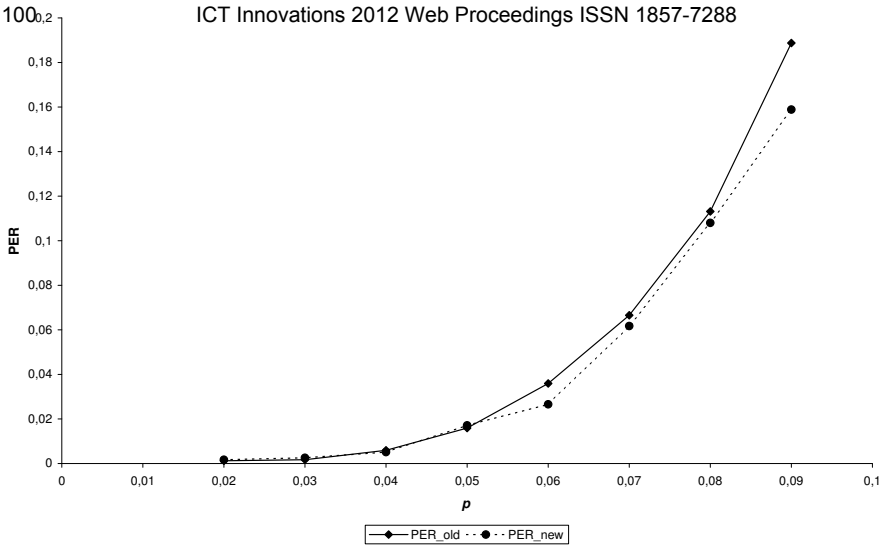


Fig. 2. Comparison of PER

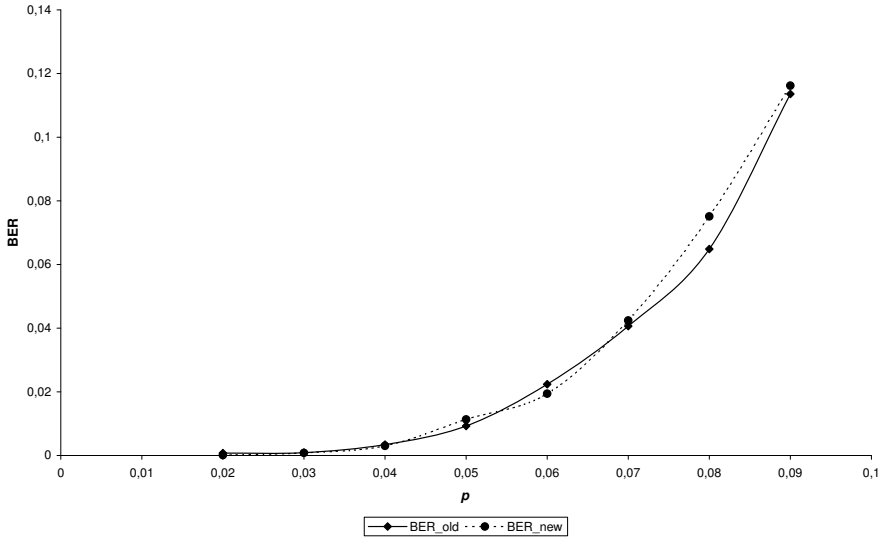


Fig. 3. Comparison of BER

No improvement was made if we use different keys if their length was at least 5 nibbles.

Experiments for rate 1/8. We have made experiments for (72, 576) codes of rate 1/8 by using both methods. In this experiments we obtained better results for values of PER and BER with the new algorithm of decoding. We also have improvement in the speed of decoding, which is more significant if we use

$B_{max} = 5$ in the process of decoding. In this case, the new decoding process is 5.2 times faster than the standard decoding process.

4 A Method for Decreasing the Number of Null-errors in the New Decoding Algorithm

In [2] we have proposed a method for decreasing the number of *null-errors* by backtracking. Namely, if in the i-th iteration of the decoding process we obtain $S_i = \emptyset$, it means that in some previous step $j < i$ we have lost the right block that had to be processed. This happened if the number of errors during transmission in some block is greater than B_{max}. Some of these errors will be eliminated if we cancel a few of iterations of the decoding process and we reprocess all of them or part of them with a larger value of B_{max}.

Now, we have made experiments using this idea in the new algorithm of decoding when a *null-error* appear. We made experiments with returning two or three iterations back and using $B_{max} + 1$ or $B_{max} + 2$ in the first of canceled iterations (the remain iterations use the previous value of B_{max}). We must note that with this backtracking in some cases instead of *null-error* we obtain *more-candidate-error*. We obtain the best results with two iterations back and $B_{max}+2$ and the percentages of eliminated unsuccessful decoding with the null-error are given in Table 3. These results are obtained by applying of this modification on the unsuccessful decoded message (with null-error) from the experiments which results are presented in the Table 2. We can conclude that this modification gives good percentage of eliminated *null errors*.

Table 3. Percentage of eliminated unsuccessfully decoding with null-error

p	Percentage of eliminated null-errors
0.03	28.57%
0.04	20.37%
0.05	24.00%
0.06	23.17%
0.07	27.23%
0.08	25.37%
0.09	22.64%

Very important for this modification is that we have only significantly small increasing of the decoding speed. Also, by eliminating some of the *null errors* we obtain better results for BER. In Table 4 we compare the values of $\mathrm{PER}_{\mathrm{new}}$ and $\mathrm{BER}_{\mathrm{new}}$ from Table 2 (obtained by the new cut-decoding) without backtracking and values of $\mathrm{PER}_{\mathrm{new+back}}$ and $\mathrm{BER}_{\mathrm{new+back}}$ obtained by using backtracking (two iterations back and $B_{max} + 2$). We conclude from the results of Table 4 that for larger values of p we have greater improvement of PER. Also, by this modification we achieve better values of BER for all p. This happens since in

the cases when the *null error* will not be eliminated by backtracking, the empty reduced sets can occur in some higher iteration, so the bigger part of the message will be decoded and the value of the bit-error will be smaller.

Table 4. Experimental results for PER and BER with and without backtracking

p	PER_{new}	$\text{PER}_{\text{new+back}}$
0.03	0.00257	0.00257
0.04	0.00514	0.00514
0.05	0.01714	0.01429
0.06	0.02657	0.02171
0.07	0.06171	0.04857
0.08	0.10800	0.08200
0.09	0.15886	0.12543

p	BER_{new}	$\text{BER}_{\text{new+back}}$
0.03	0.00083	0.00077
0.04	0.00302	0.00165
0.05	0.01134	0.00510
0.06	0.01943	0.00892
0.07	0.04243	0.02194
0.08	0.07512	0.03795
0.09	0.11621	0.05383

5 Conclusion

In this paper we proposed a new decoding algorithm of random codes based on quasigroups. The main idea was to make cuts of the sets S_i used in the standard RCBQ algorithm. Our aim was to improve the decoding speed of these codes. Several experiments were performed with different types of quasigroups, different types of patterns and different lengths of the key. The new method gives around 4.5 times faster decoding process than the older one for (72,288) codes. Also, the elimination of *more-candidate-errors* is higher.

Some other improvement of the standard decoding process are considered where improving of the packet-error probability (PER) and the bit-error probability (BER) are obtained.

These results open the problem of using cuts of three or more of the sets S_i in order the decoding speed to be more increased.

References

1. Gligoroski, D., Markovski, S., Kocarev, Lj.: Error-Correcting Codes Based on Quasigroups. In: Proceedings of 16th International Conference on Computer Communications and Networks (ICCCN 2007), pp. 165-172.
2. Popovska-Mitrovikj, A., Markovski, S., Bakeva, V.: Performances of error-correcting codes based on quasigroups. In: Davcev, D., Gomez, J.M. (eds.) ICT-Innovations 2009, pp. 377-389. Springer (2009)
3. Markovski, S., Gligoroski, D., Bakeva, V.: Quasigroup string processing: Part 1. In: Maced. Acad. of Sci. and Arts, Sec. Math. Tech. Scien, vol. XX 1-2, pp. 13-28, (1999)
4. Dimitrova, V., Markovski, J.: On Quasigroup Pseudo Random Sequence Generators. In: Proceedings of the 1^{st} Balkan Conference in Informatics, pp.393-401, Greece(2003)

Sign Language Tutor – Digital improvement for people who are deaf and hard of hearing

Nevena Ackovska, Magdalena Kostoska, Marjan Gjurovski

Faculty of Computer Science and Engineering,
St. Cyril and Methodious University,
Skopje, Republic of Macedonia

{nevena.ackovska, magdalena.kostoska}@finki.ukim.mk,
maki.gjuroski@gmail.com

Abstract. This paper is an introduction to the world of deaf and hard of hearing people and their everyday challenges (and opportunities) with technology and computer interaction. The requirements of this focus group in human-computer interaction and the currently available computers tools and technologies in the world, and especially Macedonia will be discussed. An overview of the present research directions and the possibilities of the visualization and 3D technology for this target group will be stated.

In this paper the Sign Language Tutor – sign language interactive e-learning platform will be presented. It represents a collection of modules and games dedicated to ease the learning of the Macedonian Sign Language (MSL), but also to improve the mental and memory capabilities, especially of the younger part of our target group – the deaf children. The central part of this project is 3D simulation – given a 3D model of a girl the subject should sign a chosen letter or object. Computer games to assist with the learning are used: one is a 2D adventure where the hero fights monsters and collects items as rewards – sign of the collected object, the other is memory where the subject should connect a card with a sign of the letter. The platform is built using the Microsoft XNA technology.

Keywords: deaf, hard of hearing, sign language, children, tutorial, human-computer interaction

1 Introduction

The community of deaf and hard of hearing people is not a small community. According to the World Health Organization in 2004 there were over 275 million people globally with moderate-to-profound hearing impairment [1]. Even more, this group of people don't want to be considered as a part of the community of disabled. According to the World Federation of Deaf "*Deaf and hard of hearing people do not identify as having a disability or see themselves as experiencing a limitation. Instead,*

S. Markovski, M. Gusev (Editors): ICT Innovations 2012, Web Proceedings, ISSN 1857-7288

they identify as a member of a cultural and linguistic group."[2]. However, they live isolated from the rest of the world. The reason for this is the inability for communication with the non-hard of hearing world. Regular world has little knowledge of the challenges that heard of hearing or deaf people experience in everyday activities. For a long time it have been generally accepted that the deaf and hard of hearing people are unintelligent, so that's why the offensive term "deaf and dumb" has been used. The terms deaf and hard of hearing are voted as an official designation in 1991 by the World Federation of the Deaf (WFD) [2] [3].

In a large study conducted by Conrad [4] it has been discovered that deaf and hard of hearing children attending schools using an oral approach rarely acquire sufficient lip-reading skills. The children with hearing loss greater than 85 dB could only comprehend about 25% to 28% of the words through lip-reading that they could comprehend through reading. The children with hearing loss less than 65 dB could only comprehend about 36% of the words through lip-reading that they could comprehend through reading. Also the studies have shown that many deaf and hard of hearing individuals have difficulty with reading [4][5]. This is explained by the fact that the words and the writings are based on the spoken language, something that the deaf are not truly acquainted to [6].

The language of the deaf and hard of hearing people is called Sign Language (SL). This language can be signed with usage not only of the hands, but also the head, lips or the torso. This is not universal language and the languages differ from country to country. Also the signed languages are not based on the spoken language of the region.

In Macedonia there are around 6000 deaf people [7] and according to the National association of deaf and hard of hearing of Macedonia there are only 12 licensed interpreters of the Macedonian Sign Language (MSL) [8]. The signs of the alphabet together with the terms for mother, father and some basic colors are the only information about MSL that can be, from recently, found online. Until 3 months ago this information was not available. Since then the National association upgraded their web page and published this information online. Unlike the bigger countries that promote their national sign language, there are no e-books, videos or any other type on online content that can help learning the MSL. This community in Macedonia is highly marginalized. Only one television channel (the national channel) offers once a day a MSL interpreted news and a weekly show called "The world of silence". All the other programs and TV channels don't offer even subtitles to their contents in Macedonian language. These are the reasons why we started this project. We wanted to enter this world and to offer the possibility of learning MSL to every citizen of our country.

1.1 Macedonian Sign Language

The Macedonian sign language is based on gestures and body gestures, as all the other sign languages. The hands are the basic communication means. The signs are performed with predefined movement and location using one or both hands. Not every hand movement has meaning. There are additional elements that enable more efficient

and understandable communication like head movement, facial expression, mouth or body movement.

The Macedonian sign alphabet consists of 31 signs, the same number of letters in Macedonian alphabet. There are two versions of the alphabet: one version is signed using only one hand and the other is using both hands. Figure 1 shows the MSL alphabet signed with two hands.

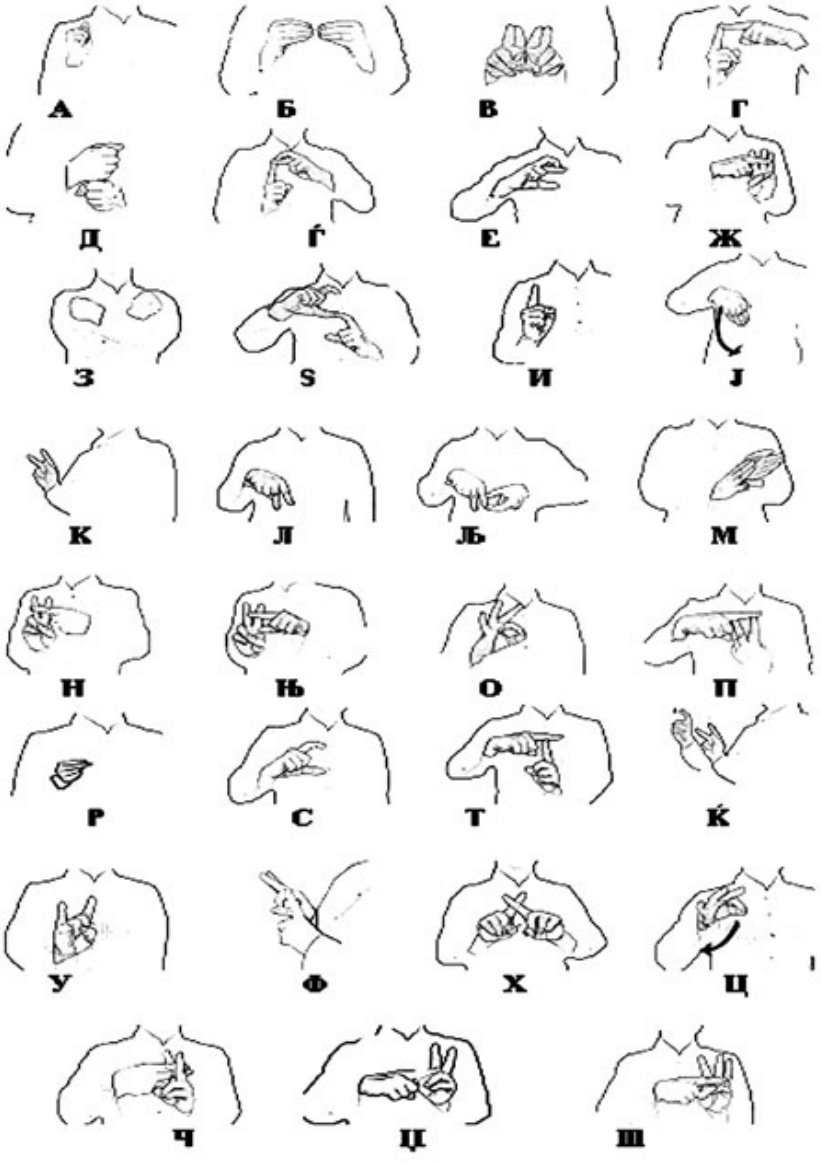


Fig. 1. The Macedonian sign language alphabet signed with two hands [8]

2 The technology possibilities and researches

The digital technology can reduce the gap between the people with physical and mental disabilities and can help the communication of deaf and hard of hearing people with the rest of the population. In the past 10 years there have been efforts in many different directions to improve the quality of life of deaf and hard of hearing.

The inclusion of this community in todays' online world has been considered for a long time. General accessibility guidelines for software application are proposed long time ago, in 1994 by Gregg C. Vanderheiden [9]. These are the general recommendations:

- Provide visual form of all auditory information
- Ensure that all visual cues are noticeable
- Offer an operation mode for noisy environments or if sound is turned off;
- Support ShowSounds feature if it exists in the operating system

There have been efforts to produce a general framework to support development of sign language Human-Computer Interaction [10] and to support sign language recognition and interaction. These researches are still in the begging phase.

2.1 Visual technologies

The visual technology has evolved. From the previous researches we can conclude that the products with visual interfaces are the most recommendable for communication with SL. A lot of researches are exploiting this technology in different ways.

One research trend is focused on improving *communication* using the visual technologies to translate sing language into text with devices like mobile phones, such as Mobile Motion Gesture Design for Deaf People [11]. Other research groups are working in the opposite direction: converting text [12] or audio into sign languages. Such examples are the applications that translate the audio signal into animated face so the deaf users can use lip-reading, like the Mobile Multimedia Application for Deaf Users [13] and the Lip Assistant [14].

Another direction of research concentrates on the learning process of the sign languages. Examples of this type of research are the System for Sign Language Tutoring [15] which evaluates users' signing and gives multimodal feedback to help improve signing, 3D Animation Editor and Display Sign Language System for Thai Sign Language [16] which exploits the XNA framework in order to train the general person and Deaf person with Thai Sign Language for ability to communicate to each other and SignOn - a Model for Teaching Written Language to Deaf People [17].

An interesting approach is the idea to develop as a prototype an augmented reality book for deaf students [18].

There are also commercial products on the market like the Vcom3D Sign Smith product which use avatar. They offer the products: Illustrated Dictionary - for learning

basic American Sign Language, Studio - to make your content sign language accessible, ASL Animations - to help students to learn English words by adding animated GIFs and MOVs to existing classroom software programs and Signing Science Dictionary - comprehensive science reference for Deaf students in grades 4-8. [19]. Figure 2 shows the signing avatar in the Illustrated Dictionary and generally shows the interface of this product. This product arranges the words both alphabetically and by category. Every word is represented by a picture, used in a sentence and signed. The Illustrated Dictionary includes the 500 signs most commonly included in beginning ASL courses. [19]

Fig. 2. The Vcom3D Illustrated Dictionary [19]

3 Sign Language Tutor

3.1 Sign Language Tutor general information and user interfaces

Sign Language Tutor is an interactive platform for MSL learning that is easily extendible. It represents a collection of games and modules that will ease the learning and increase mental and memory skills of the deaf and hard of hearing children. The central part of this project consists of 3D animations of a girl that signs the chosen alphabet character or some objects. Figure 3 shows our character performing signing.

Fig. 3. 3D animation of the girl signing

The application consists of several modules:

- 3D simulation – this module provides signing of alphabet letters and words. The user chooses the letter or the object and the 3D animation illustrates the chosen sign. The user can rotate the animation in all directions and enables different views to better capture the signing. This is of outmost importance, since the signs and words are actually expressed in 3D, usually with more than just fingers and hands. The animation can be paused and started again.
- Memory – this module represent a standard memory game where the goal is to merge an alphabet letter with the appropriate sign.
- Explore – this module is a 2D game where our hero goes in an adventure against monsters and collects objects. Each of the collected objects is an award – an animated 3D sign of that object in MSL. The game doesn't require high many skills and it is adapted for children. Figure 4 shows the first level of the Explore module.

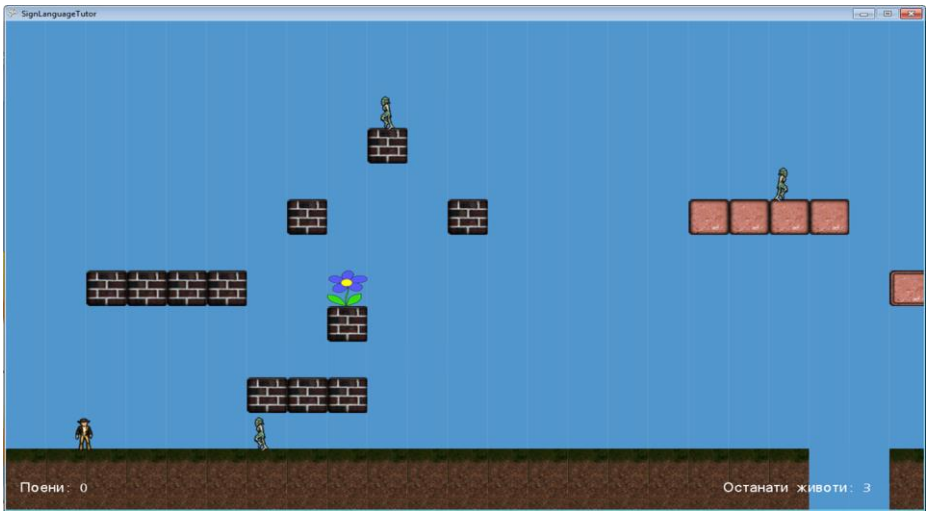


Fig. 4. The first level of the Explore module

3.2 Sign Language Tutor architecture and technical details

The Sign Language tutor is built from several software modules that are interconnected. This enables easier manipulation and expandability of the project.

The following sections will briefly describe the importance and the management of each of the modules.

3.3 SignLanguageTutor

This module along with the core part is the central part of the software solution and depends of all the other modules. It represents a Windows Game type of project that work on XNA 4.0 framework. It contains the concrete implementations of the games, and respectively the entire user interfaces.

The principle of work of this module is based on manipulation and management of the rendered views of the user interfaces. All the user screens are managed through the global object Screen Manager.

3.4 SignLanguageTutorContent

This module is responsible for the management of the resources. It represents a ContentProject type of project and it is used for resources compilation, storage and preparation for use during the execution. It is divided in parts responsible for management of models, backgrounds, textures, maps etc…

The part that manages modules contains all the 3D animation models. We use Autodesk FBX XNA importer as a specific resource to import the 3D models. The em-

bedded resource processor in XNA cannot be used because of the complexity of the 3D models (existence of animations based on skeleton movement attached to the model). That is why we have extended the embedded resource processor in XNA. The extension is represented in the SignLanguageTutor.AnimationProcessor module, described in the next section.

3.5 SignLanguageTutor.AnimationPipeline

This module is compiled to a dynamic link library. It contains the AnimationProcessor class that extends ModelProcessor. The processor contains two parts needed for texture processing: skeleton processing and animation processing part. This actions are called through the function Process that is inherited and replaced from the ModelProcessor.

First we process the skeleton attached to the model. In this process we track the skeleton bones and then we create a hierarchy of dependent bones. This is required for the movement of one animation. In this phase we transformed the model into parts that are included in one coordinate system. After that we import the model which represents the skin attached to the skeleton. For better and more real appearance we use material and textures. Finally, we process the animation. In this phase first we create a search table that will enable index-based search of the bone names in the model. After that we create an array of matrices to save the transformations of the bones. We recursively process the whole module graph of the scene and extract the data that contains animation. For each animation we process the keyframes and we define grouping of the keyframe bones. In the keyframe filtering we perform linear interpolation.

3.6 SignLanguageTutor.AnimationAux

The AnimationAux is an auxiliary module that includes animation processor connection. It contains entities for animation model storage, helpers for animation and models' binary formats loading directly in the source code.

3.7 SignLanguageTutor.TileEngine, LevelEditor and LevelEditorContent

This TileEngine module is used in the 2D game Explore. It represents a framework that defines all the basic 2D game functionalities like movement physics, game map and view. This is also popularly called game engine.

The LevelEditor and LevelEditorContent modules are in charge for game maps creation and storage. The map editor is also XNA Windows Game application.

4 Current work: application testing and improvement

At this moment we are working on getting more standard subjects in their natural environment in order to test the usability of the system. In Macedonia, since the

community of deaf and hard of hearing is much marginalized, it is very difficult to get to the children of this group. At this moment we are negotiating with the specialized institution for deaf and hard of hearing "Partenija Zografski" in order to test the system on a referent number of children

Our work on this project doesn't stop here. We have few goals for the future. Our first goal is to enrich the dictionary with more words and phrases. Another goal is to create a mobile version of our application for the popular platforms like Android and iPhone, and make the application closer to the users: the ones who are heard of hearing impairment and the ones that have no problems of this kind. We believe that it will enable the ordinary users to get more acquainted with the Macedonian Sign Language.

We would also like to explore the possibilities of automatically signing text in Macedonian language with usage of software patterns. For that purpose we have to get more acquainted to the MSL grammar.

5 Conclusion

We have seen that the scientists all around the world are working on the goal to improve the communication between the deaf and hearing people and they are trying to find out innovating approaches to reach their goal like signing tutorials, automatic convertors and avatars. In most of the develop countries the community of deaf and hard of hearing people are open to the society. They have material about their native signing language and the society pays attention to their inclusion by offering accessibility features.

Unfortunately in Macedonia the situation is far from the ideal. Poor information about MSL can be found only on the site of the National association of deaf and hard of hearing and the deaf children are isolated in only few schools. Out project offers bigger inclusion and increases the awareness of this social group by helping the deaf and hearing people to learn MSL.

The software is developed using the XNA framework which has proven to be a good platform for video games, offering wide spectrum of possibilities and fast export of the solution to any platform that supports XNA and .NET framework.

We have followed the recommendations given by research human-computer interaction with deaf children [20][21] about deaf children psychology and abilities and we created a product that deaf children will easily use and will offer education and fun combined. It will help the children or their non-deaf parent to improve SL communication skills.

6 References

1. World Health Organization. Deafness and hearing impairment. http://www.who.int/mediacentre/factsheets/fs300/en/index.html
2. World Federation of the Deaf. http://www.wfdeaf.org

3. Creighton, N.: What is Wrong With the Use of The Terms: 'Deaf-mute', 'Deaf and dumb', or 'Hearingimpaired', http://www.eamo.org/SNA/deaf%20PC%20terminology.pdf
4. Conrad, R.: The deaf schoolchild. London: Harper & Row. (1979)
5. Gallaudet Research Institute.: Literacy and deaf students. (2003) http://www.gallaudet.edu/gallaudet_research_institute/publications_and_presentations/literacy.html
6. Sears, A., Jacko, J.A.: Computer Interaction Handbook. Fundamentals, Evolving Technologies and Emerging Applications. Second Edition. Taylor and Francis Group (2008)
7. Alfa TV.: За 6.000 глуви, само 12 толкувачи на јазикот на знаците (2009) http://ww.alfa.mk/default.aspx?mId=36&eventId=13326
8. National association of deaf and hard of hearing of Macedonia, http://www.deafmkd.org.mk/
9. Vanderheiden, G.: Application Software Design Guidelines: Increasing the Accessibility of Application Software to People with Disabilities and Older Users. (1994), http://trace.wisc.edu/docs/software_guidelines/software.htm
10. Antunes, D.R., Guimaraes, C., Garcia, L.S., Oliveira, L.E.S, Fernandes, S.: A framework to support development of Sign Language human-computer interaction: Building tools for effective information access and inclusion of the deaf. In: Fifth International Conference on Research Challenges in Information Science, pp.1-12. (2011)
11. Xue, H., Qin. S.:Mobile motion gesture design for deaf people. In: 17th International Conference on Automation and Computing (ICAC), pp.46-50. (2011)
12. Grif, M., Demyanenko, A., Korolkova, O., Tsoy, Y.: Development of computer sign language translation technology for deaf people. In: 6th International Forum on Strategic Technology (IFOST), Volume 2, pp.674-677. (2011)
13. Tihanyi, A.: Mobile multimedia application for deaf users. In: ELMAR, pp. 179 182. (2007)
14. Xie, L., Wang, Y. and Liu, Z.-Q.: Lip Assistant: Visualize Speech for Hearing Impaired People in Multimedia Services. In: IEEE International Conference on Systems, Man and Cybernetics, Volume 5, pp.4331-4336. (2006)
15. Aran, O., Ari, I., Akarun, L., Sankur, B., Benoit, A., Caplier, A., Campr, P., Carrillo, A.H., Fanard, F.-X.: SignTutor: An Interactive System for Sign Language Tutoring. Multimedia, IEEE, Volume: 16, Issue: 1, pp. 81-93. (2009)
16. Ittisarn, P.; Toadithep, N.: 3D Animation Editor and Display Sign Language System case study: Thai Sign Language. In: 3rd IEEE International Conference on Computer Science and Information Technology (ICCSIT), Volume 4, pp.633-637. (2010)
17. Hilzensauer, M., Dotter, F.: “SignOn”, a model for teaching written language to deaf people. In: This paper appears in: IST-Africa Conference Proceedings, pp.1-8. (2011)
18. Zainuddin, N.M.M., Zaman, H.B., Ahmad, A.: A Participatory Design in Developing Prototype an Augmented Reality Book for Deaf Students. In: Second International Conference on Computer Research and Development, pp.400-404. (2010)
19. Vcom3D. http://www.vcom3d.com/signsmith.php
20. Zafrulla, Z., Brashear, H., Pei Yin, Presti, P., Starner, T., Hamilton, H.: American Sign Language Phrase Verification in an Educational Game for Deaf Children. In: 20th International Conference on Pattern Recognition (ICPR), pp. 3846-3849. (2010)
21. Alonso, F., de Antonio, A., Fuertes, J.L., Montes, C.: Teaching Communication Skills to Hearing-Impaired Children. In: Multimedia, IEEE, Volume 2, Issue 4, pp.5-67. (1995)

Smart Adventure: Context-aware Crowd-Sourcing Mobile Application

Marija Alagjozovska[1], Zlatko Andonovski[2] and Vladimir Trajkovikj[1]

[1] Faculty of Computer Science and Engineering, Skopje, Macedonia
marija_ala@hotmail.com trvlado@finki.ukim.mk
[2]CreationPal, Skopje, Macedonia
Zaltko.andonovski@sportypal.comi

Abstract. The biggest challenge in mobile computing is introducing new type of application that has the ability of adapting and exploiting the changing environment.
In this paper, a smart phone context-aware crowd-sourcing application SMART ADVENTURE will be presented, offering users context aware services. The main application contexts are location, time and phone orientation. According to users' location, previous usage and current date, application generates information about activities and possible threats. Congruent with the phone orientation, application displays different screen.
Motivation for developing this kind of an application lies in constant trend for improving human health by doing different kind of open-air activities. At the same time, the proposed solution represents a high level tourist guide.
The main contribution of our work is providing proof of concept of how crowd-sourcing can influence few different domains (health, ecology, human culture).
The proposed solution is service oriented, implemented in android environment.

Keywords. Crowd-sourcing, collaboration, mobile applications, geo mapping

1 Introduction

The evolution of mobile devices and their huge potential of mobile processing and services made mobile computing a field with many challenges. The biggest challenge of all is developing a new type of software that will exploit the changing environment [2]. This kind of software can be treated as an entity that adapts to its location, time, available hosts, accessible devices and other environmental states called contexts. But what is the meaning of context and how can be defined? The most accurate definition about what the word "context" really means seems to be the following: "any information that can be used to characterize the situation of the entities (person, place or object) that are considered relevant to the interaction between the user and the application, including the application and the user itself"[1]. From this definition it can be said that context is the set of environmental states and settings that either determines an application's behaviour or in which an application event occurs and is interesting to the user [3]. In other words, context is a piece of information that can be used to

characterize the situation of a participant in an interaction [2]. Context helps to e richness the communication in human-computer interaction and makes it possible to produce more useful computational services. The main three aspects that context has are: where are you, who are you with and resources nearby. But context is more than user's location; it encompasses lighting, noise level network connectivity, communication costs, communication bandwidth, and even the social situation. Software with this kind of features is characterized as context aware software. Thus, software is context aware if it uses context to provide relevant information and/or services to the user, where relevancy depends on the user's task [1].

Many of the applications use the context for providing useful data. Modelling this data, than use them in the context aware application is not a simple thing. Good modelling means choosing the right context model. Choosing it right, may depend of the architecture of the system and compatibility with the context model itself. There are many ways for modeling context information, but the following approaches are most accurate and most used by the developers.

- Key - value model
- Markup scheme models
- Graphical models
- Object oriented models
- Ontology based models
- Tree - like graph

For simplifying the process of developing a context aware application an abstract framework is needed. The framework will provide the client with the needed data and also will permit registration of new distributed heterogeneous data sources. There are several context frameworks available: SOCAM (Service-Oriented Context-Aware Middleware), CASS (Context-Awareness Sub-Structure), Hydrogen, WildCAT etc [4].

In this study, context aware crowd-sourcing application for smart phones is presented. The main application contexts are location [5], time and phone orientation. Location aware service represents one of the most crucial needs for users that own some kind of a smart phone [6] [7]. Its importance lies in necessity of finding location based information when it influence the overall behaviour of the system and might increase its performance. Location aware services can be used for finding travel information, shopping, entertainment, event information and other types of filtering [12]. In our application, location aware service filters a list of possible activities and threats in accordance with user position on the map. Positing is done by GPS receiver integrated in the phone [8]. GPS (Global positing system) represent a spaced-based satellite navigational system that provides time and location information, synchronous or asynchronous [9]. Location information based on GPS is from crucial meaning because it covers the user security issue (rescued if he is in danger) and it also renders events and places of interest (user can find the event or the place he is interested in) [10]. For more accurate data, time filter is implemented. The time filter does another filtering, optimizing the list on activities and threats that are specific for that period of time. The user's enjoyment level adjustment is done by implementing a phone orientation context. Depending on how phone is being held, different screens are displayed. If the phone is being held horizontal a map is generated, if it is being held vertical,

augmented reality is displayed, giving users enhanced information about their surroundings. Augmented reality (AR) represents a technology in which a user's view of the real world is enhanced or augmented with additional information generated by a computer. It combines real and virtual object in real environment, registers (aligns) real and virtual objects with each other; and runs interactively, in three dimensions, and in real time [20] [21]. Displaying different screen is done by the acceleration sensors (i.e., accelerometers) [24]. The accelerometer represents a device that measures the acceleration of the device on the x (lateral), y (longitudinal), and z (vertical) axes. Accelerometer is used in engineering, biology, industry, building and structural monitoring, medical applications, navigation, transport, vulcanology, consumer electronics (motion input, orientation sensing, image stabilization, device integrity, and gravimetry). In our application, the accelerometer has another function despite displaying the proper screen, it also detects fall. Fall detection is done by gathering and analyzing accelerometer data with several threshold based algorithms and position data to determine a fall. Activity of changing a phone orientation differs from activity of falling in a matter of time. Time of falling is longer than time of changing the phone orientation. When some falls it starts with free falling, causing acceleration's amplitude to drop significantly below the 1G threshold, representing the actual time of falling. For fall to be complete it must stop. This causes a spike in the graph. Then the amplitude is crossing an upper threshold which has the minimum value of around 3G, suggesting the fall. Serious injuries are represented as flat line in the graph [25] [22].

The proposed application can basically serve as a tourist guide [11], but it differs from other similar applications in terms of how it incorporates collaboration [13]. In our approach, users enter their one point of interest such as activity, animal or plant. Points of interest (POI) are entered as multimedia data by using the smart phone built in camera. User takes picture of his point of interest which is geo tagged (it has geographical coordinates), than enters its description in the application and makes it available to other users. This way, they are building a distributed collection of sharable items, making the application a crowd-sourcing system (CS). CS system enlists a crowd of users to explicitly collaborate to build a long-lasting artifact that is beneficial to the whole community. Since it enlists a crowd of users, this application face four key challenges: how to recruit contributors, what they can do, how to combine their contributions, and how to manage abuse [19]. In our approach, contributors are all users from all over the world that are keen on doing open air activities. With their contributions they can share activities specific for that region. The problem of abuse is solved by function that checks all words entered in description of the POI. If a vulgar word is detected, the POI is deleted, and never shared among users.

Depending on the context of point of interest entered, the same application can be used for environment saving, preserving national treasure, reawaking ecological conscience, visual navigation, sharing old fashion activities added by users from all over the world and for mobile learning (real time learning for local animal and plants).

SMART ADVENTURE supports two working modes: online and offline. Online mode differs from offline mode in terms of how application is used. When used online, application serves more like a nature navigator. For proper online working, GPS and IP mobile network are needed. When used offline, application serves more like a brochure for a specific region. Locations are downloaded and stored locally on

the phone memory with the first application installation and they are updated every time phone is connected to IP mobile network. This application represents a link that connects 21 century technology benefits and the old fashion way of doing activities in nature.

The main contribution of our work is to prove that by croudsorcing with geo mapping we can influence different domains that have few points of tangency (health, ecology, human culture) [15] [14] [16] [17].

In the second chapter we will give brief introduction of the system, giving a system overview (components of the system and its collaboration). The third chapter gives details of system implementation, while the forth chapter explains proposed application user interface and interaction. The final, fifth chapter concludes the paper.

2 System overview

The general architecture of the proposed application is given on **Fig.1.** : It is consists of smart phone with internet access, GPS, mobile network (GSM/WCDMA), web server that incorporates user data base.

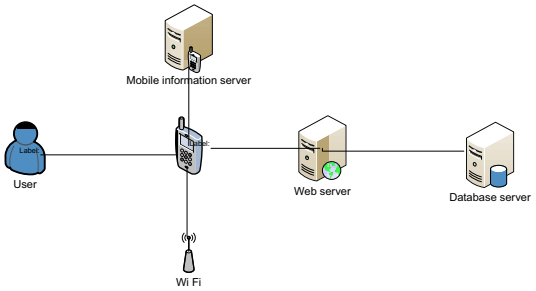


Fig. 1. System overview

Smart phone connects to an internet using mobile network and GPS. Web server handles the request for data from the user and gives him feedback. Requested data are stored on data base. When user asks for data, web server connects to data base and filters requested data. Web server is responsible for users support (member or non member), by monitoring their movement providing them with safety in every moment. The help request handles the mobile network by sending SMS alert to help centre and to predefined number (or any predefined web service). The number (web service) to whom the message will be send is set by the user. The help SMS for members is sent to both help centre and predefined number. Users without account (non members) can only send SMS to help centre.

Proposed application is implemented as a prototype on android platform. For accessing application features, users have to connect to internet using mobile network and GPS. When user is logged in he can add POIs. If user is not logged in, system treats users like non members, and provides them with few application features (finds the user position on the map which has been cashed earlier, and seeing points of interest which are saved locally on the phone). Integrated GPS in the phone sends data to the server, and server stores the data in its memory. If there is no internet access, mo-

bile network or GPS, application can be used in offline mode. When it is used in offline mode, it offers other features. Users can call for help and can see locations and points of interest that are already downloaded and stored locally on phone. In this manner, application is used more like a brochure for a specific region.

Smart adventure is defined as an alternative of classical tourist guide. It can be used as additional source of information for certain locations. The main idea of Smart Adventure design is that users with smart phones can log on or register from anywhere and get additional information for the place they are current. If there is no GPS and IP mobile network, they can inspect downloaded regions and learn about them. **Fig. 2.** represents the use case diagram of the system. A user chooses to locate or see locations. Then he logs in, skips or registers in the system. If the user is a member (user with account) he/she can see and add a point of interest (animal, plant, and activity), post on Facebook or Flicker, augment his surrounding or ask for help in case of potential danger. Otherwise if he/she is not a member he/she can just ask for help or see points of interest (animal, activity, plant). When adding points of interest, user can add a picture of the point (optional). If he/she chooses to see locations, a geotagged map is generated.

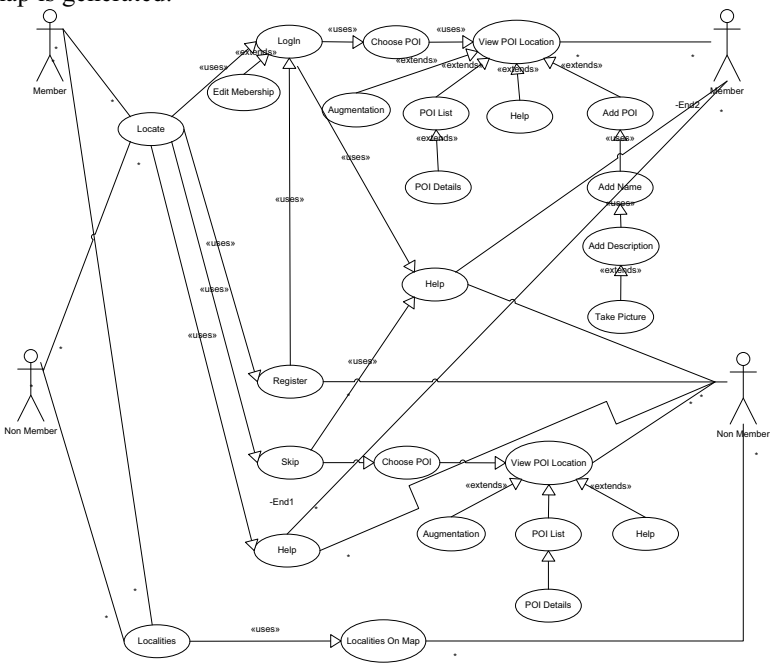


Fig. 2. Use case of the "Smart Adventure" system

3 Implementation details

The whole system is service-oriented and implemented in android environment using java for service development. It was developed in eclipse ID environment Android

1.6 SDK, and coded in java object oriented language. The application can be used only by user who has smart phones with android platform. Android represent a software stack for mobile devices that includes an operating system, middleware, and key applications. The Android SDK provides the tools and libraries necessary for developing applications that run on Android-powered devices [23] [24].

The system contains one package, "Adventure", in which classes are placed. The package contains two subpackages: "Data" and "Model". The "Data" sub-package contains classes that provide information about activities, animals, plants in accordance with their location. The "Model" sub-package contains all the classes that are not directly connected with the users' views (everything that has to present something real or abstract which existence is not direct connected to the application.) **Fig. 3.** represents the class diagram, in which the structure of the system represented by the classes, their attributes, operations and relationship among them is given. The given class diagram is truncated in order to reduce survey's fussiness. In the class diagram only the classes which have the main importance in system structure are presented.

When the application starts, the Main Activity is activated and the users can choose one of the offered system features. The Main class handles the user's choice and starts the LiveMap Activity which communicates with the LiveOverlay class. The LiveMap class determines the use of location on the map, and LiveOverlay displays icons on the map (animals, activities, plants). Our Activity class is the model class for all points of interest (animal, activity, plant); it contains all information related to POI's like coordinates, names, etc. This class also has a time filter that filters the POI's in accordance with the current system date and time. FakeDataProvider manipulates crucial data about users' points of interest (data for animals, activities, plants) in accordance with their location. It is used to provide fake data and enable us to test the application. It will later be replaced with a local database or external data source mediator.

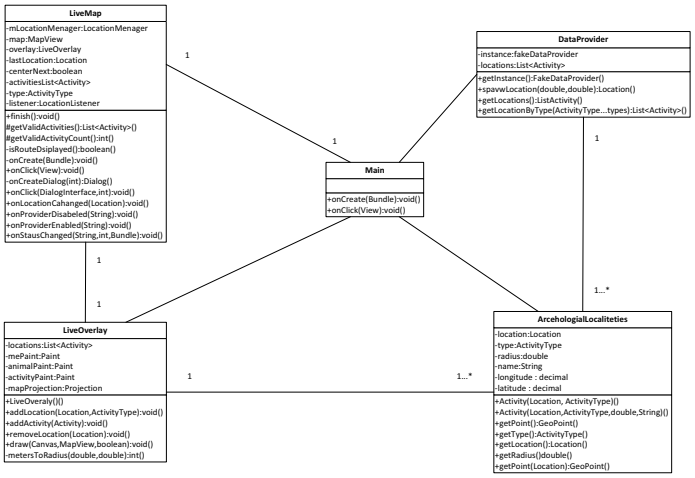


Fig. 3. Class diagram of the "Smart Adventure" system

4 User interface

Fig .4., Fig. 5., Fig. 6. and **Fig. 7.** represent the most important features of the application user interface.

Users, after they have skipped, registered or logged in the system, have to choose one of the offered application features (animal, activity, help, and plant) **Fig. 4.** The menu is organizes in this way for easy accessing. After choosing one of them, map opens and the user location is mark in it with bright blue colour dot **Fig. 5**. The list is generated by clicking the POI button positioned in the bottom. From the list user can choose one of the poi's and sees all the details **Fig. 7.** and its radius of coverage **Fig. 6.** represented by the big red circle. If the user is in any potential danger he just has to click and hold about five seconds (five second pres) the help button. The five second press is to prevent sending of false alarm. The position of the help button is set in the left and right corner for easy access for both left hander and right hander users.

For better understanding of application usage, we have chosen a concrete location, and for that location we have chosen a concrete animal – grass snake.

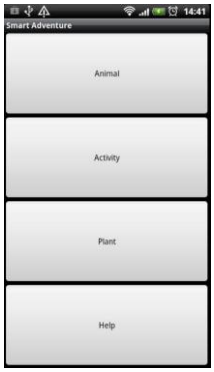


Fig. 4. Smart Adventure menu

Fig. 5. User located on map

Fig. 5. represents a located user in the nature and positioned on the map. User position is marked with bright blue dot. On this picture user is located in Skopje area, Kozle estate.

Fig. 6. Grass snake coverage

In Kozle estate (Skopje, Macedonia) grass snakes are the animals that user can come across. Grass snakes coverage is marked with big red circle **Fig. 6**.

Fig. 7. Description of grass snake

The description of the grass snake is given in **Fig. 7.** In the description are given detail information about the snake including a picture of it, so the user can recognized it very easy when he'll come across it.

5 Conclusion

In this study an Android application for smart-phones was presented.

The application is object oriented and it embeds crowd-sourcing and geo-mapping. It is context aware and main application contexts are location, time and phone orientation. For modeling the context data, the application uses an abstract framework which provides the client with the needed data and also permits registration of new distributed heterogeneous data sources. The issue of abuse control that

every crowd-sourcing has, was solved by service that checks users input and decides its correctness.

References

1. Dey, A.K., Abowd, G.D.: Towards a better understanding of context and contextawareness. In: Proceedings of the Workshop on the What, Who, Where, When and How of Context-Awareness, ACM Press, New York (2000)
2. Guanling Chen, David Kotz.: A survey of context aware mobile computing research. Tech. Rep TR2000-381, Dartmouth, November (2000)
3. Dey, A.K.: Understanding and using context. Personal and ubiquitous computing, Volume 7, Number 5, November (2001)
4. Trajkovik, V., Bakraceski, G.: Wildcast context-aware framework: case study for place selection. The 9th Conference for Informatics and Information Technology (CIIT 2012)
5. Raento, M., Oulasvirta, A., Petit, R., Toivonen, H.: Contex tPhone:A Prototyping Platform for Context-AwareMobile Applications. In: IEEE CS and IEEE ComSoc, May (2005)
6. Kaasinen, E.: User needs for location-aware mobile services. In: Personal and ubiquitous computing, Volume 7, Number 1,November (2002)
7. Aato, L., Gothlin, N., Korhonen, J., Ojala, T.: Bluetooth and WAP Push Based Location-Aware Mobile Advertising System. MobiSys'04 Proceeding of the 2nd international conference on Mobile systems, applications and services, June (2004)
8. Zaeipekis, V., Glaglis, M.G., Lekakos, G.: A Taxonomy of Indoor and Outdoor Positioning Techniques for Mobile Location Services. ACM SIGeocom Exchanges-Mobile commerce, Winter, (2003)
9. Djuknic1, G.M., Richton, R.E.: Geolocation and Assisted-GPS. In: IEEE Computer Society, August, (2002)
10. Malladi, R., Agrawal, D.P.: Current and FutureApplications of Mobile and Wireless Networks. Vol. 45, No. 10 communications of the ACM, October, (2002)
11. Mark van Setten, Pokraev, S., Koolwaaij, J.: Context-Aware Recommendations in the Mobile Tourist Application COMPASS. Adaptive Hypermedia and Adaptive Web-based System, Lecture Notes in Computer Science, (2004)
12. Baus, J., Cheverst, K., Kray, C.: A Survey of Map-based Mobile Guides. In: Meng, L., Reichenbacher, T. and Zipf, A. (Eds), Map-based mobile services - Theories, Methods, and Implementations, Springer-Verlag, (2004)
13. Oulasvirta, A., Raento, M., Tiitta, S.: ContextContacts: Re-Designing SmartPhone's Contact Book to Support Mobile Awareness and Collaboration. MobileHCI'05, Salzburg, Austria, September 19–22, (2005)
14. Parikh, T.S.: Using Mobile Phones for Secure, Distributed Document Processing in the Developing World. Pervasive Computing, IEE, May (2005)
15. Palmer, M.A., Bernhardt, E.S., Chornesky, E.A., Collins, S.L., Dobson, A.P., Duke, C.S., Gold, B.D., Jacobson, R., Kingsland, S., Kranz, R., Mappin1, M.J., Martinez, L., Micheli, F., Morse, J.L., Pace, M.L., Pascual, M., Palumbi, S., Reichman, O.J., Townsend, A., Turner, M.G.: 21st Century Vision and Action Plan for the Ecological Society of America. In: Frontiers in Ecology and The Environment, Volume 3, Issue 1, February (2005)
16. Li,C., Liu, L., Chen, S., ChenWu, C., Huang, C., Chen, X.: Mobile Healthcare Service System Using RFID. Networking, Sensing and Control, 2004 IEEE International Conference, September (2004)

17. Bellavista, P., Küpper, A., Helal, S.: Location-Based Services:Back to the Future. In: Published by the IEEE CS, (2008)
18. Motiwalla, L.F.: Mobile learning: A framework and evaluation. In: Computers and Education, Volume 49, Issue 3 November (2007)
19. Doan, A., Ramakhirnan, R., Halevy, A.Y.: Crowd-sourcing system on the World-Wide Web. In: ACM Press, Volume 54, Number 4, New York, April (2011)
20. Ronald T. Azuma, R.T.: A survey of Augmeted Reality. In: Teleoperators and Virtual Environments 6, August (1997)
21. Krevelen, D.W.F., Poelman, R.: A Survey of Augmented Reality Technologies, Applications and Limitations. The International Journal of Virtual Reality, (2010)
22. Brown, G.: An Accelerometer Based Fall Detector. Development, Experimentation, and Analysis, University of California, Berkeley, July (2005)
23. Google Inc. Android. www.android.com.
24. Google Inc. Android developers. developer.android.com
25. Tyson, G. , Sposaro, F.: iFall: An android application for fall monitoring and response. In: Engineering in Medicine and Biology Society, Annual International Conference of the IEEE, September (2009)

Efficient document retrieval using text clustering

Igor Trajkovski

Faculty of Computer Science and Engineering,
"Ss. Cyril and Methodius" University in Skopje,
Rugjer Boshkovikj 16, P.O. Box 393, 1000 Skopje, Macedonia
trajkovski@finki.ukim.mk

Abstract. Similar document retrieval is the problem of finding documents that are most similar to a given query document. In this work, we present a retrieval based on clustering of the documents that approximates the nearest neighbor search. It is done by determining the clusters that are most similar to the query document and restricting the search to the documents in these clusters. Cluster representation has an important role in the effectiveness of the search procedure, since the inclusion of a cluster in the restricted search space depends on whether its representation matches the query document. We analyse three cluster representations and their role in the performance of the proposed search procedure.

Keywords: similar document retrieval, text clustering, nearest neighbor search

1 Introduction

In this work, we address the problem of finding documents that are most similar to an input query document. Applications with this type of retrieval include: finding news articles that are similar to a given article, finding resumes that are similar to a given resume, finding patents that are similar to a specified patent, etc.

Here we present an overview of our retrieval system. We represent documents in high-dimensional vector space and employ the well-known nearest neighbor algorithm to find similar documents. For large document sets, we use the clustered search described in the next section to speed up the nearest neighbor search.

We group documents into units called clusters, and restrict the search within a claster. For example, we consider news article from the daily newspaper "Dnevnik" and look for similar documents in the daily newspaper "Vecer" collection.

The document internally is represented as a set of weighted keywords, computed by TF-IDF method. These keywords represent the most important words in the document. This set of weighted keywords is called *feature vector* of the document. The feature vector contains only the most important words that capture the essence of the document. The feature vectors are normalized to enable cosine similarity computations.

S. Markovski, M. Gusev (Editors): ICT Innovations 2012, Web Proceedings, ISSN 1857-7288

A search procedure is initiated when an input document D and a target collection C are presented. The first step in the search process is to retrieve the feature vector v of D. The next step is to compare the feature vector with the feature vectors of the other documents in the target collection C. The final step is to return the match results. The matching is based on the inner product or cosine similarity between feature vectors.

The straightforward nearest-neighbor based matching scheme compares the input vector with the feature vectors of all the documents in the collection. This can cause unacceptably slow searches for large collections, since it takes $\Theta(n)$ time, where n is the number of documents. Hence, we adopt a clustered search in which the documents in the collection are clustered in a pre-processing step, and the clusters are used in a retrieval phase to identify a subset of documents to be matched with the vector v. This scheme is a constant-time retrieval scheme, which is bounded by a parameter that specifies the maximum number of documents that can be compared with the input vector.

The subset of documents that is used in the retrieval scheme is identified as follows:

1. Compute the centroid of each cluster, where a centroid is a feature vector of unit length. The terms in the centroid are derived from the feature vectors of the documents in the cluster. Section 3 presents three ways of computing the centroid.
2. Select the clusters whose centroids are closest to v,
3. Consider only those documents that belong to the selected clusters. For example, if a collection has $100K$ documents and 300 clusters with roughly 300 documents in each cluster, we can select the *top* 20 clusters that best match v, and compare with these $20 * 300 = 6K$ documents only, as opposed to the entire set of $100K$ documents. The assumption is that the best matches from these $6K$ documents will be a good approximation to the matches obtained by examining the entire collection.

A key component of the clustered search is the centroid of each cluster. The choice of terms in the centroid is absolutely critical to obtaining good matches. The arithmetic mean of a cluster is the most popular and well-studied method of computing the centroid. In this paper, we propose two centroid computations that differ from the arithmetic mean in the way in which weights are assigned to individual terms in the centroid. Other representations for cluster centroids have been explored in [2], [3] and [4].

We use a goal-oriented evaluation method in which the goodness of a cluster centroid is based on the performance of the procedure employing this centroid approximates the nearest neighbor search over all documents in the target collection. Thus, centroid A is considered to outperform centroid B, if the same retrieval scheme employing A results in a closer approximation to the nearest neighbor search than while employing B. The evaluation is not based on the traditional measures of cluster goodness, like intra-cluster distance, inter-cluster distance and shape of the clusters. Instead it is based on the outcome of the similarity search that employs each of the representations.

The following sections discuss the clustered search in detail, how centroids are computed, our experiments and test results.

2 Clustered Search

The goal of the clustered search is to employ clustering as a means of reducing or pruning the search space. Instead of comparing with all the documents in the collection, the scheme allows a subset of the documents to be selected that have a high likelihood of matching the input document. There are two phases in this scheme: (a) a pre-processing phase in which the documents in a collection are clustered, and the centroids of the clusters are computed, and (b) a retrieval phase, in which the cluster centroids are used to retrieve similar documents.

2.1 Preprocessing

The well-known k-means clustering method [4], [5], [6] is used to cluster the documents in a collection. The number of clusters k is set to $\sqrt{n}$, where n is the number of documents in the collection. Initially, k documents are selected at random, and assigned to clusters numbered 1 through k. The feature vectors of these documents constitute the centroids of these clusters. The remaining $n - k$ documents are considered sequentially. Each document is assigned to the cluster whose centroid is closest to the input document (i.e., has the highest inner product similarity with this document). When all the documents have been examined and assigned to their closest clusters, the centroid of each cluster is recomputed (see Section 3.1). After re-computing the cluster centroids, the next iteration through all the documents is started. In this iteration, if a document is found to be closer to a cluster that is different from its current cluster, then it is removed from the current cluster and reassigned to the new cluster. Experiments suggested that 4 to 6 passes or iterations suffice to obtain good clusters.

2.2 Retrieval Phase

In the retrieval phase, a document D is received as input to the similar document search, and its feature vector v is matched with the centroids of the k clusters that were computed in the pre-processing phase. The clusters are sorted in descending order of similarity with the input vector v. The sorted list is traversed, and when a cluster is considered, all the documents that belong to it are compared with the input vector. The documents with the highest similarity seen thus far are accumulated in a results list. The traversal is stopped when the number of documents compared equals or exceeds a parameter, known as *max.comparisons*. For example, *max.comparisons* can be set to 5000 to indicate that no more than this number of documents should be compared with the input vector. This allows the similarity search to be completed in fixed time. Suppose the 10 clusters that are closest to the input document contain a total

of 5000 documents altogether, then the search is restricted to these top 10 clusters. The traversal of the sorted list of clusters is stopped after the documents in the 10-th cluster are examined. The results list of matched documents is then presented as the output of the similar document search.

3 Cluster Centroid Schemes

In this section, we describe how cluster centroids are computed. This centroid is used in both phases of the clustered search. In the pre-processing phase, it is used to assign documents to clusters, and in the retrieval phase, it is used to select clusters for further examination. The centroid is represented as a feature vector also, just like the documents in the collection. The question is: how to select the terms that constitute the centroid and how to compute their weights? The sections below present three different schemes for addressing this question.

3.1 Arithmetic Mean

The arithmetic mean of a cluster represents the average (or center of gravity) of all the documents in the cluster.

Consider a cluster with 1000 documents, each containing 25 terms in their feature vectors. Suppose that the number of unique terms over the entire space of 25 * 1000 terms is 5000. The arithmetic mean includes all 5000 terms, and the weight of a term is the average weight over all documents in the cluster. For example, if the term finance appears in 5 of the 1000 documents in this cluster, with weights 0.2, 0.3, 0.4, 0.1 and 0.8 respectively, the weight of this term in the centroid is (0.2 + 0.3 + 0.4 + 0.1 + 0.8) / 1000, which is 0.0018.

For large document collections characterized by high dimensional sparse vectors, the arithmetic mean has the following disadvantage: if a term is found only in a small fraction of the documents in a cluster, the weight of this term in the arithmetic mean is drastically reduced. The next two schemes overcome this limitation.

3.2 Maximum Weight

In this representation, the weight of a term in the cluster centroid is defined as the maximum weight of this term, over all occurrences of this term in the cluster. In the above example, for the term finance, the weight of this term in the centroid is 0.8. Thus, if a term has a high weight in any document in the cluster, its weight in the centroid is also high. This heuristic overcomes the weight reduction problem caused by averaging, but has the opposite effect of boosting weights, even if there is insufficient evidence to warrant the boosting. The next scheme addresses this problem.

3.3 Penalty Weight

In this scheme, if a term occurs infrequently in a cluster, it is penalized, but the reduction is not as steep as in the arithmetic mean. The weight of a term in the centroid is given by

$$max_w * p^m$$

where max_w is the maximum weight of this term over all occurrences of this term in the cluster, p is a number < 1.0 and m is the number of documents in the cluster that did not contain this term. In the above example, the maximum weight of the term finance is 0.8; m is $1000 - 5 = 995$; if p is 0.9999, the weight of this term in the centroid is $0.8 * 0.9999^{995}$, which is 0.7242.

The centroids obtained from all three methods are truncated to include the 200 terms with the highest weights. This is the so-called limited feature representation of the centroid. The truncation is performed to reduce the time spent in comparing the input document with the cluster centroids (see [4] for a discussion on truncating features during clustering). The number 200 was obtained after experimentation that indicated that increasing this number beyond 200 does not result in improved search performance.

4 Experiments

The primary objective of the tests was the comparison of the restricted search involving a subset of the documents with the baseline search involving the entire collection. Experiments were conducted by utilizing the three cluster representations described above.

4.1 Data Set

We used a document collection consisting of 1 million news articles. These articles were crawled from the archives of all newspapers and TV stations in Macedonia. The archives were from the period $1998 - 2010$.

4.2 Pre-Processing

We compute the feature vectors of all the documents, and cluster them. The feature vector length for documents is fixed at 25. The arithmetic mean centroid representation is used during the clustering.

4.3 Baseline Search

A set of 1000 input documents is selected at random from the collection. For each input document, the following steps are performed:

1. Similar documents are retrieved from the collection by performing a baseline search (or nearest neighbor search) through all the documents in the collection.
2. The results are sorted in decreasing order of similarity.
3. The top 20 results are recorded. Since we restrict our evaluation to the 20 documents that are most similar to the input document, it suffices to store these results only.

Let

$$F = \{D_{f_1}, D_{f_2}, D_{f_3}, ..., D_{f_i}, ..., D_{fm}\}$$

be the result set for a given input document. The top 3 matches for this document is given by

$$F(top\ 3) = \{D_{f_1}, D_{f_2}, D_{f_3}\}$$

.

In general, we can define $F(top\ x)$, where x is the number of results being considered. The tables shown in the next section present results related to $F(top\ 3)$, $F(top\ 10)$ and $F(top\ 20)$.

4.4 Clustered Search

We fix a cluster representation, for example, the Maximum Weight representation for the cluster centroid. The centroids of all the clusters are recomputed using this scheme. The clustered search described in Section 2.1 is performed for the same set of 1000 documents that was used in the baseline search. While performing the search, the parameter max. comparisons is used (for example, 10,000), to truncate the search (see Section 2.1 for more details).

For this cluster representation and value of max. comparisons, the results list for a given document is

$$C = \{D_{c_1}, D_{c_2}, D_{c_3}, ..., D_{c_i}, ..., D_{c_m}\}$$

Similar to the baseline search we have

$$C(top\ 3) = \{D_{c_1}, D_{c_2}, D_{c_3}\}$$

to include the 3 documents that are most similar to the input document.

We define a *precision* that determines the degree to which the clustered search results agreed with the baseline search. In other words, *precision* measures the extent to which the clustered search mimics the baseline search. Since the clustered search makes fewer comparisons than the baseline search, it is expected that the results will degrade. The *precision* measures this degree of degradation.

4.5 Precision

The *precision* is defined as the intersection of the results from the baseline search and the clustered search, for a given input document, a cluster representation, a value of max. comparisons ($mxCmp$), and a $topx$ level, where x can be 3, 10, 20 etc. The level restricts the results sets to include the first 3, 10 or 20 documents in the sorted results list.

$$Precision(mxCmp, topx) = \frac{|F(topx) \cap C(topx)|}{|F(topx)|} \times 100\ \%$$

The *precision* for a single document is averaged over all the input documents (in our case, 1000).

The tests are performed for all three cluster representations, for 3 different comparison bounds: 10,000 (1%), 30,000 (3%) and 100,000 (10% of the documents in the collection).

The *precision* averaged over the input documents for the top 3, top 10 and top 20 results levels are computed.

5 Results and Discussion

Table 1 shows the results of the test conducted using the arithmetic mean as the cluster representative for clustered search. The entry in the first row, first column, with a value of 74.7 indicates that on average, 74.7% of the top 3 results from the baseline search were among the top 3 results in the clustered search, when the number of comparisons allowed is 10,000.

Table 1. Average precision for arithmetic mean centroid representation

MaxCmp	Top 3	Top 10	Top 20
10,000	74.7	75.1	75.2
30,000	83.1	82.9	82.4
100,000	92.5	92.3	91.8

As expected, the results show that the *precision* increases as the number of documents compared increase. Also a noticeable characteristic is that the *precision* is fairly constant over the number of documents in the retrieved set i.e., *precision* calculated for the top 3, top 10 and top 20 documents considered.

Table 2 shows the results of the test conducted on the data set using the maximum weight centroid representation scheme described in section 3.2.

The results of maximum weight centroid representation show that the *precision* is higher than for the arithmetic mean centroid representation (Table 1). This indicates that the clustered search using maximum weight centroid representation is more effective in approximating the nearest neighbor search. However,

Table 2. Average precision for maximum weight centroid representation

MaxCmp	Top 3	Top 10	Top 20
10,000	89.1	85.4	82.2
30,000	92.3	88.7	87.7
100,000	97.9	96.1	95.3

one observation can be made. The average *precision* rates decreases as the size of the retrieved document set increases. The *precision* for top 20 is lower than for the top 3 retrieved documents considered for calculation of the *precision*.

Table 3 show the test results for the penalty weight centroid representation described in section 3.3.

Table 3. Average precision for penalty weight centroid representation

MaxCmp	Top 3	Top 10	Top 20
10,000	92.5	87.0	85.2
30,000	96.9	94.3	91.7
100,000	99.2	98.1	97.9

Let us consider the following questions:

1. When can a document appear in the results list for baseline search only, and
2. Can the relative order of results be different for baseline and clustered searches?

The first situation can happen in one of two cases: (a) The document D was included in a cluster C, and the centroid of this cluster did not match the input document, or (b) The centroid matched, but the inner product value was so low, that by the time this cluster was considered, the max. comparisons value was exhausted. Recall that in the clustered search, the clusters are traversed in sorted order and the traversal ends when the allowed number of comparisons is exhausted.

The second situation cannot happen since the same inner product measure is used in both types of searches, to match the input document with a document from a target collection.

6 Conclusions

We have studied three different cluster representations and their impact on similar documents search. The results indicate that the alternatives to the arithmetic mean centroid representation presented in this paper are very effective in approximating nearest neighbor search, even when the number of documents being

examined is less than 1% of the total. The reduction in the average *precision* for higher retrieval set sizes (top 10, top 20) as compared to values for top 3 needs to be explored further.

References

1. Witten I.H., Moffat. A. and Bell T.C., Managing Gigabytes - Compression and Indexing of Documents and Images (1999)
2. Bhatia S.K., Sanjiv K. and Deogun J.S., Cluster Characterization in Information Retrieval. ACM-SAC Indiana USA, 721-727. (1993)
3. Croft, W.Bruce. On the Implementation of some Models of Document Retrieval. ACM SIGIR, 71-77 (1977)
4. Jain, A.K., Murty, M.N. and Flynn. P.J. Data Clustering: A Review. ACM Computing Surverys (CSUR), Vol 31, No. 3, 264 - 323 (1999)
5. Duda R., Hart P., Stork D. G., Pattern Recognition, Wiley-Interscience; 2 edition; New York. (2000)
6. Thordoridis S., Koutroumbas K., Pattern Recognition, Academic Press. (2008)

S. Markovski, M. Gusev (Editors): ICT Innovations 2012, Web Proceedings, ISSN 1857-7288

ICT Innovations 2012 Web Proceedings ISSN 1857-7288 133

Linked Data-Based Social Bookmarking and Recommender System

Vladimir Apostolski[1] , Ljupco Jovanoski[1] , Dimitar Trajanov[1]

[1]Faculty of Computer Science and Engineering - Karpoš II bb, PO Box 574, 1000 Skopje, Macedonia
{vladimir.apostolski, ljupco.jovanoski, dimitar.trajanov}@gmail.com

Abstract. Social Bookmarking services have spread over the last few years. People often use tagging to organize and share their bookmarks. But this process can also be overloading the users and recommender systems are a popular approach to address this issue. Our goal was to explore the potential of services for semantic annotation and entity extraction to assist tagging and generate recommendations in social bookmarking communities. For that purpose, we have built a prototype of a Linked Data-based recommender and social bookmarking system. The system uses Zemanta to generate semantic tags and later the tags are the basis upon which recommendations are calculated. After that, a set of people used the application, gave feedback and evaluated the recommendations the system generated. In addition, we give a proposal of how tag connections to Linked Data entities expressed with MOAT ontology represent boost reusability and interoperability of gathered information

Keywords: Social bookmarking, semantic web, semantic annotation, web services, data, information, sharing, entity extraction, Linked Data, collaboration

1 Introduction

Social bookmarking is a concept that has gained in popularity in the last few years. Due to the rapid expansion of the Web, and share web resources they found relevant. But the popular social bookmarking services easily became overflowed with bookmarks, making it difficult to navigate through them. A partial solution was adding meta-data to the stored web resources, known as tagging.

Tags represent a flexible and popular way to describe and share web resources in communities on the Web. They are simple to use and can connect resources across different categories. Despite that, due to tags' increasing number, users are prone to information overload and counterproductive effects. These characteristics make recommender systems a candidate for filtering and discovering relevant content in social bookmarking environments. But problems in social tagging systems are also yielded to recommender systems that use tags in order to make recommendations.

As denoted in [1][2], collaborative tagging systems have downsides such as defining vocabulary, multiple tags with same meaning, misspelling and ambiguity.

S. Markovski, M. Gusev (Editors): ICT Innovations 2012, Web Proceedings, ISSN 1857-7288

S. Markovski, M. Gusev (Editors): ICT Innovations 2012, Web Proceedings, ISSN 1857-7288
© ICT ACT – http://ictinnovations.org/2012, 2012

134 ICT Innovations 2012 Web Proceedings ISSN 1857-7288

There are efforts to correlate tags between different social online communities, but with limited success because of the lack of meaning of the provided tags [3][4].

Consequently, these tag correlations are difficult to provide personalized recommendations based on the meaning of the tags. Furthermore, in [5] Alag states that the collective set of terms or tags in an application defines the vocabulary for the application. When this same vocabulary is used to describe both the user and the items, we can compute the similarity of items with other items and the similarity of the item to the user's metadata to find content that's relevant to the user.

The goal is to propose a way of implementing a semantic tag-based recommender system in social bookmarking communities that will not suffer from the issues of social tagging with reusable vocabularies. This paper covers three aspects: (1) How can services for semantic annotation assist tagging in social bookmarking communities, (2) how can the semantic tags generated by these services be used to make relevant recommendations to the users and (3) how linking semantic tags with Linked Data concepts can improve interoperability and reusability of gathered information.

1.1 Related work

There are related systems that include work on recommender systems like the work of Marinhoet al. [6] who find that tags can be used in order to make recommendations for users and web pages. Song [7] shares the same aspect towards recommendation types, but applies machine learning techniques to automate the process. Algorithms like SocialRank [8] are proposed to overcome inefficiencies caused by typical problems in social tagging systems. It ranks the recommendations based on the inferred semantic distance of the query to the tags weighted by the similarity of the querying user and the user who created the tags. In [9], data mining approach is used to merge tags in hierarchical clusters, hence dealing with tag redundancy. But these approaches lack the reusability of tags among different communities.

On the other hand, Semantic Web technologies are also proposed as candidates for recommender systems that could overcome tagging problems and make data reusable. The semantic recommender systems, are described in [10]. Their anatomy is described by Dell'Aglio [11], as well as their major drawback: modeling, building and maintenance of the knowledge base. She also emphasizes the usage of the Linked Open Data cloud to partially solve the problem of maintenance and data interoperability. Authors of ConTag [12], claim that services for semantic annotation and entity extraction can provide relevant tag recommendations for documents, solving the tagging problems. It uses services like Yahoo Term Extraction service to extract entities ,WordNet and custom DefTag service to bind them with meanings and align them to an ontology. Other such services that extract Linked Data entities are reported to be applied in a variety of applications. For example, OpenCalais[1], has been used in Tell Me More, an application that given an input text, mines the web for similar stories to extract entities [13]. Also OpenCalais is reported to make alerts to Semantic Me-

[1] http://opencalais.com

S. Markovski, M. Gusev (Editors): ICT Innovations 2012, Web Proceedings, ISSN 1857-7288

S. Markovski, M. Gusev (Editors): ICT Innovations 2012, Web Proceedings, ISSN 1857-7288
© ICT ACT – http://ictinnovations.org/2012, 2012

ICT Innovations 2012 Web Proceedings ISSN 1857-7288 135

diaWiki administrators when certain entities occur in RSS feeds [14]. Zemanta[2] is reported to be used in a social bookmarking platform with semantic tags, bound to Wikipedia articles, called Faviki [15][16]. NERD is an API that unifies results from various entity extracting services in one ontology and classifies the content according to it [17]. Its authors also make a feature comparison matrix for existing services and also adopt the idea to link extracted entities to Linked Data concepts [18]. Similarly, Halb and team [19] focus on the business potential of linking entities to Linked Data cloud in enriching content for general news articles. Moreover, TagMe! [20] is an application that attempts to assign DbPedia URIs to tags on Flickr images, so that external information could be extracted. All this research outlines the potential and need for connecting entities to Linked Data concepts.

1.2 Our Contributions

Having the related work in mind, we built a prototype of Linked Data-based social bookmarking and recommender application that integrates Zemanta, using the retrieved entities as (semantic) tag recommendations basis. This way we addressed both traditional tagging problems, as well as data reusability. Among the reasons we took this approach of using semantic annotation services for social bookmarking are: automated entity extraction and disambiguation, reduced amount of tedious work from user perspective, reusability and tag data openness in a lightweight manner. We estimated the extent to which these semantic tags can be used to provide relevant personalized recommendations to the users for bookmarking pages, help them discover users with similar bookmarking habits as well. The system is built under two assumptions: (1) services for semantic annotation and entity extraction are treated as black box; (2) Bookmarked pages are articles from different topics or user generated content in English. The recommendations are content-based. We conducted an experiment with real users to evaluate the recommendations by providing positive or negative feedback.

Tagging issues like synonyms, misspellings and defining contextual meaning were addressed by the services for semantic annotation and entity extraction behind the scenes. Tags' meanings were expressed with resources from popular sources like DbPedia and Freebase. Tags were saved for each bookmarked web page and they were treated as "similarity" factor between them. We built term vectors for each bookmark (vector of tags and their associated weights) and calculated cosine similarity, which is determined by the dot product of term vectors.

We also explored how Zemanta can be used to give relevant recommendations for external resources i.e. web pages that are not bookmarked by any user yet (external recommendations). We chose Zemanta based on researches and conclusions of other authors; for example Faviki, NERD's and other comparisons [21][22][23].

[2] http://zemanta.com

S. Markovski, M. Gusev (Editors): ICT Innovations 2012, Web Proceedings, ISSN 1857-7288

S. Markovski, M. Gusev (Editors): ICT Innovations 2012, Web Proceedings, ISSN 1857-7288
© ICT ACT – http://ictinnovations.org/2012, 2012

136 ICT Innovations 2012 Web Proceedings ISSN 1857-7288

Also we found that a benefit of using service-generated semantic tags that point to Linked Data entities, not only ease traditional tagging problems, but opens space for other semantic applications to leverage that data. That makes data re-usable and interoperable with other systems. To illustrate this, we used ontology for semantic tags to express the relationships between the tags, users and bookmarked web pages with the Linked Data cloud.

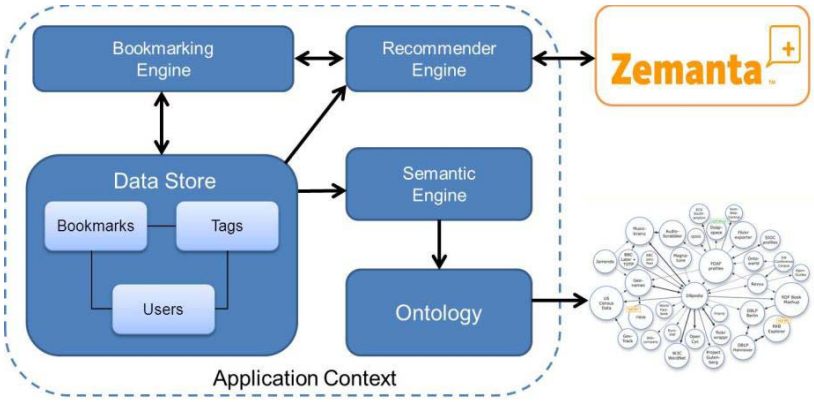


Fig. 1. System architecture Overview

2 System Overview

2.1 Architecture

The architecture of the system is given on **Fig. 1**. The main stored concepts are tags, bookmarks and users. For each user, it is known which pages are bookmarked, along with the tags used. There are three engines in the system: bookmarking, recommender and semantic engine. The bookmarking engine enables users to create, store and tag bookmarks. The recommender engine provides recommendations. The bookmarking engine retrieves tag recommendations provided by the recommender engine. For that purpose, the recommender engine interacts with Zemanta, the service for entity extraction. For other recommendation types, it uses the stored data in the system. The semantic engine transforms data to machine-readable format. It establishes the connections to the Linked Data cloud.

Bookmarking Engine.
Users interact with the bookmarking engine when they want to bookmark a web page via a bookmarklet. When a web page is selected for bookmarking by the user, the bookmarking engine asks the recommender engine to suggest tags to the user. After the user chooses the tags to describe the web page she wants to bookmark, the bookmarking engine stores the bookmarked web page along with the associated tags.

S. Markovski, M. Gusev (Editors): ICT Innovations 2012, Web Proceedings, ISSN 1857-7288
© ICT ACT – http://ictinnovations.org/2012, 2012

ICT Innovations 2012 Web Proceedings ISSN 1857-7288 137

Recommender Engine.
In our system, there are three types of recommendations: for tags, bookmarks and users. Tag recommendations are suggested tags when the user wants to bookmark a page. Bookmark recommendations help users discover web pages that might interest them and user recommendations refer to suggesting other users with similar interests. We will show how each recommendation type is generated in the following subsections:

Recommendations of Tags.
When the user wants to bookmark a web page, tag recommendations are automatically generated. A HTTP request is made to the bookmarking web page and its HTML response is passed to the service for semantic annotation and entity extraction. The semantic annotation and entity extraction service (Zemanta) processes the request and the returned results are parsed. Entities returned are treated as tags with description, meaning, relative importance towards the submitted URL (in range [0,1]).We use this weight to calculate term vectors in recommendations for bookmarks and users. In cases where the user inputs a tag, its relative weight is set to 1, because we consider that tag to be specifically relevant for the user.

Recommendations of Bookmarks.
Bookmark recommendations represent a feature to offer additional web pages for bookmarking to a given user. They can be internal or external. Internal bookmark recommendations refer to web pages that are already present in the system (meaning another user already bookmarked the same page). Internal bookmark recommendations can be personalized and non-personalized. Personalized internal recommendations are calculated based on the cosine similarity of the term vectors of the bookmarks and the term vector of the logged-in user. The term vector of the bookmark consists of the weight (W) of each tag that it is tagged with. Each tag represents a dimension in the vector. We adapted the method proposed in [24] to our model. The term vectors for bookmarks are represented in **Table 1**. m is the number of bookmarks and n is the number of tags in the system. Weight values assigned are the relative importance values while originally the author uses the number of times a tag is used for a given bookmark.

Table 1. Calculation of term vectors for bookmarks based on tag weights.

	Tag_1	Tag_2	Tag_3	…	Tag_n
$Bookmark_1$	W_{11}/M_1	W_{12}/M_1	W_{13}/M_1	...	W_{1n}/M_1
$Bookamrk_2$	W_{21}/M_2	W_{22}/M_2	W_{23}/M_2	...	W_{2n}/M_2
…	...	...	...	...	...
$Bookmark_m$	W_{m1}/M_m	W_{m2}/M_m	W_{m3}/M_m	...	W_{mn}/M_m

A normalization magnitude is calculated for each $vector_i$ in range $i \in [1, m]$:

$$M_i = \sqrt{\sum_{j=1}^{n} W_{ij}^2} \quad (1)$$

S. Markovski, M. Gusev (Editors): ICT Innovations 2012, Web Proceedings, ISSN 1857-7288

S. Markovski, M. Gusev (Editors): ICT Innovations 2012, Web Proceedings, ISSN 1857-7288
© ICT ACT – http://ictinnovations.org/2012, 2012

138 ICT Innovations 2012 Web Proceedings ISSN 1857-7288

Then, all weights are normalized to a vector with magnitude equal to 1. After that, we compute the term vector for the currently logged-in user as in **Table 2**:

Table 2. Computing Term vector for the user that is logged in

	Tag_1	Tag_2	Tag_3	…	Tag_n
User	W_1/M	W_2/M	W_3/M	...	W_n/M

Each of the weights ($W_1 ... W_n$) represent the number of times that user has used that tag when bookmarking a page.
We calculate the normalization magnitude:

$$M = \sqrt{\sum_{i=1}^{n} W_i^2} \quad (2)$$

In order to generate the recommendations, we calculate the dot products ($P_1..P_m$) between each bookmark term vector and the user term vector **Table 3**:

Table 3. Calculated dot products ($P_1..P_m$) between term vectors for bookmarks and term vector for the current user

	$Bookmark_1$	$Bookmark_2$	$Bookmark_3$	…	$Bookmark_m$
User	P_1	P_2	P_3	…	P_m

Recommendations are sorted in descending order by the dot product P_i where $i \in [1, m]$. Of course, pages that are already bookmarked by the user are filtered out and do not belong to the recommendations.

Non-personalized internal recommendations occur when the user inspects the details of a given bookmarked page. The process is the same except that the dot product is calculated between term vectors for bookmarks.

When the user is inspecting the details of the bookmark, recommendations for similar bookmarks (ordered by the dot product values) appear to the user. Of course, the similarity between the bookmark itself is 1 and is not included in the list of recommendations.

External recommendations refer to web pages that are not already present in the system. They represent a way to introduce new bookmarks in the system. External recommendations leverage Zemanta's suggest[3] method to get web pages that are not bookmarked yet, but might be of interest to the users. When analyzing content, this method returns related articles which it aggregates from various sources.

Recommendations of People.
To extend the social aspects of the application, it recommends other users who could have similar interests. Like for the bookmarks, the calculation of these recommendations consists of constructing term vectors for all users, normalizing them and then calculating the dot product between them. In this case, the weight for each tag represents the number of times a user has used that tag.

[3]http://developer.zemanta.com/docs/suggest/

S. Markovski, M. Gusev (Editors): ICT Innovations 2012, Web Proceedings, ISSN 1857-7288
© ICT ACT – http://ictinnovations.org/2012, 2012

ICT Innovations 2012 Web Proceedings ISSN 1857-7288 139

Semantic Engine.
As outlined in the introductory part of this paper, one of the problems with tags in online communities is the ambiguity and difficulty of their re-usability and interoperability of tag data with other systems or communities. This piece of the application covers those aspects. In order to make the tags and bookmarks externally accessible, we used the MOAT ontology[4]. The MOAT ontology is chosen because it is easily extendable to bridge the gap between tagging and Linked Data, according to Passant and Laublet [25] and this is the lastly developed ontology and suitable for modeling tags, meanings and concepts [26].

The semantic engine of our system exposes data in OWL 2 XML format with the MOAT ontology to enable external interaction with the data. It is available via URL endpoint.

Disambiguated tags have at least one meaning accompanied with URI to Linked Data cloud.

The service for semantic annotation and entity extraction we used includes entity URIs from sources like: DBpedia, Freebase, LinkedIMDB, MusicBrainz etc. We can dereference the URIs of the semantic tags to pull data from other applications (for example extracting information from Wikipedia). From that, one can tell what entities is a web page about, rather than raw unstructured text tags.

Due to openness, external applications can extract tag data that is present in our system. For example, external application can query for web pages tagged with Paris, defined as resource with the following URI: http://dbpedia.org/resource/Paris . Since this is a global URI, both applications can identify the exact resource they refer to, rendering the data interoperable between them. This also opens space for data translation engines like Mosto and LDIF to leverage the data with other vocabularies [27].

3 Evaluation and Results

A set of 41 users was asked to bookmark the sites they find relevant for a period of 7 days. The users were free to bookmark any content that is an article, user-generated content that contains text and that is written in English. The assumption is also that all users behave as instructed. The point is for each user to give to rate the recommendations that are provided. Every time the user logs in, he is asked by the system to rate the recommendations he got with positive or negative mark. That applied to all types of recommendations. We considered each giving of a mark to be one feedback. After that we collected the results and got the following facts:

[4]http://moat-project.org/ontology

S. Markovski, M. Gusev (Editors): ICT Innovations 2012, Web Proceedings, ISSN 1857-7288

S. Markovski, M. Gusev (Editors): ICT Innovations 2012, Web Proceedings, ISSN 1857-7288
© ICT ACT – http://ictinnovations.org/2012, 2012

140 ICT Innovations 2012 Web Proceedings ISSN 1857-7288

Table 4. Facts gathered from the evaluation of the recommendations

Total number of users	41
Number of bookmarked pages	97
Total number of tags	951
Tags connected to Linked Data entities	859
User-generated tags	69
Total Feedbacks	171
Positive feedback for non-personalized internal bookmark recommendations	82%
Positive feedback for personalized internal bookmark recommendations	84%
Positive feedback for external bookmark recommendations	78%
Positive feedback for recommendations for users	81%

Based on the facts given in **Table 4**, we can report that users gladly accept the service-generated tags (7% of the tags are user-generated). That indicates that existing tags were either reused or accepted the tags that the service generated. Because of this, common tagging problems like ambiguities are less likely to occur. Regarding the semantics of the tags and their interconnectivity with the Linked Data cloud, we notice that approximately 90% of the tags are accompanied by DbPedia and/or Freebase resources. This clearly shows that the tags and other data can be reused in other applications and vice versa – external information can be added to the existing application. Also, rules for addressing synonyms in the set of tags can be applied. For example, tags linked to the same Linked Data resources but with different labels should be treated as equivalent.

Regarding the feedback provided from the users: the average acceptance rate of 83% shows that semantic tags are a solid base for calculating recommendations for bookmarking pages. Results also show that the web service for entity extraction and semantic annotation helped the users discover relevant content in 78% of the cases, which is significant. Also users consider recommendations for similar users to be relevant in 81% of the cases, so that is a strong indicator of quality.
Based on the conducted evaluation, one can summarize that web services for entity extraction are mature enough to be used in recommender systems for social bookmarking.

4 Conclusion

In this paper, we showed how semantic tags generated by web services for entity extraction can improve tagging in social bookmarking environments. In addition, we proved the potential of those tags for making relevant recommendations to the users. Lastly, we propose means of leveraging connections of tags to popular Linked Data hubs, such as DbPedia.

The results we got lead us to the conclusion that : (1) services for semantic annotation and entity extraction generate relevant tag suggestions. (2) Recommendations

S. Markovski, M. Gusev (Editors): ICT Innovations 2012, Web Proceedings, ISSN 1857-7288
© ICT ACT – http://ictinnovations.org/2012, 2012

ICT Innovations 2012 Web Proceedings ISSN 1857-7288 141

based on that tags users also find relevant. (3) Using ontologies like MOAT to represent URIs to popular vocabularies like DbPedia can help overcoming tagging problems like disambiguity and reusability in different systems.

Future work may include leveraging the semantic tags for collaborative-filtering approaches when generating recommendations. Furthermore, we could explore the possible benefits of combining results from multiple web services like in the NERD platform. Also our team is thinking towards addressing common problems in recommender systems like the cold-start problem by using the links to the Linked Data cloud.

5 References

1. Braun S., Schora C., and Zacharias V. Semantics to Bookmarks: A Review of Social Semantic Bookmarking Systems, 5th International Conference on Semantic Systems. (2009)
2. Majid, A.; Khusro, S.; Rauf, A. Semantics in social tagging systems: A review, Computer Networks and Information Technology (ICCNIT), International Conference (2011)
3. Ying D., Elin J., Michael F., Ioan T., Erjia Y., Schubert F., Milojević S.. Upper tag ontology for integrating social tagging data, Journal of the American Society for Information Science and Technology, Volume 61, Issue 3, pages 505–521 (2010)
4. Marlow C., Naaman M. and Davis M. Collaborative Web Tagging, Proceedings of the 15th International World Wide Web Conference. (2006)
5. Satnam A. Collective Intelligence in Action, Manning Publications, November 4, 2008 pages 50-81 (2008)
6. Marinho Leandro B., Hotho A., Jäschke R., Nanopoulos A., Rendle S., Schmidt-Thieme L., Stumme G. and Symeonidis P.. Recommender Systems for Social Tagging Systems, Springer, pages 5 – 12. (2012)
7. Song Y., Zhang L., Giles C. L. Automatic tag recommendation algorithms for social recommender systems, ACM Transactions on the Web (2011)Volume: 5, Issue: 1, ACM, Pages: 1-31(2011)
8. Zanardi V., Capra L. Social ranking: uncovering relevant content using tag-based recommender systems, Proceeding Rec Sys '08 Proceedings of the 2008 ACM conference on Recommender systems (2008)
9. Shepitsen A., Gemmell J., Mobasher B., Burke R. Personalized recommendation in social tagging systems using hierarchical clustering, Proceeding Rec Sys '08 Proceedings of the 2008 ACM conference on Recommender systems (2008)
10. Peis E., Morales-del-Castillo J. M., Delgado-López J. A.. Semantic Recommender Systems. Analysis of the state of the topic, Hipertext.net Yearbook, volume 8 (2008)
11. Dell'Aglio D., Celino I., Cerizza D. Anatomy of a Semantic Web-enabled Knowledge-based Recommender System, Proceedings of the 4th international workshop Semantic Matchmaking and Resource Retrieval in the Semantic Web, at the 9th International Semantic Web Conference 2010, Shangai, China,(2010)
12. Adrian B., Sauermann L., Roth-Berghofer T. ConTag: A Semantic Tag Recommendation System, Proceedings of I-MEDIA ’07 and I-SEMANTICS ’07 Graz, Austria (2007)

S. Markovski, M. Gusev (Editors): ICT Innovations 2012, Web Proceedings, ISSN 1857-7288

S. Markovski, M. Gusev (Editors): ICT Innovations 2012, Web Proceedings, ISSN 1857-7288
© ICT ACT – http://ictinnovations.org/2012, 2012

142 ICT Innovations 2012 Web Proceedings ISSN 1857-7288

13. Iacobelli F., Birnbaum L., Hammond K. Tell me more, not just "more of the same", IUI '10 Proceedings of the 15th international conference on Intelligent user interfaces (2010)
14. Ulicnya B., Matheusa C. J, Kokar M. M. A Semantic Wiki Alerting Environment Incorporating Credibility and Reliability Evaluation, Proceedings of the 5th International Conference on Semantic Technologies for Intelligence, Defense, and Security *(STIDS 2010)* (2010)
15. Faviki Case Study: Semantic Tags http://www.w3.org/2001/sw/sweo/public/UseCases/Faviki/
16. Zaino J. Putting Wikipedia to Work for the Semantic Web http://semanticweb.com/putting-wikipedia-to-work-for-the-semantic-web_b416
17. Rizzo G., Troncy R. NERD: A Framework for Evaluating Named Entity Recognition Tools in the Web of Data (poster). Demo Session at the 10th International Semantic Web Conference (ISWC'2011) (2011)
18. Rizzo G., Troncy R., Hellmann S. and Bruemmer M. NERD meets NIF: Lifting NLP Extraction Results to the Linked Data Cloud. (LDOW'12) Linked Data on the Web (WWW'12) (2012)
19. Halb W., Stocker A., Mayer H., Mülner H., Ademi I. Towards a commercial adoption of linked open data for online content providers. I-SEMANTICS '10 Proceedings of the 6th International Conference on Semantic Systems (2010)
20. Abel F., Henze N., Kawase R., Krause D., and Patrick S. TagMe!: Enhancing Social Tagging with Spatial Context. WEBIST Selected Papers, Vol. 75 Springer (2010) , p. 114-128 (2010)
21. Entity Extraction & Content API Evaluation http://blog.viewchange.org/2010/05/entity-extraction-content-api-evaluation/
22. Nowack B. Linked Data Entity Extraction with Zemanta and OpenCalais - benjamin-nowack's blog http://bnode.org/blog/2010/07/28/linked-data-entity-extraction-with-zemanta-and-opencalais
23. Puzzle Pieces Comparing NLP APIs for Entity Extraction http://faganm.com/blog/2010/01/02/1009/
24. Satnam A. Collective Intelligence in Action, Manning Publications, November 4, 2008 pages 31-34 (2008)
25. Passant A., Laublet P. Meaning Of A Tag: A Collaborative Approach to Bridge the Gap Between Tagging and Linked Data, Proceedings of the WWW 2008 Workshop Linked Data on the Web (LDOW2008) (2008)
26. Kim Hak L., Scerri S., Breslin J. G., Decker S., Kim Hong G. The State of the Art in Tag Ontologies: A Semantic Model for Tagging and Folksonomies, Proc. Int'l Conf. on Dublin Core and Metadata Applications (2008)
27. Rivero Carlos R., Schultz A., Bizer C. Benchmarking the Performance of Linked Data Translation Systems, Linked Data On the Web (LDOW) workshop (2012)

S. Markovski, M. Gusev (Editors): ICT Innovations 2012, Web Proceedings, ISSN 1857-7288
© ICT ACT – http://ictinnovations.org/2012, 2012

Security and Privacy Issues and Requirements for Healthcare Cloud Computing

Goce Gavrilov[1], Vladimir Trajkovik[2]

[1]Health Insurance Fund of Macedonia, Macedonia bb, Skopje, Macedonia
gavrilovgoce@yahoo.com
[2]Faculty of computer science and Engineering, "Ss. Cyril and Methodius" University, Rugjer Boskovikj 16, Skopje, Macedonia
trvlado@finki.ukim.mk

Abstract. Information technology is increasingly used in healthcare with the goal to improve and enhance medical services and to reduce costs. One of the areas with greatest needs having available information at the right moment and with high accuracy is healthcare. With the widespread use of electronic health record (EHR), building a secure EHR sharing environment has attracted a lot of attention in both healthcare industry and academic community. Cloud computing paradigm is one of the popular health IT infrastructure for facilitating EHR sharing and EHR integration. Healthcare clouds offer new possibilities, such as easy and ubiquitous access to medical data, and opportunities for new business models. However, they also bear new risks and raise challenges with respect to security and privacy aspects. Ensuring the security and privacy is a major factor in the cloud computing environment.

In this paper, we will present current state of the art research in this field. We focused of several shortcomings of current healthcare solutions and standards, particularly for platform security, privacy aspect and requirements which is a crucial aspect for the overall security of healthcare IT systems.

Keywords: cloud computing, electronic health record (EHR), privacy, security

1 Introduction

Using of information technology in the healthcare (healthcare IT) has become increasingly important in many countries in the recent years. There are continuing efforts on national and international standardization for interoperability and data exchange. Cloud computing aims to incorporate the evolutionary development of many existing computing approaches and technologies such as distributed services, applications, information and infrastructure consisting of pools of computers, networks, information and storage resources [17]. It is still an evolving paradigm but has shown tremendous potential to enhance collaboration, agility, scale, and availability although its definitions, issues, underlying technologies, risks, and values need to be refined. Cloud computing has been defined by the US National Institute of Standards and Technology (NIST) defines cloud as follows:

S. Markovski, M. Gusev (Editors): ICT Innovations 2012, Web Proceedings, ISSN 1857-7288

"Cloud computing is a model for enabling convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction. This cloud model promotes availability and is composed of five essential characteristics, three delivery models, and four deployment models." [14].

In traditional IT environments, clients connect to multiple servers located on company premises. Clients need to connect to each of the servers separately. In Cloud Computing clients connect to the Cloud. The Cloud contains all of the applications and infrastructure and appears as a single entity. Cloud Computing allows for dynamically reconfigurable resources to cater for changes in demand for load, allowing a more efficient use of the resources. Virtualization in Cloud Computing allows distributing computing power to cater for load fluctuations. Standard web protocols provide access to Cloud Computing and control is centrally managed in various data centers.

In the healthcare field, cloud computing offers great potential for quick access to healthcare information. IT infrastructure in the healthcare is very complex and for this reason United States Congress has taken additional measures to protect the patient's private data under HIPAA (Health Insurance Portability and Accountability Act). Cloud computing can help patients to gain access to their medical data from anywhere in the world via the Internet. The healthcare domain needs increased security and privacy levels. In order to achieve this requirements cloud computing technology has to be more carefully managed. The matter is less technical and more ethical and legal. On the international basis the ISO (Technical Committee 215) [24] and the Health Level 7 consortium (HL7) [25] define standards for e-health infrastructures.

In this paper, we will present current state of the art research in this field. We present an overview of the security and privacy issues in the healthcare cloud and requirements for implementation of cloud computing in healthcare. The paper is organized as follows: In Section 2, we present an overview of abstract model of healthcare cloud. Section 3, gives some aspect of the security and privacy issues in the healthcare cloud. Also, in this section we present the requirements for building a healthcare cloud. In section 4, we give a systematic overview of the threats in the privacy and security sensitive context of healthcare clouds. Section 5 discusses cloud computing as a solution supporting healthcare information systems, and Section 6 concludes the suggested solution.

2 Model of the Healthcare Cloud: Overview

By NIST, cloud computing model consists of five characteristics, three delivery models, and four deployment models [3]. The five key characteristics of cloud computing are: location-independent resource pooling, on-demand self-service, rapid elasticity, broad network access, and measured service [26]. The author of the [3], [5] and [14] describe each of these characteristics and models. These five characteristics represent the first layer in the cloud environment architecture (see figure 1).
According to the different types of services offered, cloud computing can be considered to consist of three layers: IaaS- Infrastructure as a Service is the lowest

layer that provides basic infrastructure support service, PaaS – the Platform as a Service layer is the middle layer, which offers platform oriented services, besides providing the environment for hosting user's applications and SaaS - Software as a Service is the topmost layer which features a complete application offered as service on demand [5].

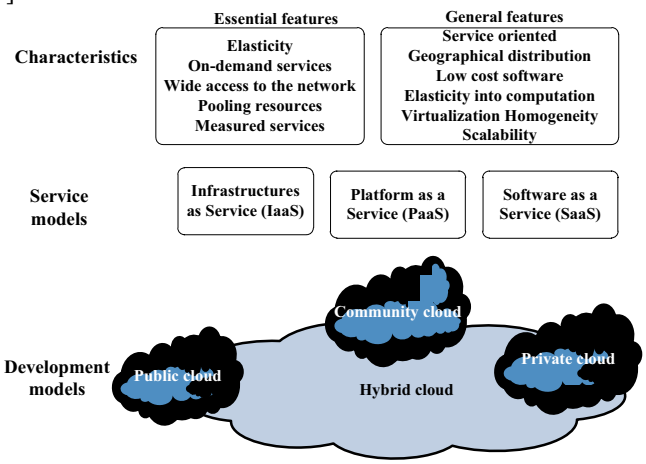


Fig. 1. Element and characteristics of the cloud

IaaS model provides the capability for consumers to provision processing, storage, networks, and other fundamental computing resources, in which consumer is able to deploy and run arbitrary software, including operating systems and applications. IaaS refers to the sharing of hardware resources for executing services, typically using Virtualization technology. With IaaS approach, potentially multiple users use available resources. The consumer does not manage or control the underlying cloud infrastructure but has control over operating systems, storage, deployed applications, and possibly limited control of select networking components. In the IaaS cloud model, the healthcare application developers hold full responsibility for protecting patients' security and privacy.

PaaS offers an integrated set of software that provides everything that a software developer needs to build an application- an online environment for quick development of web applications using browser-based development tools. PaaS model aims to protect data, which is especially important in case of storage as a service. The data needs to be encrypted when hosted on a platform for security reasons. The consumer does not manage or control the underlying cloud infrastructure including network, servers, operating systems, or storage, but has control over the deployed applications and possibly application hosting environment configurations. In this type of cloud service model, two levels of protection for security and privacy are required.

SaaS– business applications hosted and delivered as a service via the web. These kinds of applications do not require installation of additional computer programs, the most popular being the e-mail in a web browser. This layer provides capability for consumers to use the provider's applications running on a cloud infrastructure. It eliminates the need to install and run the application on the customer's local

computer, thus alleviating the customer's burden for software maintenance. The consumer does not manage or control the underlying cloud infrastructure including network, servers, operating systems, storage, or even individual application capabilities. In this type of cloud service model, the security and privacy protection is provided as an integral part of the SaaS to the healthcare consumers.
Cloud computing is offered in four different forms: Private clouds– are owned by a single organization and are being used only in that organization; Community clouds – belonging to several organizations and allowing access only to those concerned for certain actions; Public clouds – are held by a company selling cloud services to the general public; Hybrid clouds – a composition of two or more types of clouds (private, public or community) that remain unique entities but are linked by standard technologies that enable portability of applications [3], [17].

Zhang and Liu [5] define the concept of Personal Health Record (PHR), Electronic Health Record (EHR), and Electronic Medical Record (EMR) in the healthcare cloud computing. The terms of EHRs and EMRs are used alternately in both: healthcare industry and the press or health science literature. Both EMRs and EHRs are critical to the grand vision of healthcare digitization for improving safety, quality and efficiency of patient care and reducing healthcare delivery costs [5]. Furthermore in this paper we use the following terms:

- Health professional: person who delivers health care services, e.g., physician, dentist, pharmacists, etc.
- Health care provider: organization that provides services of health professionals, e.g., doctor's practice or hospital.

The terms of EHRs and EMRs are used alternately in both: healthcare industry and the press or health science literature. Both EMRs and EHRs are critical to the grand vision of healthcare digitization for improving safety, quality and efficiency of patient care and reducing healthcare delivery costs.

- EMR is the legal record of what happened to the patient during their encounter at a Care Delivery Organization (CDO) across inpatient and outpatient environments and is owned by the CDO. EMR is created, used and maintained by healthcare practitioners to document, monitor, and manage health care delivery within a CDO.
- EHR- database of medical data objects and health-related data managed by health professionals. EHR is a subset of EMR record maintained by each CDO and is created and owned by the patient.
- Personal Health Record (PHR): database of medical data objects and health-related data managed by a patient.

It is very important to ensure the availability of medical data to all the locations a patient is present in. Several examples and developments are already available in literature and presented in the following.

In [1], a model is presented as an integrated EMR sharing medical data between medical units. The application is developed on a cloud platform that keeps the EMR system on the form of SaaS and can be used by Government, Hospitals, Doctors, Patients, Pharmacies and Health Insurance Organizations, through the Internet. This system allows access to national data sharing; the data center is common to all units. Communication between the data center and the healthcare organizations is done via HL7 messages. All patient data are stored and accessed in the same location over the Internet from any healthcare organizations.

In the paper [2], [4] are presented two examples of using cloud computing in Indian healthcare based on SaaS types of cloud computing. First presents the architecture and implementation of cloud computing system in India healthcare system and discusses its strengths and weaknesses. The author of the [7], [12] presents the utilization of telemedicine and wireless sensor network and devices in the cloud computing environment.
Hans at all in [9] discuss the general problems of e-health systems and provide a technical solution for the protection of privacy-sensitive data, which has not been appropriately addressed yet for end-user systems (simple model of e-health cloud and advanced e-health cloud Infrastructure). They describe an abstract model of e-health clouds, which comprehends the common entities of healthcare infrastructures.

3 Problems of Healthcare Clouds

In this section, we give an overview of the threats in the privacy context and security aspect of e-healthcare clouds. The processing of healthcare data of patients has technical, but also legal problems that one has to deal with them. It seems unlikely that any technical means could completely prevent cloud providers from abusing customer data in all cases, so we need a combination of technical and nontechnical means to achieve this. We focus a little on the technical aspects but more on the legal aspects.
Löhr at all [9] have developed the concept of technical aspect of the privacy context and security aspect in the e-health cloud computing. The key segments that affect the technical aspects of security and privacy in healthcare clouds are: data storage and processing (data centers where store data, client platforms and mobile storage devices), management of e-health Infrastructure (cryptographic key management, management of certificates and hardware/software component), usability and user experience. Most of these problems can be overcome by the providers of cloud computing services, maintenances of the e-healthcare infrastructure and users themselves.
The legal aspects refer to the privacy legislation and regulations in the countries where use cloud computing. Policies on the creation of privacy legislation in the European Union (EU) and the United States (US) are differing. Privacy in the US is dispersed among various different sector specific laws and these sectors include the health care sector for the HIPAA. The EU has a different approach concerning legislation. The EU approach to legislation favors participation among businesses and governments as opposed to the US self-regulation approach. The EU set the privacy regulations up to the front as opposed to relying on industry self-regulation.

Research on the various security issues surrounding healthcare information systems has been developed over the last few years. ISO/TS 18308 standard gives the definitions of security and privacy issue for EHR. The Working Group 4 of International Medical Informatics Association (IMIA) was set up to investigate the issues of data protection and security within the healthcare environment. Its work to date has mainly concentrated on security in EHR networked systems and common

security solutions for communicating patient data [5]. Choice of the International Classification of Diseases, 10th revision (ICD-10) and ICD-10-CM, for the standard of codification in the EMR system are very important in the healthcare information system [1]. For the data exchange between the entities involved in the healthcare cloud computing, it is necessary to use of HL7 standard which is widely being adopted by health care institutions in several nation-wide. Countries that wish to allow use of cloud computing in healthcare information systems, in their laws, have to provide appropriate provisions to include the previously mentioned standards and some new that relate to the security and privacy of data.

4 Healthcare Cloud Privacy and Security Concept

Security and privacy in the healthcare cloud computing are more than just user privileges and password enforcement. Cloud computing platforms are multi domain environments in which each domain can use different security, privacy, policies and procedures, and trust requirements and potentially employ various mechanisms, interfaces, and semantics, secure data-backup strategy, third-party certification. Such domains could represent individually enabled services or other infrastructural or application components. In healthcare cloud, security should be the top priority from day one. In healthcare cloud applications, some of the security and privacy issue and requirements are orthogonal to the concrete cloud service model or cloud deployment model used. In this section, we briefly present these issue and requirements. Security in Cloud Computing consists of established security solutions such as encryption, access management, firewalls and intrusion detection. In internal Clouds computing the IT department has the ability to install all available security solutions it sees fit but in external Cloud Computing the security depends on the Cloud Service Provider (CSP). Some CSPs do not provide flexibility in the choice of security solutions, while others allow the implementation of client security requirements.

The emergence of cloud computing is a recent development. The security issues in Cloud Computing [14] are organized into several general categories: trust, architecture, identity management, software isolation, data protection, and availability. Cloud computing has grown out of an amalgamation of technologies, including service oriented architecture, virtualization, Web 2.0, and utility computing, many of the security issues involved can be viewed as known problems cast in a new setting. Jansen [14] describes all of the security issue in detail. Security of the cloud infrastructure relies on trusted computing and cryptography.

Some countries in the world [9], like Austria, the German electronic Health Card (eHC) system under development, or the Taiwan Electronic Medical Record Template (TMT), took some activities of the e-healthcare security. In Germany each insured person will get a smartcard that not only contains administrative information (name, health insurance company), but also can be used to access and store medical data like electronic prescriptions, emergency information like blood group, medication history, and electronic health records. The smartcard contains cryptographic keys and functions to identify the patient and to encrypt sensitive data. Recalling the definition of EHR, PHR, EMR, securing issue from the previous paragraph and the possibilities

of the eHC, we can define the minimum requirements for cloud computing security concept.

The concept of Cloud Computing brings many uncertainties with respect to compliance with privacy regulations. There are no clear answers on which privacy regulation requirements apply to Cloud Computing. CSPs must assure their customers and provide a high degree of transparency into their operations and privacy assurance. Privacy-protection mechanisms must be embedded in all security solutions. In a related issue, it's becoming important to know who created a piece of data, who modified it and how, and so on. Ensuring privacy and security of health information, including information in EHR, PHR and EMR is a key component to building the trust required to realize the potential benefits of health information exchange in cloud computing. CSPs must provide some Service Level Agreements (SLA) issues and requirements if it wants to offer services of cloud computing.

The migration into a cloud computing environment is in many ways an exercise in risk management. The risks must be carefully balanced against the available safeguards and expected benefits, with the understanding that accountability for security remains with the organization. Both qualitative and quantitative factors apply in an analysis. Too many controls can be inefficient and ineffective, if the benefits outweigh the costs and associated risks. An appropriate balance between the strength of controls and the relative risk associated with particular programs and operations must be ensured.

5 Cloud Computing as a Solution Supporting Healthcare Information System

In the healthcare field, cloud computing offers great potential for quick access to medical information and healthcare information. Healthcare IT infrastructure is very complex and for this reason organization has taken additional measures to protect the patient's private data. Cloud computing can support different healthcare information systems by sharing information stored in diverse locations. All the medical data (EMR, EHR, PHR) are stored in a private cloud and all the participants in the healthcare can access medical patient data when is needed according to their privileges of access. In this case, the medical act is performed quickly, and the typing errors reduced, all of this driving to higher quality. This is one of example of using cloud computing in healthcare.

The authors of [2], [4] give description of some examples of using cloud computer technology in the Indian healthcare sector. In the healthcare area exist many examples of using cloud computing technology, some of them related of using wireless network technology [7] and so-called "Health ATM" [12] based on Google CSPs. All of these examples use some type of cloud computing model or hybrid of them.

The huge benefit of using cloud computing in healthcare, is the possibility of all stakeholder can find data from anywhere and from any place. The costs of the IT infrastructure will be cheaper because the healthcare units will only rent the infrastructure to store healthcare data as it need and will no longer need the latest equipment for the applications used, managed or maintained. They need only computers or devices with access to Internet. In emphasizing the cost and

performance benefits of the cloud computing in healthcare, some fundamental security problems have been left unresolved. Determining the security of complex computer systems is also a long-standing security problem that overshadows large scale computing in general. Attaining the high assurance qualities in implementations has been an elusive goal of computer security researchers and practitioners, and is also a work in progress for cloud computing. Few research papers have systematically studied the impact of cloud computing on healthcare IT in terms of its opportunities and challenges. Table 1 shows some of opportunities and challenges from the viewpoint of management, technology, security, and privacy.

Table 1. Cloud computing opportunity and challenges

Aspects	Opportunities	Challenges
Technology	Infrastructure scalability and flexibility Reduction of IT maintenance Advantage for green computing	Bugs in large-scale distributed cloud systems Unpredictable performance Data transfer bottlenecks Resource exhaustion issues
Management	Computing resources available on demand Lower cost of new IT infrastructure Payment of use as needed	Organizational inertia Lack of trust by health care professionals Loss of governance Provider's compliance
Privacy	Development of guidelines and technologies Protect customer's data and privacy from provider's commitments Fostering of regulations by government for data and privacy protection	Privacy issues Data jurisdiction issues
Security	Increasing data security at replication of data in multiple locations Strengthening resilience- dynamically scaled defensive resources More resources available for data protection	Privilege abuse Poor encryption key management Public management interface issues Failure separation

Security of the cloud infrastructure relies on trusted computing and cryptography. Healthcare data must be protected in a manner consistent with policies, whether in the organization's computing center or the cloud. No standard service contract exists that covers the ranges of cloud services available and the needs of different organizations.

6 Conclusion

Using the cloud computing technology in a healthcare may considerably improve the access to information, which can be done be much easier. The scalability, that is the key of the cloud computing, can offer more resources needed for certain operation at any time. The collaboration between healthcare units is an opportunity offered by

cloud computing for healthcare staff. With this technology can be checked the availability of a physician, a medical specialist, a product or a service at different times and in different cases.

Security and privacy issues of cloud computing are delaying its fast adoption, but it has become very popular and we needed to provide security mechanisms to ensure its secure adoption. While security and privacy services in the cloud computing can be fine-tuned and delivered in new ways and by new types of service providers, there need to be frameworks that efficiently deliver cloud-based security services and provide a desirable solution to customers based on their requirements.

Although the use of cloud computing has rapidly increased, cloud computing security is still considered the major issue in the cloud computing environment, especially when it comes to medical data. Customers do not want to lose their private information as a result of malicious insiders in the cloud. In addition, the loss of service availability has caused many problems for a large number of customers recently. Implementation of the privacy and security standards that are currently under development within the cloud community, including business associate contracts that specify auditable, enforceable performance metrics and sharing of liabilities, should allow such a system to achieve compliance with federal privacy and security regulations. There are still many challenges to fostering the new model of cloud computing in healthcare.

Cloud computing is a new model of computing that promises to provide more flexibility, less expense, and more efficiency in IT services to end users. It offers potential opportunities for improving EHR adoption, health care services, and research. When a healthcare organization considers moving its service into the cloud, it needs strategic planning to examine environmental factors such as staffing, budget, technologies, organizational culture, and government regulations that may affect it, assess its capabilities to achieve the goal, and identify strategies designed to move forward. Cloud computing presents a compelling opportunity for consumers of IT and producers of information services.

References

1. Pardamean, B., Rumanda, R. R.: Integrated Model of Cloud-Based E-Medical Record for Health Care Organizations. In 10th WSEAS International Conference on E-Activities, pp. 157-162, December 2011
2. Srivastava, P., Jada, R., Razdan, P.: Cloud Computing in Indian Healthcare Sector. In Proceedings of ASCNT-2011, CDAC Noida, India, 2011
3. Lupşe, O.S., Vida, M. M., Tivadar, L. S.: Cloud Computing and Interoperability in Healthcare Information Systems. In INTELLI 2012: The First International Conference on Intelligent Systems and Applications, 2012
4. Karthikeyan, N., Sukanesh, R.: Case Study on Software as a Service (SaaS) Based Emergency Healthcare in India. In European Journal of Scientific Research, ISSN 1450-216x, Vol.69 No.3, pp. 461-472, 2012
5. Zhang, R., Liu, L.: Security Models and Requirements for Healthcare Application Clouds. In IEEE 3rd International Conference on Cloud Computing, July 2010
6. Rolim, C. O., Koch, F. L., Westphall, C. B., Werner, J., Fracalossi, A., Salvador, G. S.: A Cloud Computing Solution for Patient's Data Collection in Health Care Institutions. In Second International Conference on e-Health, Telemedicine and Social Medicine, 2010
7. Perumal, B., Rajasekaran, P. M., Ramalingam, H. M: WSN Integrated cloud for automated telemedicine based e-healthcare applications. In 4th International Conference on

Bioinformatics and Biomedical Technology, IPCBEE vol.29. IACSIT Press Singapore, 2012
8. Nalin, M., Baroni, I., Sanna, A.: E-health drivers and barriers for cloud computing adoption. In International Conference on Cloud Computing & Services Science, Netherlands, 2011
9. Löhr, H., Sadeghi, A. R., Winandy, M.: Securing the E-Health Cloud. In Proceedings of the 1st ACM International Health Informatics Symposium, IHI 2010
10. Dandapani, A., Palani, D.: Cloud Computing (Implementation in Health Care Technology and Solution for Instant Medication using Cloud). In International Conference on Computing and Control Engineering, April 2012
11. Schweitzer, E. J.: Reconciliation of the cloud computing model with US federal electronic health record regulations,Journal of American Medical Information Association, may 2012
12. Botts, N., Thoms, B., Noamani, A., Horan, T. A.: Cloud Computing Architectures for the Underserved: Public Health Cyberinfrastructures through a Network of HealthATMs. In Proceedings of the 43rd Hawaii International Conference on System Sciences, 2010
13. Rashid Al Masud, S. M.: A Novel Approach to Introduce Cloud Services in Healthcare Sectors for the Medically Underserved Populations in South Asia. In International Journal of Engineering Research and Applications, Vol. 2, Issue 3, pp.1337-1346, May-Jun 2012
14. Jansen, W. A.: Cloud Hooks:Security and Privacy Issues in Cloud Computing. In Proceedings of the 44th Hawaii International Conference on System Sciences, 2011
15. Clarke, A., Steele, R.: Secure and Reliable Distributed Health Records: Achieving Query Assurance Across Repositories of Encrypted Health Data. In 45th Hawaii International Conference on System Sciences, 2012
16. Al Zain, M. A., Pardede, E., Soh, B., Thom, J. A.: Cloud Computing Security: From Single to Multi-Clouds. In 45th Hawaii International Conference on System Sciences, 2012
17. Takabi, H., Joshi, J. B. D.: Policy Management as a Service:An Approach to Manage Policy Heterogeneity in Cloud Computing Environment. In 45th Hawaii International Conference on System Sciences, 2012
18. Morin, J. H., Gateau, B.: Towards Cloud Computing SLA Risk Management: Issues and Challenges. In 45th Hawaii International Conference on System Sciences, 2012
19. Huu, T. T., Koslovski, G., Anhalt, F., Montagnat, J., Vicat, P., Primet, B., Elastic, J.: Cloud and Virtual Network Framework for Application Performance-cost Optimization. Published online: 4 November 2010, © Springer Science+Business Media B.V. 2010
20. Rimal, B. P., Jukan, A., Katsaros, D., Goeleven, Y.: Architectural Requirements for Cloud Computing Systems: An Enterprise Cloud Approach. Published online: 7 December 2010, © Springer Science+Business Media B.V. 2010
21. Caron, E., Desprez, F., Muresan, A.: Pattern Matching Based Forecast of Non-periodic Repetitive Behavior for Cloud Clients. Published online: 6 January 2011, © Springer Science+Business Media B.V. 2011
22. Diaz, R. G., Ramo, A. C., Agüero, A. C., Fifield, T., Sevior, M.: Belle-DIRAC Setup for Using Amazon Elastic Computer Cloud Providing Homogeneous Access to Heterogeneous Computing Resources. Published online: 13 January 2011, © Springer Science+Business Media B.V. 2011
23. Vaquero, L., Caceres, J., Lindner, M., Merino, L. R.: A Break in the Clouds: Towards a Cloud Definition. In ACM SIGCOMM Computer Communication Rev, pp. 50–55, 2009
24. Health Level Seven International (HL7), http://www.hl7.org/
25. International Organization for Standardization (ISO). Technical Committee 215, Health Informatics, http://www.iso.org/iso/iso_technical_committee?commid=54960
26. Hassan, T., James, B. D., Gail-Joon, A.: Security and Privacy Challenges in Cloud Computing Environments. In IEEE journal & magazines, vol. 8 Issue 6, pp. 32-39, 2010
27. Sunyaev, A., Pflug, J.: Risk evaluation and security analysis of the clinical area within the German electronic health information system. Published online: February 2012, Springer

Context-aware QoS: Different Approaches for Classification and Provisioning

Toni Malinovski, Vladimir Trajkovik

Faculty of Computer Science and Engineering,
Ss. Cyril and Methodius University, Skopje, 1000, R. Macedonia

tmalin@nbrm.mk, trvlado@finki.ukim.mk

Abstract. Coorporate networks are running diffrent applications which support the execution of business processes. These networks have to provide proper level of Quality of Service (QoS) according to the requirements of an entire business process. Still, this can not be established using traditional QoS mechanisams which do not take into the account the requirements of the bussiness side. This paper presents two different approaches for classification and provisioning of novel QoS schemes which integrate with the emerging business processes and dynamic requirements from the infrastructure. Each approach focuses on different parts of the system where context-aware intelligence can be applied to meet the QoS needs, for better Quality of Experience (QoE) for the end-user. The advantages and disadvantages of each approach are presented through results gathered from an environment which runs several business applications, where QoS is applied.

Keywords: QoS, QoE, Context-aware, Infrastructure, Middleware, Classification, Provisioning.

1 Introduction

Modern network infrastructures are rapidly changing their behavior and outlook. Different media - data, video, and voice are already widely consolidated into a converged network creating single, standard based, modular infrastructure that gives organizations more efficiency and simplified management. The convergence of the network infrastructure has not only changed the network models, but it has also affected the way the network supports services and applications. The challenges for the future development in the field of network management include the need to make the network application aware, for the optimization and cross-layers security, which should result in more efficient delivery of business applications to the end-users. Business processes relay on different applications running within the network infrastructure, so different approaches must be evaluated and adopted for successfully reaching business objectives with positive expirience for the end-user. The large amount of data within the converged network may introduce bottlenecks at certain part of the infrastructure, so appropriate measures must be taken in advance, to avoid

S. Markovski, M. Gusev (Editors): ICT Innovations 2012, Web Proceedings, ISSN 1857-7288

the problems that may occur and provide stable, efficient, cost-effective solutions. QoS is mechanism, usually defined as a set of requirements that need to be met by the infrastructure while transporting a data stream from the source to the destination. The traditional QoS mechanism deal with the technical behavior of the protocols, services and the application within the network, so upgrade of QoS approaches based on context-base recognition must be defined, evaluated and tested, to ideally give solution that is not application dependent and can meet the end-user requirements.

In our mothodology, the traditional QoS mechanisms serve as a base, which is built on, by adding context-aware QoS-drivers and Middlewares into the infrastructure. Still, we must realize that the end-users are not interested in the provisioning and control mechanisms that are implemented in the infrastructure. They expect proper implementations that will deliver best services which will support their bussines processes. In the last years, Quality of Experience (QoE) [17] has emerged as a full scale evaluation of systems‘ performance in terms of end-user expectations. Therefore, beside the techincal setup, we investigate the relations among QoS variables that may impact the degree of end-users' satisfaction and perceived quality.

The main contribution of this paper is the presentation of two different approaches for classifications and provisioning, that can give added value to the existing QoS mechanism and built context-aware QoS, which can meet the business needs at the end and provide positive QoE. These approaches focus on different points of the infrastructure and based on preliminary tests within proper environment, generate results that produce concepts and guidelines how the QoS schemes should develop in the future. Furthermore, application and services are accessed by different kind of device like desktops, laptops, ip phones etc, so the context-aware QoS should provide different aspects for each one. Focusing on the users‘ point of view, the two approaches are evaluated in a form of end-users‘ perceived QoS, referred as QoE.

In section 2, the system infrastructure and the different approaches for context-aware QoS are presented; analyses and related work are given in section 3, while section 4 concludes the paper.

2 Different Approaches for Classification and Provisioning

In this paper, the context of the business process, as an input to a context-aware classification and provisioning, is researched through two different approaches in the same system infrastructure in order to provide results and determine the correct concept for development. We focus on different parts of the system where context-aware intelligence can be applied, try to improve the performance based on the bussines needes and evaluate the system’s performance and subjective end-users indicators in terms of QoE. The system infrastructure is consisted of a live environment with several switches, routers and client devices, application and database servers. It represents a wired local and wide area network, which has congestion at certain points, during specific time periods during business hours. This system was used to test the behaivor of the proposed approaches for classification and provisioning,. We focused on achieving transparency to already implemented QoS

mechanisms in the network, by adding an additional layer for context-awareness. A good system infrastructure, where different scenarios can be tested, has to be defined with several traffic classifications which should be provisioned accordingly, to provide consistent and predictable performance.

Different production environments can have different protocols, services and applications running inside. So the number of classes of the traffic that should be taken into the consideration, which should have good representation of the actual situation, can vary from two-three to nine-ten different classes.

Having all this in consideration, we have chosen to classify the traffic evaluated in our system infrastructure in six different classes, as shown on Figure 1.

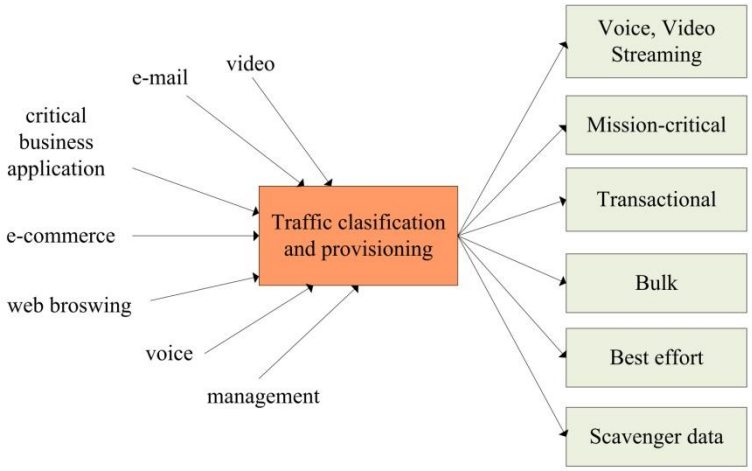


Fig. 1. Traffic classification in the system infrastructure

In our opinion, the number and the nature of the chosen classes, is sufficient to cover broad range of production infrastructures in different corporate organizations.

The system runs different business applications for everyday operations, so it can provide accurate results based on the business needs. For example, on of them is the business application for "on-line auction of state bills", which is implemented on the application level. This application requires strict priority and mission critical provisioning for half an hour twice a week, while online auctions are in place, and should have bulk provisioning for the rest of the business hours while reports and statistics are being retrieved. After business hours this application should have best effort provisioning. Besides this time-based context behavior, this example was tested for specific tasks and steps within the application. Some tasks and steps are more important and mission critical, so they should be treated accordingly. The proposed approaches for classification and provisioning should generate results for the behavior of the context-aware QoS, and possibile provide efficiency and excellent user experience.

2.1 Context-aware Approach with QoS Middleware

This approach is designed to implement transparency for the existing QoS mechanisms in place and introduce additional QoS Middleware layer [2] positioned on the application servers or gateways into the infrastructure [3], [4]. The network infrastructure and the transport layer are already configured according to the predefined policy for provisioning of different classes through the slow links, so this layer will introduce a business intelligence and classify the traffic depending on the business needs.

The design of the Context-aware approach with QoS Middleware is shown on Figure 2

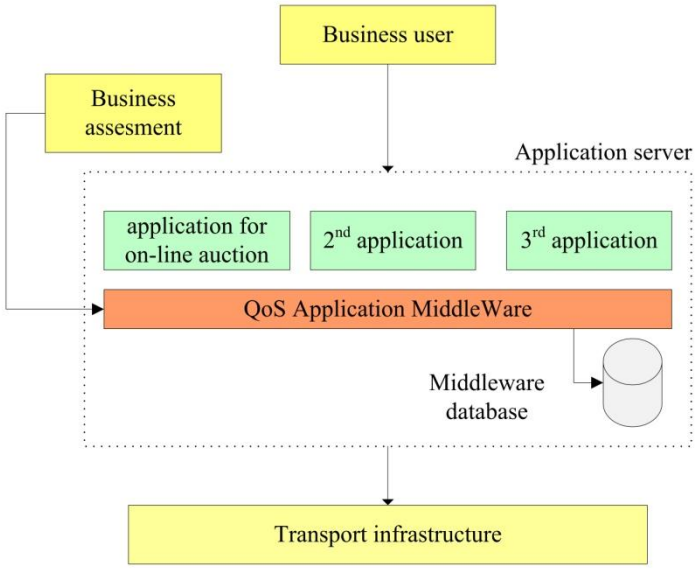


Fig. 2. QoS Middleware within the application server

This layer is placed on the application servers in the infrastructure and is configured to reflect the business needs [8], [10]. Since IP is chosen as a network protocol in our research, which is the most common protocol these days, within the Middleware packets are marked using the DSCP field in the IP header [1]. Depending of the business logic at specific moment, the Middleware application marks (colors) the packets accordingly. For the "on-line auction of state bills" example, the built in intelligence into this layer, was programmed to classify the data differently for predetermined time periods during the work day, so the packet leave the application server already classified. It also reacts at specific steps and tasks where possible.

In this way, the networking infrastructure and its' predefined policy for provisioning, follows the desired classification and handles the business application as need, introducing a context-aware infrastructure and QoS in the places of the network

where needed. The different types of traffic depending of the classification are colored to our needs using the DSCP field. Table 1 shows the markings we have chosen for the six different classes presented in our scenario [1].

Table 1. Classification and markings for different types of traffic.

	Traffic class	DSCP Layer 3	CoS Layer 2
1.	Voice, Video, Streaming	46	5
2.	Mission-critical	26	3
3.	Transactional	22, 20	2
4.	Bulk	12, 14	1
5.	Best effort	0	0
6.	Scavenger data	2	0

Most of the standardized equipment is able to recognize the markings and provision the traffic accordingly [12]. For the ISO/OSI reference model Layer 2 devices, traffic is being marked at Layer 2 by copying the DSCP markings [1] (that come from the Middleware) when they enter the device, so they can follow the QoS policy at this layer too.

For this approach, the existing QoS policy was altered to embrace the new QoS Middleware, so the overall policy can now reflect the business needs by possible providing an incresed level of QoE.

The results of this approach have shown that the whole infrastructure can change and meet the business needs. The Middleware positioned on the application servers [3], [4], [11], [14] is a place where most of the development should be done, since the overall QoS policy is now stored in the Middleware database. In our environment this layer was able to deal with the applications that were passing through the application servers and provide adequate treatment. Diffrent devices present in the environment like the IP phones, multimedia streaming servers, mark the traffic accordingly, are part of the overall QoS policy and have their place in the Middleware database only for informational purposes, since their traffic classification is still static.

The results have shown that the infrastructure was behaving to our expectation and was providing results for context classification and provisioning. The Middleware layer needs to be further developed to be more integrated into the business application and provide even deeper understanding of the application itself when needed. The Middleware was able to follow the predefined behavior of the applications and also produce results that are aligned to the dynamic requirements of the users while using the application by situation analysis. These preliminary results give positive impulse for developing this approach into a larger concept for context-aware QoS classification and provisioning of the traffic through the overall infrastructure.

2.2 Context-aware QoS Approach within the Infrastructure Equipment

The second approach we suggest is to implement context-awareness within the infrastructure equipment itself. The system infrastructure in our environment has

sophisticated networking routers on the bottleneck points of the network, which are designed and configured to have collaboration with the application layer of the traffic that passes through. Following this design, the infrastructure itself has application recognition Middleware, which can be programmed to the business needs and tested for results.

The already predefined policy for classification and provisioning used in our previous approach was reflected in this approach too, so the results can be evaluated and compared. The application for "on-line auction of state bills" was also tested in this approach and the infrastructure was programmed to follow the mentioned business needs. The application recognition layer of the infrastructure was configured to be context-aware and to classify and provision different protocols and applications.

The design for this approach is shown on Figure 3.

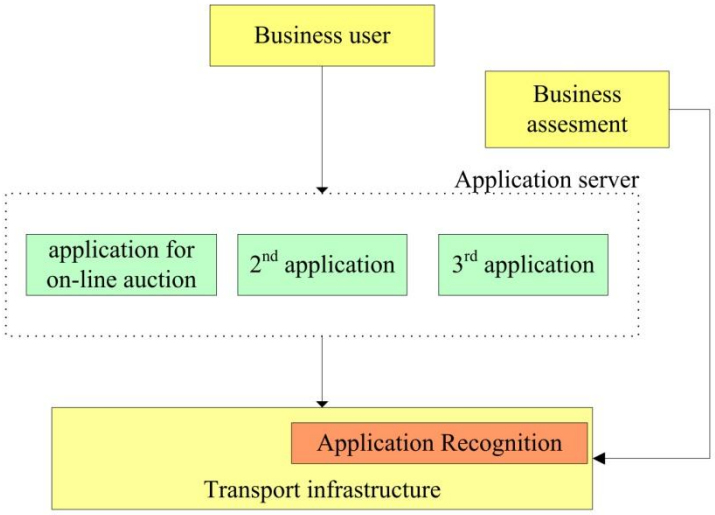


Fig. 3. Application recognition within the transport infrastructure

With this approach the packets are not marked using the DSCP field in the IP header, because the context-aware intelligence and the provisioning devices (routers) are in fact the same infrastructure [9]. In this way, the decisions could be made directly on the device itself by assessing the business needs and provision the traffic accordingly.

When the intelligence is implemented within the infrastructure, the capabilities of the devices to recognized different types of traffic can be used in advance, when the QoS classification and provisioning mechanisms are being planned. Therefore, in this approach we have designed two phases scenario which included:

- Discovery phase, which was running for one week, while the routers were gathering information and building statistics for the traffic in the infrastructure;

- Classification and provisioning phase, which used the gathered information of the previous phase, so we can configure the devices to classify the traffic according to the behavior of the network.

Most of the vendors that offer networking infrastructure devices that support application recognition and intelligent integration with the infrastructure, are able to recognize large amount of standardized protocols, services and applications, with low impact of devices' resources. The discovery phase produced information about the traffic based on this intelligence. It also produced information that the devices could not automatically understand which came from the non standardized business application. So within this approach, through configuration of the devices, additional filters are implemented to produce additional layer, which can recognize the desired application and meet the user requirements at the end [13]. The previously mentioned six different classes are easily implemented within this approach too.

The results of this approach have shown that the infrastructure itself can be tailored to meet the business needs. The tested applications showed that the system was positioned to their QoS requirements and could be considered as a concept that can be treated as a context-aware QoS approach. The additional devices present in the researched environment like the IP phones, multimedia streaming servers were easily integrated and covered with the overall QoS policy through this approach.

On the other hand, this approach was showing less flexibility and scalability than the previous approach. Its capability to take into account the dynamic requirements of the business process and ideally permit application-independent usage was limited and was showing difficulty for rapid development. The results have shown that this approach can match most of the protocols, services and applications and can be aware of their behavior, but mostly in a predetermined way. For example, it could follow the time-base behavior of the application which was predefined in advance, but showed little inside understanding of certain tasks and steps within the business process itself. The implemented filters within the infrastructure, which have produced the additional context-aware layer, were difficult to program to meet the complete requirements as mentioned.

Still this scenario have shown that this concept has room for development and can produce better results in the future to provide fully and easily managed context-aware QoS system.

3 Analysis and Related Work

Efficient access to the information which should be available for the relevant recipients is decisive competitive success factor. The business processes are prioritized as a result of the business assessment in the organization, so the infrastructure must be aligned to meet the business objective.

By covering different approaches, we have tried to give relevant results that can develop concepts for QoS context-awareness with a novel QoS schemes.

The first approach was better in providing results for context classification and provisioning than the second approach. By focusing on the application Middleware

layer directly, we were able to create intelligence that builds an overall QoS policy in direct link with the business assessment processes. Even more, this approach may eventually produce full concept for solutions that will ideally create application-independent system that focuses only on business requirements.

The second approach provided interesting results as well. By positioning the intelligence in the infrastructure layer, which provided a discovery phase of the underlying traffic, we were able to gather more information about the actual network itself. This information was used when the QoS policy was created for better alignment towards the business processes. Still it lacked the dynamic behavior of the previous approach.

To provide quantitative measures for each approach from the end-users' point of view, we have tested the behavior of the "on-line auction of state bills" application and gathered statistical data for the experience of 39 different users. Graded by humans, and therefore, somewhat subjective, the range of QoE score was from 1 to 5, where 1 is unsatisfactory and 5 is excellent. The infrastructure equipment reported high network utilization (close to 100% of the available bandwidth), while the users were evaluating the behavior of the application. These end-users expressed their QoE while evaluating the application on two different aspects:

- Predefined time-based behavior of the application;
- Random steps and tasks with critical importance within the application.

The summary of the surveys' response data is presented in Table 2 (predefined time-based behavior) and Table 3 (random steps and tasks) represented as mean, standard deviation, skew and kurtosis.

Table 2. QoE grades from both approaches for the tested predefined time-based behavior of the application.

	Mean	Standard Deviation	Skew	Kurtosis
First approach	4.64	0.639	-1.598	1.445
Second approach	4.08	0.906	-0.658	-0.399

Table 3. QoE grades from both approaches for the tested random steps and tasks of the application.

	Mean	Standard Deviation	Skew	Kurtosis
First approach	4.28	0.741	-0.505	-0.978
Second approach	3.53	0.810	0.076	-0.362

From statistical point of view the results in both tables illustrates the normality of the surveys' data. Since absolute values of skew > 3.0 are described as "extremely" skewed and kurtosis absolute values > 8.0 suggest a problem [16], surveys' data presented in the study provide relevant information.

The presented results illustrate how the end-users' QoE was influenced while classification and provisioning was implemented within the system following both approaches. In this paper, we didn't focus on the technical behavior of the system, since the subjective QoE of the end-users provides relevant information from the business point of view. The end-users were not aware of the technical setup, so the produced results could provide information about the relation between the actual QoS implementation and the end-users' QoE. The data show that the first approach provided higher level of positive QoE (higher mean scores) for the end-users when the application was tested during the predefined time-based behavior and for the random steps and tasks. The end-users reported that the second approach also provided high level of positive QoE for the predefined time-based behavior of the application, but didn't perform adequately while random steps and tasks were tested within the application.

In our opinion, this research has shown that a combined approach can be designed, and if planed well, it can produce even better result for the context-aware QoS.

The system infrastructure in our scenarios represents a wired local and wide area network. Additional results can be achieved by extended the scenarios and introducing Wireless Mesh Networks [6] or Mobile Networks [7] in both approaches. Different authors have already explored support for discovering context, based on the needs of various mobile applications/agents, acquiring contexts from various sources, and deliver the acquired context data to mobile applications/agents.

Furthermore, existing Wireless or Mobile Middleware [5] can be incorporated in the infrastructure as added layer for context-aware intelligence in the system, which can give additional relevant information to broaden the concepts that have been explored in our work.

4 Conclusion

In this paper we have presented two different approaches for context-aware QoS which are providing positive QoE for the end-users. Each approach was tested with the real business case scenario and preliminary results were obtained. The advantages and the disadvantages of each approach are explained through presentation of the results. The paper follows user oriented approach, by focusing on positive end-users' experience, a key driver of technology acceptance, adoption and usage behavior.

In our future work, we plan to continue focusing on QoE of the end-user in different production system [15], while using these approaches and provide additional results. We will provide some quantitative measures that can specify if a specific system can reach the business objective while further researching both approaches. This future work will focus on development of neuro-fuzzy model, which can be used to identify the causal relationship between input parameters of both objective (technical) and subjective (end-user oriented) nature, towards resulting QoE measurement.

References

1. IETF RFC 2474: Definition of the Differentiated Services Field (DS Field) in the IPv4 and IPv6 Headers. Available at http://tools.ietf.org/html/rfc2474 (1998)
2. Baldauf, M., Dustdar, S., Rosenberg, F.: A Survey on Context-aware Systems. In: Journal on Ad Hoc and Ubiquitous Computing (2007)
3. Chaari, T., Laforest, F., Celentano, A.: Design of Contextaware Application Based on Web Services. In: Technical Report CS-2004-5, Universit`a Ca'Foscari di Venezia, Venezia, Italy (2004)
4. Zeng, L., Lei, H., Chang, H.: Monitoring the QoS for Web Services. In: Lecture Notes in Computer Science, 2007, vol. 4749/2007, pp. 132-144 (2007)
5. Romero, D.: Context-aware Middleware: An overview, In: Paradigma 2,3 (2010)
6. Lee, M., Copeland, J.A.: An Adaptive End-to-end Delay Assurance Algorithm with DiffServ Architecture in IEEE 802.11e/IEEE 802.16 Hybrid Mesh/Relay Networks. In: Computer Communications and Networks (ICCCN). Proceedings of 18th Internatonal Conference on Computer Communications and Networks (ICCCN) (2009)
7. Yau, S.S., Karim, F.: A Context-sensitive Middleware-based Approach to Dynamically Integrating Mobile Devices into Computational Infrastructures. In: J.Paralle and Distributing Computing, 64(2), 301 (2004)
8. Liang, Q., Lau, H.C., Wu, X.: Robust Application-level QoS Management in Service-oriented Systems. In: 2008 IEEE International Conference on e-Business Engineering, pp. 239-246 (2008)
9. Zander, S., Nguyen, T., Armitage, G.: Automated Traffic Classification and Application Identification Using Machine Learning. In: LCN '05 Proceedings of the The IEEE Conference on Local Computer Networks 30th Anniversary, pp. 250 - 257 (2005)
10. Shirazi, B., Kumar, M., Sung, B.Y.: QoS Middleware Support for Pervasive Computing Applications. In: System Sciences (HICSS), Proceedings of the 37th Annual Hawaii International Conference on System Sciences (2004)
11. Xiong, P., Fan, Y., Zhou, M.: QoS-aware Web Service Configuration. In: Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on (2008)
12. IETF RFC 3246: An expedited forwarding PHB (Per-Hop behavior). Available at http://www.ietf.org/rfc/rfc3246.txt (2002)
13. Sidiroglou, S., Keromytis, A.D.: A Network Worm Vaccine Architecture. In: TWETICE '03 Proceedings of the Twelfth International Workshop on Enabling Technologies: Infrastructure for Collaborative Enterprises (2003)
14. Halima, R.B., Drira, K., Jmaiel, M.: A QoS-oriented Reconfigurable Middleware for Self-healing Web Services. In: 2008 IEEE International Conference on Web Services, pp.104-111 (2008)
15. Patrick, A. S., Singer, J., Corrie, B., No, S., El Khatib, K., Emond, B., Zimmerman, T., Marsh, S.: A QoE Sensitive Architecture for Advanced Collaborative Environments. In: First International Conference on Quality of Service in Heterogeneous Wired/Wireless Networks (QSHINE'04), pp.319-322 (2004)
16. Curran, P. J., West, S. G., Finch, J. F.: The robustness of test statistics to nonnormality and specification error in confirmatory factor analysis. In: Psychological Methods, Vol. 1, pp. 16–29 (1997)
17. Wikipedia: Quality of experience (QoE). Available at http://en.wikipedia.org/wiki/Quality_of_experience

Investigating the Users' Behavioral Intention Toward Using 3G Mobile VAS in Macedonia

Arsim Fidani

Natural Science and Math Faculty, Department of Informatics, State University of Tetova, 1200 Tetovo, Macedonia

arsim.fidani@unite.edu.mk

Abstract. The launch of 3G networks in Macedonia opened up a new revenue opportunity in mobile broadband offerings. Although the importance of 3G mobile value-added services (MVAS), it is surprising that Macedonia's society show little interest towards them. Based on extant literature, the factors which are observed as more influential in the 3G MVAS adoption, together with the factors derived from previously proposed models on technology adoption, such as technology acceptance model and Innovation Diffusion Theory (TAM, IDT), constitute the research model of this study.The results reveal that the total effect of the factors that influence the behavioral intention of Macedonian consumers to use 3G MVAS is ranked as follows: Compatibility, Perceived Enjoyment, Self – Efficacy, Perceived Cost, Perceived Ease of Use, Perceived Usefulness and Attitude. The results give telecom service providers in Macedonia an insight on the behavioral intention criteria of their customers, suggesting them to focus on educating their consumers about the offered services, foster the expansion of their 3G network coverage and also introduce new versatile and entertaining services which are easy to use in order to attract new customers and also to retain the old ones.

Keywords: 3G, mobile, value-added, services, Macedonia

1 Introduction

Mobile phones have become ubiquitous in our society and an integral part of people's everyday life. The advances in the mobile technology have increased the number of people using mobile services [1]. The growing number of mobile users and the decline in conventional voice service tariffs have gradually reduced average revenue per user (ARPU), thus decreasing the service providers profits [2]. In a 3G market, the major revenue source for telecommunications operators will originate from packet-based value-added services provided by independent value-added service providers, rather than traditional voice telephony [3]. Thus, imposing the mobile service providers to introduce various 3G MVAS, such as Mp3 ring tone download service, MMS, video news, photo download, mobile Java-based games, mobile TV etc. which have become a new opportunity for providers to create revenue. However, ARPU

S. Markovski, M. Gusev (Editors): ICT Innovations 2012, Web Proceedings, ISSN 1857-7288

could be substantially elevated when consumers are willing to use 3G MVAS and utilize them [2].

The annual report of AEC [4] in Macedonia reveals that the number of the mobile users is increasing, and until the 4th quarter of 2009 the penetration of mobile subscribers is 95%. Regarding the 3G services, in spite of the considerable investments by the service providers to take advantage offered by the new technologies, people in Macedonia show little interest towards adopting these services. According to the agency for electronic communication in Macedonia the penetration of the 3G users is about 15 % [4]. Given the difference between rapid growth rates in the adoption of mobile technologies and associated services in some countries and the relatively slow growth rates in others, such as Macedonia, is the first reason that makes the research worthwhile to conduct. The second reason is that 3G mobile services adoption and acceptance have been at the forefront of several research projects in different geographical and social context [5-7], however the use and the adoption intentions of the Macedonian users have not been investigated.

To take an extended perspective for examining consumers behavior [8], this study integrates Technology Acceptance Model [9], IDT [8, 10] and other significant external factors.

The results from this study give service providers in Macedonia an insight on the behavioral intention criteria of their customers regarding the offered 3G services, how to tailor particular services, understand the customers' needs and measures that telecom service providers should take to handle the adoption of these services.

2 Theoretical Framework and Hypothesis Development

Since the mid-1970s several theoretical models have been proposed [11] which are developed gradually and built up on each other [12].

TAM is intended to provide a conceptual model featuring a theoretic foundation and parsimony, to explain and predict the behavioral intention and practical behaviors of information technology (IT) users, based on the acceptance and use of IT [8]. On the other hand, IDT is also a theory associated with research on technology innovation, which tries to explain the innovation decision process, the determining factors of rate of adoption, and different categories of adopters. In addition to these theoretical models, there are some studies that have focused specifically on the users' intention to adopt mobile services [2, 7, 13-20].

To enhance the prediction of consumers' behavioral intention towards using of 3G MVAS in Macedonia while maintaining the model simplicity at the same time, four constructs from TAM and IDT are selected, such as: perceived ease of use, perceived usefulness, attitude and compatibility. Five other constructs, which are not found in traditional TAM and IDT model and are selected based on an extant literature review, are the following: self-efficacy, service availability, perceived cost, perceived enjoyment, and social influence. (see Fig.1)

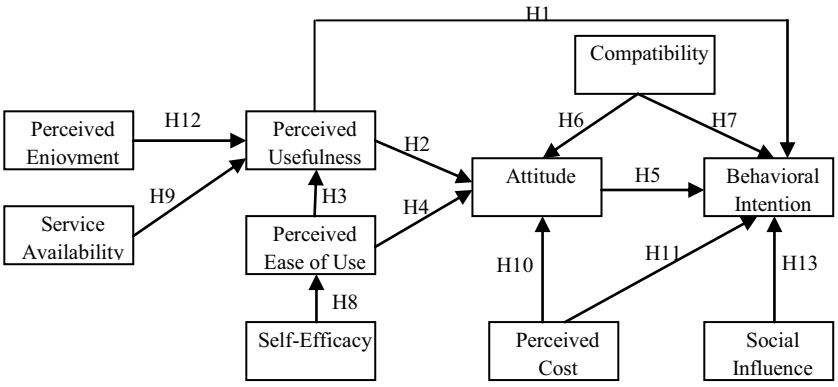


Fig. 1. Research framework

2.1 Perceived Usefulness, Perceived Ease of Use, Attitude and Compatibility

Perceived usefulness, first explored in the TAM, is found to be a crucial construct affecting the acceptance of the innovations. From the review of the existing literature is derived that perceived usefulness has a significant positive effect on user attitude [2, 15, 20]. On the other hand, higher perceived ease of use leads to higher perceived usefulness and it has a significant positive effect on user attitude [2]. Moreover, attitude is found to be determinant factor in individuals use intention and adoption of 3G MVAS [16] and having significantly positive effects on behavioral intention [2]. Compatibility, which has been selected from the IDT, is best described where an innovation is more likely to be adopted when it supports an individual's lifestyle, needs, job responsibilities and values [21]. Compatibility has been found to have a strong effect on users' attitudes and their behavioral intention [8].

The hypotheses are as follows:
H1: Perceived usefulness is positively related to behavioral intention of 3G MVAS
H2: Perceived usefulness has a positive influence on attitude of 3G MVAS
H3: Perceived ease of use has a positive influence on perceived usefulness of 3G MVAS.
H4: Perceived ease of use has a positive influence on attitude of 3G MVAS
H5: Attitude is positively related to behavioral intention of 3G MVAS.
H6: Compatibility is positively related to attitude of 3G MVAS
H7: Compatibility is positively related to behavioral intention of 3G MVAS

2.2 Self Efficacy, Service Availability, Perceived Cost, Perceived Enjoyment and Social Influence

In the context of this study, the more confident an individual is in his/her technical skills or the greater the experience he/she has with a cell phone, the more likely that 3G MVAS would be adopted [22]. On the other hand, service availability is defined

as the extent to which an information appliance is perceived as being able to provide pervasive and timely connections [13]. Regarding cost, many scholars considered it as an important factor affecting user's behavioral intention [2, 7, 13, 15, 18], to adopt 3G mobile services. The greater the cost the less likely would be the intention to adopt the technology. Perceived enjoyment, on the other hand, was proven to be an important antecedent of usage intentions and had a significant effect on mobile services [15, 23], while social influence has been found to have significant influence on the behavioral intention to use 3G mobile phones.

Thus, the following hypotheses are developed:
H 8: Self efficacy is positively related to perceived ease of use of 3G MVAS
H9: Service availability significantly affects perceived usefulness of 3G MVAS
H10: Perceived cost is negatively related to attitude of 3G MVAS
H11: Perceived cost is negatively related to behavioral intention of 3G MVAS
H12: Perceived enjoyment positively affects the perceived usefulness of 3G MVAS
H13: Social influence positively affects the behavioral intention of 3G MVAS

3 Research Methodology

This research adopts the positivist paradigm as it involves hypothesis testing through data collection and statistical analysis [24]. The quantitative approach was found to be most suitable for the purpose of this study, as the main objective of this study is to investigate the behavioral intention towards using 3G MVAS by individuals in Macedonia, by applying survey as a research strategy.

3.1 Sample, Questionnaire and Data Collection

The population of interest in this study is, all the individuals in Macedonia who own a 3G enabled mobile device. Because of the time constraints, and the fact of the respondents' availability in the specified places, the sampling frame is limited to the cities in western Macedonia Gostivar, Tetovo, including the capital Skopje, thus convenience sampling method best fits in this study. Beside the criticisms about this sampling technique, that the sample may not represent the whole population, as a non probabilistic sampling method does not mean that the sample is not a representative of the population, although it is very hard to represent the whole population well [25]. A sample of 500 respondents is used to gather the empirical data.

The questionnaire consists of two sections: the first section gathers the demographic information of the respondents, while the second section is about the respondents' perception about the 3G mobile services. The data collection is conducted with more focus on young people, mainly in universities; various mobile phone shops as well as prepaid phone reload shops, located in these cities. From the 500 distributed questionnaires, 361 valid responses were collected.

3.2 Validity and Reliability

Regarding the validity of this study, all the measures for constructs in this research are obtained from instruments which were already used and validated in existing and related literature [25], and it has been approved with the help of the research advisor and supervisor by which one type of validity - content validity is achieved. Moreover, to ensure validity of the study, factorial validity, as a favored technique in Information Systems [26], a subtype of construct validity is examined using factor analytic techniques.

Since the questionnaire used in this research contains constructs with multiple items the internal consistency reliability has been applied for each construct independently and also for all the items by using the Cronbach's alpha value, which range from 0.780 to 0.936 and the value for all the items in this study is equal to 0.877 that means it is internally consistent and acceptable. (see Table 2).

4 Data Analysis

The research model and the proposed hypotheses in this study are evaluated by the structural equation model. For the reliability coefficients, variance analysis, explanatory factor analysis, SPSS version 18 was used. Confirmatory factor analysis was conducted by LISREL version 8.80.

4.1 Demographic Characteristics

The demographic results reveal 76.5% of the respondents are male are and 23.5 % female. In terms of age most of the respondents from 19-23 years, which constitute 64.5 % of the total respondents.. For the educational background, most of the respondents have a bachelor degree constituting 67.9 % of the total number of the respondents, 21.9 % have finished secondary school, 8.6 % of the respondents have a master degree and the rest have doctorate 1.7 % and none of the respondents' is with a primary school. Among the 361 valid responses, only 32 respondents were reported to have used 3G MVAS. Out of 32, 27 users reported that they used 3G MVAS for more than 1 hour in a month, 3 users used the services for 30 min – 1 hour, and 2 users used the services for less than 10 min in a month. Moreover, among 32 users, 17 of them spent more than 600 denars on 3G MVAS each month, 7 respondents spent 300 to 600 denars, and 8 respondents have stated that they spent 100 to 300 denars on 3G MVAS each month.

4.2 Factor Analysis

The results' from the factor analysis reveal that partial correlations among variables are all greater than 0.5 and the factor model is appropriate whereby all the relationships are significant ($p < 0.05$). Moreover, convergent validity is adequate

when constructs have an Average Variance Extracted (AVE) of at least 0.5 [27] as shown in Table 2. Having the AVE values greater than the variance shared between a particular construct and other constructs in the model [28], the results reveal a satisfactory discriminate validity.

Table 1. Descriptive statistics of items

	KMO > 0.5	Bartlett's Test of Sphericity p < 0.005	Total Variance Explained	Cronobach alpha > 0.7	Average Variance Extracted > 0.5
PU	0.797	0.000	72 %	0.870	0.788
PEU	0.683	0.000	73 %	0.808	0.777
A	0.690	0.000	69 %	0.780	0.743
C	0.755	0.000	88 %	0.933	0.907
SE	0.730	0.000	80 %	0.875	0.840
SA	0.739	0.000	80 %	0.854	0.843
PC	0.680	0.000	74 %	0.823	0.783
PE	0.716	0.000	87%	0.927	0.903
SI	0.703	0.000	72 %	0.810	0.770
BI	0.756	0.000	88 %	0.936	0.910

4.3 Hypothesis Testing

In the confirmatory factor model, the researcher imposes substantively motivated constraints that determine which pairs of common factors are correlated, which observed variables are affected by which common factors, which observed variables are affected by a unique factor, and which pairs of unique factors are correlated [29]. Therefore the estimated path coefficients of the structural model were studied to evaluate the hypotheses presented in Fig 1. The hypothesis results are summarized in Table 3.

Table 3**.** Hypothesis results

Hypothesis	Effect	T-value	Results
H1	PU → BI	2.50	Accepted
H2	PU → A	1.35	Rejected
H3	PEU → PU	3.26	Accepted
H4	PEU →A	0.38	Rejected
H5	A → BI	2.13	Accepted
H6	C → A	0.06	Rejected
H7	C → BI	12.41	Accepted
H8	SE → PEU	6.65	Accepted
H9	SA → PU	1.23	Rejected
H10	PC → A	4.09	Accepted
H11	PC → BI	1.71	Rejected
H12	PE → PU	11.46	Accepted
H13	SI → BI	0.08	Rejected

4.4 Goodness of Fit

Confirmatory factor analysis (CFA) is used to assess the structural model fit. Table 4 shows the common model-fit indices, recommended values and results of the test of structural model fitness, indicate a good model fit.

Table 4. Fit indices for structural model

Fit Indices	Recommended Value	Result
Chi-square/Degree of Freedom (χ^2 / df)	<3	2.288
Goodness of-Fit Index (GFI)	>0.8	0.86
Adjusted Goodness of Fit Index (AGFI)	>0.8	0.83
Root Mean Square Error of Approximation (RMSEA)	<0.08	0.060
Root Mean Square Residual (RMR)	<0.08	0.08
Normed Fit Index (NFI)	>0.9	0.92
Comparative Fit Index (CFI)	>0.9	0.95

5 Discussion

The results show that Perceived Usefulness (**H1**), consistent with prior studies [30], in this study is found to be significant determinant to predict the Behavioral Intention of 3G MVAS. Thus, in order for users to use the 3G MVAS, they must feel that these services are useful to them, such as increasing their efficiency in their life and work, increasing their mobility and providing them with better internet surfing. Hence, telecom service providers should focus on informing and educating their users about the usefulness of the offered 3G MVAS in order to retain and also attract new consumers. The relationship between Perceived Usefulness and Attitude (**H2**) is not significant. The outcome of these results in this study could be ascribed to the fact that a small number of respondents use the actual offered 3G MVAS, or the range and the availability of the services offered by the telecom service providers in Macedonia.

The results reveal that Perceived Ease of Use has a significantly positive effect on Perceived Usefulness (**H3**). These results imply that telecom service providers need to consider the ease of use of services when identifying the 3G MVAS that can offer practical values to consumers and also telecom service providers should focus on developing customized user interfaces that are easy to use. On the other hand, the relationship between Perceived Ease of Use and Attitude (**H4**) is not found significant.

Next, the results indicate that Attitude (**H5**) significantly affects Behavioral Intention of 3G MVAS. As there are Attitude's antecedents or factors that influence the consumers' attitudes, therefore telecom service providers should take into consideration all the recommendations presented for each factor in order for the Macedonian consumers to have positive attitudes towards behavioral intention of 3G MVAS.

Compatibility is found to have an insignificant effect on users' Attitude (**H6**), but on the other hand it shows a high effect on Behavioral Intention (**H7**), implying that

telecom service providers have to emphasize how their offered services fit with the targeted group's lifestyle.

Self-Efficacy in this study is found to have a significant influence on Perceived Ease of Use (**H8**), which is consistent with previous studies [19]. These results from the third highest ranked factor, Self-Efficacy, are attributed to the fact that respondents educational level is high, consistent with the theoretical argument [9] and [31] that an individual with high expertise might rate a system as easier to use than an individual with lower expertise. This implies that telecom service providers should focus on developing customized user interfaces that are easy to use in order to attract new consumers.

The relationship between Service Availability and Perceived Usefulness is found to be insignificant (**H9**), which contradicts to findings that Service Availability to have a strong effect on Perceived Usefulness [13]. Consistent with the results of a prior study [2], (**H10**) Perceived Cost is found to have a significant influence on Attitudes, but contradicting to their results regarding the relationship between Perceived Cost and Behavior Intention (**H11**), which is in this study is found to be insignificant. The result from this relationship are consistent with prior [20] findings, that the results might indicate that the users are more concerned with what 3G can offer rather than its costs. However, in this study the results from Perceived Cost are attributed to the fact that Macedonian consumers perceive that the tariff of these services is still high. Therefore, telecom service providers are recommended to adopt promotions to reduce the threshold of service tariff in order to promote 3G MVAS. If the service tariff cannot be reduced, then service providers should develop more valuable and special services, so that consumers can enjoy the benefits or effectiveness of 3G MVAS at a lower cost.

Regarding Perceived Enjoyment, the results in this study indicate that this construct has a significant influence on Perceived Usefulness (**H12**) which in turn had a significant influence on Behavioral Intention of 3G MVAS. These findings suggest that users of 3G mobile services need to be provided with more diverse and entertaining ways, which can greatly contribute to the efficiency and convenience of communicating. Social influence, in this study, exhibits insignificant correlation on Behavioral Intention of 3G MVAS (**H13**), which contradicts with prior findings [20] and [32]. The outcome of these results in this study could be explained by the statement that the relationship between social influence and behavioral intention exhibited different results for different service categories, in some, the relationship was significant and in others it was insignificant [13]. This implies that the service category of mobile services can be an important boundary condition in explaining consumers' behavioral intention.

6 Conclusion

The results of this study reveal that the total effect of the factors that influence the behavioral intention of Macedonian consumers to use 3G MVAS, is ranked as follows: Compatibility, Perceived Enjoyment, Self – Efficacy, Perceived Cost, Perceived Ease of Use, Perceived Usefulness and Attitude. Thus implying that telecom service providers in Macedonia should consider the above mentioned

recommendations, such as focus on educating their consumers about the offered services, foster the expansion of their 3G network coverage and also introduce new versatile and entertaining 3G MVAS which are easy to use, in order to attract new customers and also to retain the old ones. Moreover, as a theoretical contribution, contrary to many technology adoption models in the past studies [2, 7, 9, 13, 15, 19], the model used in this study shows that both perceived usefulness and perceived ease of use have an insignificant effect on attitude, rather, only perceived usefulness shows to be significantly related to behavioral intention directly, which is in line with the extended technology acceptance model [30].

Since the respondents that constitute the main user group of 3G MVAS are relatively young, future studies should include users who are older. Further, the model used in this study measures perceptions and intentions at a single point in time, and the 3G telecom service market in Macedonia is still under development, implying for a longitudinal study, and also an in-depth investigation is also required in order to acquire more objective arguments when consumers have a higher level of involvement in 3G. Future research should also consider the demographic characteristics of the respondents, and may also consider other external factors such as culture [20], which allows the researchers to conduct a multi-country comparison study.

Acknowledgements

I would like to express my gratitude to my supervisor Dr. Jaime Campos, and also to Dr. Anita Mirijamdotter and Dr. Jan Aidemark for their support, insight and diligent guidance. This research was done at Linnaeus University as part of master's thesis.

References

1. Tang, L.: Key success factors in 3G services adoption: a consumer perspective. (2008) 1-7
2. Kuo, Y., Yen, S.: Towards an understanding of the behavioral intention to use 3G mobile value-added services. Computers in Human Behavior 25 (2009) 103-110
3. Gazis, V., Koutsopoulou, M., Farmakis, C., Kaloxylos, A.: A flexible charging and billing approach for the emerging UMTS network operator role. SIMULATION SERIES 33 (2001) 85-94
4. AEC: Report on the development of the electronic communication market in the 4th quarter,. AEC Macedonia Retrieved, from: http://www.aek.mk/ (2009)
5. López-Nicolás, C., Molina-Castillo, F., Bouwman, H.: An assessment of advanced mobile services acceptance: Contributions from TAM and diffusion theory models. Information & Management 45 (2008) 359-364
6. Rao Hill, S., Troshani, I.: Factors influencing the adoption of personalisation mobile services: empirical evidence from young Australians. International Journal of Mobile Communications 8 (2010) 150-168
7. Sun, Q., Cao, H., You, J.: Factors influencing the adoption of mobile service in China: An integration of TAM. Journal of Computers 5 (2010) 799
8. Chen, L., Gillenson, M., Sherrell, D.: Enticing online consumers: an extended technology acceptance perspective. Information & Management 39 (2002) 705-719
9. Davis, F.: Perceived usefulness, perceived ease of use, and user acceptance of information technology. Mis Quarterly 13 (1989) 319-340

10. Hsu, C., Lu, H., Hsu, H.: Adoption of the mobile Internet: An empirical study of multimedia message service (MMS). Omega 35 (2007) 715-726
11. Compeau, D., Higgins, C.: Computer self-efficacy: Development of a measure and initial test. Mis Quarterly 19 (1995) 189-211
12. Sendecka, L.: Adoption of mobile services: Moderating effects of service's information intensity. NORGES HANDELSHOY University, Norway (2006)
13. Hong, S., Tam, K.: Understanding the adoption of multipurpose information appliances: The case of mobile data services. IS research 17 (2006) 162
14. Wu, Y., Tao, Y., Yang, P.: Using UTAUT to explore the behavior of 3G mobile communication users. (2007) 199-203
15. Pagani, M.: Determinants of adoption of third generation mobile multimedia services. Journal of Interactive Marketing 18 (2004) 46-59
16. Bouwman, H., Carlsson, C., Walden, P., Molina-Castillo, F.: Reconsidering the actual and future use of mobile services. Information Systems and E-Business Management 7 (2009) 301-317
17. Luarn, P., Lin, H.: Toward an understanding of the behavioral intention to use mobile banking. Computers in Human Behavior 21 (2005) 873-891
18. Agarwal, N., Wang, Z., Xu, Y., Poo, D.: Factors Affecting 3G Adoption: An Empirical Study. PACIS 2007 Proceedings (2007) 3
19. Wang, Y., Lin, H., Luarn, P.: Predicting consumer intention to use mobile service. Information Systems Journal 16 (2006) 157-179
20. Chong, A., Darmawan, N., Ooi, K., Lin, B.: Adoption of 3G services among Malaysian consumers: an empirical analysis. International Journal of Mobile Communications 8 (2010) 129-149
21. Agarwal, R., Prasad, J.: The role of innovation characteristics and perceived voluntariness in the acceptance of information technologies. Decision Sciences 28 (1997) 557-582
22. Taylor, S., Todd, P.: Understanding information technology usage: A test of competing models. Information systems research 6 (1995) 144-176
23. Nysveen, H., Pedersen, P., Thorbjørnsen, H.: Intentions to use mobile services: antecedents and cross-service comparisons. Journal of the Academy of Marketing Science 33 (2005) 330-346
24. Orlikowski, W., Baroudi, J.: Studying information technology in organizations: Research approaches and assumptions. Information systems research 2 (1991) 1-28
25. Trochim, P.: Research methods. Dreamtech Press (2003)
26. Straub, D., Boudreau, M., Gefen, D.: Validation guidelines for IS positivist research. Communications of the Association for IS 13 (2004) 380-427
27. Fornell, C., Larcker, D.: Structural equation models with unobservable variables and measurement error: Algebra and statistics. Journal of Marketing Research 18 (1981) 382-388
28. Chin, W.: Commentary: Issues and opinion on structural equation modeling. Vol. 22. JSTOR (1998)
29. Long, J.: Confirmatory factor analysis: A preface to LISREL. Sage Publications, Inc (1983)
30. Venkatesh, V., Davis, F.: A theoretical extension of the technology acceptance model: four longitudinal field studies. Management science 46 (2000) 186-204
31. Mathieson, K.: Predicting user intentions: comparing the technology acceptance model with the theory of planned behavior. Information systems research 2 (1991) 173-191
32. Venkatesh, V., Morris, M., Davis, G., Davis, F., DeLone, W., McLean, E., Jarvis, C., MacKenzie, S., Podsakoff, P., Chin, W.: User acceptance of information technology: Toward a unified view. Information & Management 27 (2003) 425-478

An Improved 3-Quasigroup based Encryption Scheme

Sucheta Chakrabarti[1], Saibal K. Pal[1], Sugata Gangopadhyay[2]
[1]SAG, DRDO, India
suchetadrdo@hotmail.com, skptech@yahoo.com
[2]Indian Statistical Institute, India
gsugata@gmail.com

Abstract. The Crypto-community is always in search of new strong crypto-primitives to handle the present security threats and for providing efficient secure digital communication. One of the main goals of the cryptographer is to make an encryption scheme computationally fast with optimized use of memory and high cryptographic complexity. In this direction n-quasigroups (n = 2, 3) are considered as a class of new strong crypto-primitives. In this paper we propose an improved version of 3-quasigroup based encryption scheme given by Petrescu. Here we only consider reducible 3-quasigroup as the seed. It is randomly generated based on a secret key. The process of deriving different 3-ary operations for construction of reducible 3-quasigroups is described together with some related issues. We also present experimental results on 4-order reducible 3-quasigroups which are generated to find the different cases suitable for cryptographic applications.

Keywords: Quasigroup, quasi algebra, reducible 3-quasigroup, isotopy, order of quasigroup, encryption, decryption, stream cipher

1. Introduction

Stream ciphers [7], [11], [12] play an important role in secure digital communication. This category of encryption schemes is fast in implementation and can be used in applications with less computational resources. Stream ciphers form a class of symmetric key cryptographic algorithms. Here the crypto primitives are mainly the Pseudo-Random Bit Generators (PRBG) based on nonlinear combinations of different Linear Feedback Shift Registers (LFSR). Sequences based on LFSRs [7] have been widely used in the design of stream ciphers in the past decades. Today, one of the main research areas is to explore and identify new crypto primitives which are optimised both in terms of security & efficiency. In this direction, present research shows that algebraic structures based on quasigroups or n-quasigroups [4], [5], [6], [8] are suitable crypto primitives for stream ciphers and other cryptographic applications. Quasigroups [1]

S. Markovski, M. Gusev (Editors): ICT Innovations 2012, Web Proceedings, ISSN 1857-7288

have applications in coding theory also. Different transformations of quasigroups play an important role in ensuring security of the crypto algorithms based on these structures. Isotopy is one of the widely used transformation which is used to generate large number of quasigroups [3] and increase the security of the scheme.

Quasigroup based encryption scheme working as a stream cipher was first proposed by Koscielny [2] in 1996. Design and analysis of different encryption schemes for stream ciphers based on quasigroups have motivated researchers to generalize it for 3-quasigroups. These schemes have been able to exponentially increase the number of 3-quasigroups due to isotopic operations. Petrescu [8] has given a 3-quasigroup based encryption scheme for stream ciphers. The author considers the 3-quasigroup which acts as publicly known seed. In this paper, we have improved 3-quasigroup based encryption scheme given in [7] by randomly generating the quasigroup based on a key and then deriving the reducible 3-quasigroup (which plays the role of the seed for the encryption scheme). Section 2 contains the definitions and brief descriptions of the fundamental concepts related to this paper and Petrescu's encryption scheme. In Section 3 the process to randomly generate a quasigroup based on the key [10] is given. The process to derive reducible 3-quasigroups is discussed in Section 4. In Section 5 description of the improved version of the 3-quasigroup based encryption scheme is given. We have reported observations and inferences based on experimental results in Section 6 and conclusions are drawn in Section 7.

2. Preliminaries

In this section we present fundamental definitions related to this work and provide brief description of Petrescu's encryption scheme [9].

Definition 1. A Quasigroup $\langle A,\circ \rangle$ is a groupoid consisting of elements of A w.r.t. a binary operation $'\circ'$ such that $\forall a,b \in A$ there exist unique $x,y \in A$ for which it satisfies the identities $a \circ x = b \ \& \ y \circ a = b$

This means that every row and every column of the Cayley's table is a permutation of A. The cardinality of A is called the order of the quasigroup. For every finite quasigroup of order m, given by the Cayley table, it can be equivalently associated with the combinatorial design viz. a $m \times m$ Latin square.

Definition 2. An algebraic quasigroup $(A,\circ,/,\backslash)$ is an algebra with 3 binary operations which satisfy the following identities:

$(x \ / \ y) \circ y = x = (x \circ y) \ / \ y \ \ \& \ \ x \circ (x \backslash y) = y = x \backslash (x \circ y)$

Definition 1 & 2 are same when A is finite and hence both of them are called quasigroup. By quasigroup we mean the binary quasigroup denoted by 2-quasigroup. An algebraic quasigroup can be represented in general as follows:
$\langle A, \alpha, \alpha_1, \alpha_2\rangle$ is an algebra where $\alpha, \alpha_1, \alpha_2$ are three binary operations satisfying the following identities:

$$\alpha(\alpha_1(x_1,x_2),x_2) = x_1 = \alpha_1(\alpha(x_1,x_2),x_2) \;\&\; \alpha\left(x_1,\alpha_2(x_1,x_2)\right) = x_2 = \alpha_2\left(x_1,\alpha(x_1,x_2)\right)$$

Definition 3. A Ternary quasigroup (3-quasigroup) is a finite algebra $\langle A, \alpha, \alpha_1, \alpha_2, \alpha_3\rangle$ consisting of the elements of A and 4 ternary operations satisfying the following identities:

$$\alpha(\alpha_1(x_1,x_2,x_3),x_2,x_3) = x_1 = \alpha_1(\alpha(x_1,x_2,x_3),x_2,x_3)$$
$$\alpha(x_1,\alpha_2(x_1,x_2,x_3),x_3) = x_2 = \alpha_2(x_1,\alpha(x_1,x_2,x_3),x_3)$$
$$\alpha(x_1,x_2,\alpha_3(x_1,x_2,x_3)) = x_3 = \alpha_3(x_1,x_2,\alpha(x_1,x_2,x_3))$$

If the 3-quasigroup is derived from the quasigroup then it is called the **reducible 3-quasigrouop.** We can generalise this definition to n-quasigroup.

Definition 4. Isotopy of 3-quasigroup is defined as follows:
Let $\langle A_1,\alpha\rangle$ & $\langle A_2,\beta\rangle$ be two 3-quasigroups. A_1 is isotopic to A_2 if there are four bijections
$f, f_1, f_2, f_3 : A_1 \to A_2$ such that $f(\alpha(x_1,x_2,x_3)) = \beta(f_1(x_1), f_2(x_2), f_3(x_3))$. The ordered quadruple (f, f_1, f_2, f_3) is called an isotopism or isotopy.

Next, we briefly describe Peterscu's [9] **3-quasigroup based encryption scheme** for stream ciphers. Let $\langle A, \alpha, \alpha_1, \alpha_2, \alpha_3\rangle$ be a publicly known 3-quasigroup which is used as seed and isotopic carrier. Let $K = A^8 \times \{1,2,3\}$ be the key space. The key is represented as $k = a_1a_2 \ldots a_8 i$ and it determines another isotopic quasigroup operation β on A as follows: $\beta(x_1, x_2, x_3) = f_4\left(\alpha\left(f_1^{-1}(x_1), f_2^{-1}(x_2), f_3^{-1}(x_3)\right)\right)$, where $f_j = f_{a_j}$ are permutations on A based $a_1a_2a_3a_4$ of k and $a_5a_6a_7a_8$ of k are Initial Value (IV) for Encryption / Decryption (E_k/D_k). Every key k uniquely determines the E_k/D_k s.t. $D_k(E_k(m)) = m$.
Let $m = m_1m_2$ ... and the encryption function $E_k(m)$ =c_1c_2 ... is defined as follows:

$i = 1$	$i = 2$	$i = 3$
$c_1 = \beta(m_1, a_5, a_6)$	$c_1 = \beta(a_5, m_1, a_6)$	$c_1 = \beta(a_5, a_6, m_1)$
$c_2 = \beta(m_2, a_7, a_8)$	$c_2 = \beta(a_7, m_2, a_8)$	$c_2 = \beta(a_7, a_8, m_2)$
$for\ j > 2$	$for\ j > 2$	$for\ j > 2$
$c_j = \beta(m_j, c_{j-2}, c_{j-1})$	$c_j = \beta(c_{j-2}, m_j, c_{j-1})$	$c_j = \beta(c_{j-2}, c_{j-1}, m_j)$

The Decryption function $D_k(c_1, c_2, \dots) = m_1 m_2 \dots$ is defined as follows:

$i = 1$	$i = 2$	$i = 3$
$m_1 = \beta_1(c_1, a_5, a_6)$	$m_1 = \beta_2(a_5, c_1, a_6)$	$m_1 = \beta_3(a_5, a_6, m_1)$
$m_2 = \beta_1(c_2, a_7, a_8)$	$m_2 = \beta_2(a_7, c_2, a_8)$	$m_2 = \beta_3(a_7, a_8, m_2)$
$for\ j > 2$	$for\ j > 2$	$for\ j > 2$
$m_j = \beta_1(c_j, c_{j-2}, c_{j-1})$	$m_j = \beta_2(c_{j-2}, c_j, c_{j-1})$	$m_j = \beta_3(c_{j-2}, c_{j-1}, m_j)$

In the next section, we describe the process to generate a quasigroup randomly based on a secret key [5].

3. Randomly Generated Quasigroup based on a Key

We present two methods to generate any m-order quasigroup based on a key.

Method I - Let $A = \{1,2, \dots, m\}$ be a set of m elements. To randomly generate the quasigroup of order m, the required key length is $2m$. So the key space is as follows: $K = \{a_1 \dots a_{2m} | \ a_j \in A, 1 \le j \le 2m \}$. Let the key be $k = a_1 \dots a_{2m}$. First, we take the basic quasigroup $\langle A, \alpha \rangle$ and represented by the matrix A_s whose 1st row is the identity permutation of A and other $m - 1$ rows are derived by left shift of one character of the previous row

$$A_s = \begin{bmatrix} 1 & 2 & 3 & \cdots & m \\ 2 & 3 & \cdots & m & 1 \\ & & \vdots & & \\ m & 1 & \cdots & m-1 & \end{bmatrix}$$

Now, by applying the following row swapping process based on $a_1 \cdots a_m$ of the key k we derive the matrix A_{sr}. Here we take $x \bmod m \equiv m \ if \ m|x$

$$temp = A_s$$
$$for\ \ j = 1{:}\, m$$
$$\quad r_1 = (j + a_j) mod\ m$$
$$\quad r_2 = \left(\left\lceil \frac{m}{2} \right\rceil * j + 1 + a_j\right) mod\ m$$
$$Swap\ \ r_1\ and\ r_2\ of\ temp$$
$$A_{sr} = temp$$

Similarly, as above based on $a_{m+1} \cdots a_{2m}$ of the key k the computation is carried out on A_{sr} by swapping the column m times. This process will generate a quasigroup $\langle A, \gamma \rangle$ which we called the Initial Quasigroup and represented by the matrix A_f.

If the order of the quasigroup $m > 16$ then it requires a key of more than 128 bits in length. In this case for practical reasons we can follow the 2nd Method which is given below for any $m = 2^n > 16$.

Method II – Let $A = \{1,2, \dots, m\}$ *be t*he set of $m = 2^n > 16$ elements. The key space consists of
$K = \{a_1 \dots a_{16} | \, a_j \in A, 1 \leq j \leq 2m\}$. Let the key $= a_1 \dots a_{16}$. First, take $a_1 \cdots a_8$ of the key k as the seed of a PRBG to generate $2 * 8n$ bits. Then we take n bits at a time and convert it to the decimal value $a'_j \in A$. So we derive a new key $k' = a'_1 \cdots a'_{16}$ .. Based on the key k' we calculate two row values as follows:

$for \;\; j = 1{:}\,16$
$for \; i = 1{:}\,100$
$r_1 = (j + a_j) mod \; m$
$r_2 = \left(\left\lceil \frac{m}{2} \right\rceil * j + 1 + a_j\right) mod \; m$
$Swap \;\; r_1 \; and \; r_2 \; of \; A_s$ to derive A_{sr}

Similarly, based on $a_9 \cdots a_{16}$ of the key k we generate another key k'' and calculate two column values as above and swap the columns of A_{sr} to generate the Initial Quasigroup $\langle A, \gamma \rangle$. Example of generating initial quasigroup of order 4 based on a key by using Method I is given below:

Example: Let the order of $|A| = 4$ and let the key $k = a_1 \cdots a_8 = 4\,1\,3\,1\,2\,1\,4\,3$

Here , $\left\lceil \frac{m}{2} \right\rceil = 2$ and

$$A_s = \begin{bmatrix} 1\,2\,3\,4 \\ 2\;\;3\,4\,1 \\ 3\,4\,1\,2 \\ 4\,1\,2\,3 \end{bmatrix}$$

After swapping the rows for 4 times as described in Method I

$$A_{sr} = \begin{bmatrix} 1\,2\,3\,4 \\ 3\,4\,1\,2 \\ 2\,3\,4\,1 \\ 4\,1\,2\,3 \end{bmatrix}$$

Now we apply the column swapping as described in Method I and finally derive the Initial Quasigroup

$$A_f = \begin{bmatrix} 3\,1\,4\,2 \\ 1\,3\,2\,4 \\ 4\,2\,1\,3 \\ 2\,4\,3\,1 \end{bmatrix}$$

In the next section we explain the process to derive the 3-quasigroup from the Initial Quasigroup.

4. Generation of Reducible 3-quasigroup based on a Key

Based on a key we generate the Initial quasigroup $\langle A, \gamma \rangle \equiv \langle A, \gamma, \gamma_1 \gamma_2 \rangle$. Now by using these 2-ary operations $\gamma, \gamma_1 \gamma_2$ on A we define 3-ary operations $\alpha, \alpha_1 \alpha_2 \alpha_3$ on A. There are many ways we can define 3-ary operation α from $\gamma, \gamma_1 \; and \; \gamma_2$ by using their non-associative & non-commutative properties. So, for each initial quasigroup generated based on a key the following ternary operations are defined and are arranged in the lexicographical order as follows:

1. $\alpha(x_1, x_2, x_3) = \gamma(\gamma(x_1, x_2), x_3)$
2. $\alpha(x_1, x_2, x_3) = \gamma\big(x_1, \; \gamma(x_2, x_3)\big)$
3. $\alpha(x_1, x_2, x_3) = \gamma_1(\gamma_1(x_1, x_2), x_3)$
4. $\alpha(x_1, x_2, x_3) = \gamma_1\big(x_1, \; \gamma_1(x_2, x_3)\big)$
5. $\alpha(x_1, x_2, x_3) = \gamma_2(\gamma_2(x_1, x_2), x_3)$
6. $\alpha(x_1, x_2, x_3) = \gamma_2\big(x_1, \; \gamma_2(x_2, x_3)\big)$
7. $\alpha(x_1, x_2, x_3) = \gamma(x_1, \; \gamma_1(x_2, x_3),)$
8. $\alpha(x_1, x_2, x_3) = \gamma(\gamma_2(x_1, x_2), x_3)$
9. $\alpha(x_1, x_2, x_3) = \gamma_2(\gamma(x_1, x_2), x_3)$
10. $\alpha(x_1, x_2, x_3) = \gamma_1(\gamma_2(x_1, x_2), x_3)$
11. $\alpha(x_1, x_2, x_3) = \gamma_2(\gamma_1(x_1, x_2), x_3)$
12. $\alpha(x_1, x_2, x_3) = \gamma_2\big(x_1, \; \gamma_1(x_2, x_3)\big)$

We derive different 3-ary operations on the set A and then for each 3-ary operation α we derive $\alpha_1, \alpha_2, \alpha_3$ by using the identities of $\gamma, \; \gamma_1, \gamma_2$ so that $\langle A, \alpha, \alpha_1, \alpha_2, \alpha_3 \rangle$ is a 3-quasigroup. Hence from an Initial Quasigroup we can generate a large number of 3-quasigroups which plays the role of the seed for the encryption scheme. Here we present an example of generation of a reducible 3-quasigroup.

Example: Let the Initial Quasigroup generated from a key be represented by a matrix as follows:

$$P = \begin{bmatrix} 4\,1\,3\,2 \\ 3\,2\,1\,4 \\ 2\,3\,4\,1 \\ 1\,4\,2\,3 \end{bmatrix}$$

It represents a Latin square which is the inner body of the Caley's table and hence it defines γ. Now we define two matrices which represent γ_1, γ_2 of the Initial Quasigroup $\langle A, \gamma, \gamma_1 \gamma_2 \rangle$ as follows:

$$P_1 = \begin{bmatrix} 4\,1\,2\,3 \\ 3\,2\,4\,1 \\ 2\,3\,1\,4 \\ 1\,4\,3\,2 \end{bmatrix} \qquad P_2 = \begin{bmatrix} 2\,4\,3\,1 \\ 3\,2\,1\,4 \\ 4\,1\,2\,3 \\ 1\,3\,4\,2 \end{bmatrix}$$

Define $\alpha = \gamma(\gamma(x_1, x_2), x_3)$ and consequently $\alpha_1 = \gamma_1(\gamma_1(x_1, x_3), x_2)$, $\alpha_2 = \gamma_2(x_1, \; \gamma_1(x_2, x_3),)$ and $\alpha_3 = \gamma_2(\gamma(x_1, x_2), x_3)$. Then $\langle A, \alpha, \alpha_1, \alpha_2, \alpha_3 \rangle$ is a reducible 3-quasigroup. In the next section we present our modified 3-quasigroup based encryption scheme for stream cipher.

5. Improved Version of 3-quasigroup based Encryption Scheme

The quasigroup based encryption scheme given by Peterscu [9] considered that the seed quasigroup is known to public. So it is to be sent to the receiver which increases the communication overhead. In this case only the key which control the isotopy, Initial Value (IV) and Encryption (E_k) / Decryption (D_k) functions are unknown. The key complexity in this case is $|A|^8 * \{1,2,3\}$. This is not sufficient for providing security under practical scenario with respect to the present computational power (assumed to be) available to the attacker. We have modified this scheme to improve its cryptographic strength.

Let $K = K_1 \times K_2$ where the first part of the key space K_1 is as described in Section 3 & $K_2 = \{b_1 \cdots b_8 i \mid b_j \in A\ ,\ i \in \{1,2,3\}\}$. So for $m \leq 16$, we take $k_1 = a_1 \cdots a_{2m}$ & $k_2 = b_1 \cdots b_8\, i$ and for $m > 16$ & $m = 2^n$ where $n \in Z_+$ we take $k_1 = a_1 \cdots a_{16}$ & $k_2 = b_1 \cdots b_8\, i$. So the key is represented as $k = k_1 k_2$.

Now based on the key k_1 the random initial quasigroup is generated as described in Section 3. Also, the list of 12 different 3-ary operation α is assumed to be known. Now, based on the key $k_1$we get order as $\sum_{i=1}^{2m} a_i \bmod 12$ or $\sum_{i=1}^{16} a_i \bmod 12$ where we denote $mod\ 12 \equiv 12\ if\ 12|\ x$. So we fix α by the key k_1. By using the initial quasigroup and α we generate the reducible 3-quasigroup which is taken as the seed 3-quasigroup for the Encryption scheme. For each element $\in A$, let f_a denotes the permutation of A. We define the isotopy of $\langle A,\ \alpha, \alpha_1, \alpha_2\ , \alpha_3\rangle$ as follows:

The permutation $f_j = f_{b_j}$, $1 \leq j \leq 4$ is defined by $b_1\, b_2\, b_3\, b_4$ of k_2 as given below:

$f_j = f_{b_j} = \alpha_j(b_j, b_{j+1}, x)$; $1 \leq x \leq |A|$ and $j + 1 = j + 1\ mod\ 4\ \ if\ j + 1 > 4$

So $(f_4\ ,\ f_1\ ,\ f_2\ ,\ f_3)$ is an isotopy of 3-quasigroup $\langle A,\ \alpha, \alpha_1, \alpha_2\ , \alpha_3\rangle$. So, by applying the isotopy on the seed 3-quasigroup we generate the new isotopic 3-quasigroup $< A, \beta, \beta_1, \beta_2, \beta_3 >$ as follows:

$$\beta(x_1, x_2,\ x_3) = f_4\left(\alpha\left(f_1^{-1}(x_1), f_2^{-1}(x_2), f_3^{-1}(x_3)\right)\right)$$
$$\beta_1(x_1, x_2,\ x_3) = f_1\left(\alpha_1\left(f_4^{-1}(x_1), f_2^{-1}(x_2), f_3^{-1}(x_3)\right)\right)$$
$$\beta_2(x_1, x_2,\ x_3) = f_2\left(\alpha_2\left(f_1^{-1}(x_1), f_4^{-1}(x_2), f_3^{-1}(x_3)\right)\right)$$
$$\beta_3(x_1, x_2,\ x_3) = f_3\left(\alpha_3\left(f_1^{-1}(x_1), f_2^{-1}(x_2), f_4^{-1}(x_3)\right)\right)$$

Now $i = 1\ ,\ 2,\ or\ 3\ of\ k_2$ uniquely determines the (E_k, D_k) as $(\beta\ , \beta_1)$, $(\beta\ , \beta_2)$ or $(\beta\ , \beta_3)$ respectively (as defined in Section 2) where $b_5\, b_6\, b_7\, b_8$ of k_2 play the role of IV for the encryption scheme.

The key complexity of our scheme for 3-quasigroups of order 4 is $2^{32} \times \{1,2,3\}$ where as in earlier case it is only $2^{16} \times \{1,2,3\}$.However due to order 4 total number of

3-quasigroups is $\ll 2^{32}$ For practical purposes, 3-quasigroups of order $256 = 2^8$ are commonly used. In our proposed modified scheme the key complexity is $(2^8)^{16} * (2^8)^8 \times \{1, 2, 3\} \approx 2^{195}$ instead of 2^{65}.For such cases number of 3-quasigroups is $\gg 2^{195}$. This key complexity is considered secure even in today's computational scenario. The block diagram of the modified scheme is given in the Appendix.

In the following section we present observations and some inferences from the experimental results.

6. Observations & Experimental Results

We have carried out experiments on the algorithm based on different reducible 3-quasigroup seed and different type of inputs under different keys. Here we discuss some important observations and inference from these experiments. Experimental results of different cases of a seed are given below:

Experiment

We take the quasigroup $\langle A, \gamma \rangle \equiv \langle A, \gamma, \gamma_1, \gamma_2 \rangle$ of order 4 where $A = \{1,2,3,4\}$ and the binary operation γ represented by matrix $P = [4\ 1\ 3\ 2;\ 3\ 2\ 1\ 4; 2\ 3\ 4\ 1; 1\ 4\ 2\ 3]$ which is a Latin Square of order 4 and hence derive P_1, P_2 which represented γ_1, γ_2. We define 3-ary operation as $\alpha = \gamma(\gamma(x_1, x_2), x_3)$ and hence calculate $\alpha_1, \alpha_2, \alpha_3$ to derive the 3-quasigroup $\langle A, \alpha \rangle \equiv \langle A, \alpha, \alpha_1, \alpha_2, \alpha_3 \rangle$, which is a reduced 3-quasigroup to be considered as a seed for the algorithm.

Case 1

(a) **Binary operation representation** P: [4 1 3 2 ; 3 2 1 4 ; 2 3 4 1;1 4 2 3]

2nd part of the KEY (k_2): [2 4 1 3 3 4 2 1 1]

Message: [1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1]

P_1 = [4 1 2 3 ; 3 2 4 1; 2 3 1 4 ; 1 4 3 2]

P_2 = [2 4 3 1 ; 3 2 1 4 ; 4 1 2 3 ; 1 3 4 2]

Cipher:

1 4 4 2 3 2 4 3 4 1 1 3 1 2 2 1 4

4 2 3 2 4 3 4 1 1 3 1 2 2

(b) **Binary Operation** – same P as in Case1 (a)

2nd part of the KEY (k_2): [1 1 1 1 2 2 2 2 1]

Message: message:[1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1]

Cipher:

4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4
4 4 4 4 4 4 4 4 4 4 4 4

(c) **Binary Operation –** same P as in Case1 (a)

2nd part of the KEY (k_2): [3 4 3 4 4 4 3 3 1]

Message:[1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1]

Cipher:

3 2 3 3 2 3 3 2 3 3 2 3 3 2 3 3 2 3
3 2 3 3 2 3 3 2 3 3 2 3

(d) **Binary Operation –** same P as in Case1 (a)

2nd part of the KEY (k_2): [3 4 3 4 4 4 3 3 1]

Message: [3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3]

Cipher:

1 3 2 1 2 3 1 4 4 1 3 2 1 2 3 1 4 4
1 3 2 1 2 3 1 4 4 1

From these experimental results it is inferred that there exists k_2 which can make the crypts almost flatten (statistically uniform frequency distribution) even if the message is a sequence of a single element of the set. It also shows that there exists k_2 which returns the output of a sequence of single element only (which is undesirable). These features are depicted in (a) & (b) respectively. In (a) the frequency of occurrence of the element '1' i.e. freq(1) = 8, freq(2) = 8, freq(3) = 6, freq(4) = 8 and in (b) freq(1) = freq(2) = freq(3) = 0 , freq(4) = 30.

It has been also inferred from observations of different experimental results that crypts of some messages are almost flat whereas for some messages crypts are not flat under the same key. In this case (c) & (d) support the fact. In (c) freq(1) = 0, freq(2) = 10, freq(3) = 20 & freq(4) = 0 and in (d) freq(1) = 10 , freq(2) = 6 , freq(3) = 6 , freq(4) = 6.

Case2

(a) **Binary Operation –** same P as in Case1 (a)

2nd part of the KEY (k_2): [1 1 1 1 2 2 2 2 1]

Message: [1 2 3 2 2 1 1 1 3 3 2 2 3 4 4 4 3 2 1 2 3 4 3 2 1 2 3 4]

Cipher:

4 2 2 2 2 4 2 1 4 1 2 4 3 1 1 4 1 2
2 2 1 2 3 3 4 3 3 3

(b) **Binary Operation –** same P as in Case1 (a)

2nd part of the KEY (k_2): [1 1 1 1 3 3 3 3 1]

Message: [1 2 3 2 2 1 1 1 3 3 2 2 3 4 4 4 3 2 1 2 3 4 3 2 1 2 3 4]

Cipher:

4 2 2 2 2 4 2 1 4 1 2 4 3 1 1 4 1 2
2 2 1 2 3 3 4 3 3 3

It has been inferred from the experiments that there are equivalent keys in the set of keys for this scheme. In this case (a) & (b) shows that crypts of any message under two different keys are same.
Sometimes it may happen that crypts of some particular message under two different keys are same. The presence of equivalent keys reduce the key space and hence the key complexity.

Case 3

Binary Operation – same P as in Case1 (a)

2nd part of the KEY (k_2): [1 4 3 2 2 2 2 2 1]

Message: [2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2]

Cipher:

2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
2 2 2 2 2 2 2 2 2 2 2

It shows that there exist keys which returns the message even by using sophisticated operations on a 3-quasigroup. This is a cryptographic weakness. These are weak keys which should not be considered as a part of the key space and should not be used for encryption.

7. Conclusions

In this paper we have proposed a modified 3-quasigroup based stream cipher. The scheme is designed to increase the key complexity exponentially so that it may be used for present day practical applications. The first part of the key randomly generates the initial quasigroup. It selects different parameters for generation of the seed 3-

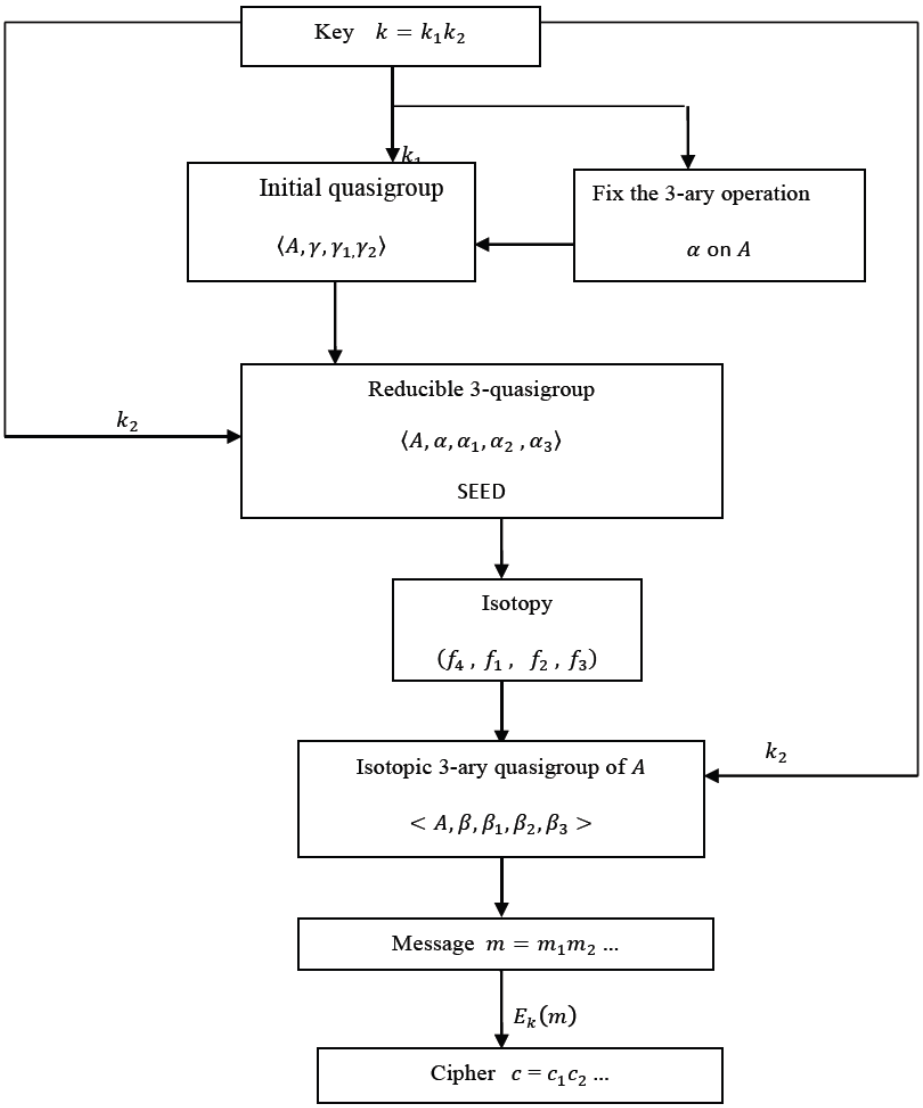
quasigroup and in a way acts as a different algorithm for encryption. Since the choice of 3-quasigroups is large even for small order and it also increases exponentially, we include this structure and use it to customize the algorithm based on the key used for encryption. In this paper, we have considered only reducible 3-quasigroups. Extension of this scheme for selecting suitable randomly generated 3-quasigroup based on the key is in progress. This would help to improve the cryptographic strength of our scheme. We have carried out large number of experiments on different order quasigroups and specifically on 4 order reducible 3-quasigroups for ease of visualization. Theoretical research based on observed results is also in progress which would help us to derive different suitable cases for cryptographic applications. Our future work in this direction also includes automatic property testing of large quasigroups generated using this process and thorough security analysis of this scheme.

References

1. Bruck R. H.: Simple Quasigroup, Bull. Amer. Math. Soc., 50, pp.769-781, (1944).
2. Koscielny C.: A Method of Constructing Quasigroup-based Stream Ciphers, Int. Journal Applied Math and Computational Sciences, Vol. 6, No. 1, 1996, pp. 109-121, (1996).
3. Koscielny C.: Generating quasigroups for cryptographic applications , Int. Journal Applied Math and Computational Sciences, Vol 12, No. 4,pp 559-569, (2002).
4. Markovski S., Gligoroski D. & Andova S.: Using Quasigroups for One-one Secure Encoding, Proc. VIII Conf. Logic and Computer Science (LIRA), Novisad, (1997).
5. Markovski S., Dimitrova V. and Mileva A.: A new method for computing the number of n-quasigroups, Buletinul Academiei De Stiinte A Republicii Moldova, Matematica, Vol 52, No 3, pp. 57-64 , (2006).
6. Markovski S.: Quasigroup String Processing and Applications in Cryptography, In 1st conference of Mathematics and Informatics for Industry, pp. 278-290, Thessaloniki, (2003).
7. Menezes A.,V Oorschot P., and Vanstone S.: Handbook of Applied Cryptography, CRC Press,(1997).
8. Petrescu A.: Applications of Quasigroups in Cryptography, Proc. Inter-Eng 2007, Univ. Petru Maior of Tg. Mures, Romania, (2007).
9. Petrescu A.: n-quasigroup Cryptographic Primitives: Stream Ciphers, Studia Univ. Babes Bolyai, Informatica, Vol. LV, No 2, pp.27-34, (2010).
10. Pal S.K., Jaiswal A. & Chamoli V.: Quasigroup based Design of New Cryptographic Schemes, Aryabhatt Journal of Mathematics & Informatics, Vol.3, No.2, pp. 277-294, (2011).

11. Stinson D. R.: Cryptography: Theory and Practice, Chapman & Hall / CRC, 3rd Edition, (2006).
12. Stallings W.: Cryptography and Network Security, Fifth Edition, Pearson,(2011).

Appendix



MMCacheSim: A Highly Configurable Matrix Multiplication Cache Simulator

Blagoj Atanasovski, Sasko Ristov, Marjan Gusev, and Nenad Anchev

Ss. Cyril and Methodious University, Faculty of Information Sciences and Computer Engineering,
Rugjer Boshkovikj 16, 1000 Skoipje, Macedonia
blagoj.atanasovski@gmail.com, sashko.ristov@finki.ukim.mk,
marjan.gushev@finki.ukim.mk, nenad_ancev@hotmail.com

Abstract. Memory access is the bottleneck of all computations. CPU cache is introduced to speed up accessing reused and local data. Matrix multiplication is the most common representative of many linear algebra algorithms which performance directly depends of the cache. Many cache parameters exist and impact the overall computing performance such as cache type, line, size, level, associativity, and replacement policy. Therefore an optimal architecture to execute certain compute and memory intensive algorithm is desirable in most applications.
We have developed MMCacheSim simulator to predict matrix multiplication performance on particular existing or non-existing multiprocessor. MMCacheSim simulates the execution time and number of cache misses that matrix multiplication algorithm performs with particular matrix size and element size executing on processor with different cache size, line, level associativity, and replacement policy.

Keywords: CPU Cache, Multiprocessor, HPC, Simulation

1 Introduction

Basic computer system is built on the Von Neumann concept (or Eckert-Mauchly as recently recognized) with CPU, main memory and bus. The problem of matching the speed of the instruction execution with the speed of fetching and storing the data / instruction degrades the overall performance of the computer system. Modern multiprocessors use multilayer cache memory system [9] to balance the gap between CPU and main memory and to speedup data access. The *cache size* is one of the most important cache parameters since larger caches reduce miss rates, but unfortunately, require greater data access times.

The matrix multiplication algorithm provides similar performance (speed) in the same cache region [14]. n-way set associative caches produce huge performance drawbacks for cache intensive algorithms regardless of cache size [7]. *Cache line* speeds up the time locality, i.e. if sometimes a particular memory location is referenced, then it is likely that near or even the same location will

S. Markovski, M. Gusev (Editors): ICT Innovations 2012, Web Proceedings, ISSN 1857-7288

be referenced again in the near future. The decision which cache line to be replaced if all the cache lines are fulfilled in the particular set depends on *cache replacement policy*.

All these cache parameters impact the algorithm overall performance and it is difficult to select the cache with optimal parameters for particular algorithm. Even more, the same algorithm behaves differently for different input size data. Applications provide better performance when they are executed on flexible cache with reconfigurability [17]. Using a proper simulators to predict the algorithm performance can save time and wasted money for unnecessary hardware. They can be used to measure the performance of new proposed schemes [1]. The authors in [2] propose techniques to predict the performance impact using hybrid analytical models. The authors in [19] propose a technique to overcome inter-thread cache conflict misses on shared cache and develop a highly configurable multi-core cache contention MCCCSim simulator that reproduces parallel instruction execution. A predictive model is proposed in [18] to allow fast and accurate estimation of system performance degradation also due to shared cache contention in parallel execution. The authors in [5] propose a statistical cache model Statstack that models a fully associative cache with LRU replacement policy and compared the results with the traditional cache simulator.

In this paper we present a trace driven simulation based MMCacheSim simulator that takes a list of memory addresses that represent the calls to main memory and tracks the changes in the cache. An overview of several existing cache simulators is presented in Section 2. The rest of the paper is organized as follows: In Section 3 we describe the MMCAcheSim architecture and design. Section 4 describes the real and simulated experiment environments and Section 5 presents the results of the simulation and experiments. Section 6 is devoted to conclusion and future work.

2 Related Work

This section presents different purpose cache simulators that we found in the literature. Dinero IV is the cache simulator that simulates a memory hierarchy with various caches [4]. A DEW strategy [8] speeds up the simulation of multiple combinations of cache parameters. It simulates only FIFO replacement policy. The authors in [6] define a fully parameterizable models applicable to n-way associative caches, but only for LRU replacement policy. Our MMCacheSim simulates both FIFO and LRU cache replacement policies for all cache levels.

The authors in [10] propose a CMP$im simulator based on the Pin binary instrumentation tool. It is a better simulator offering multi core support and data gathering for all levels of the cache. However, the capturing the results is more complex than our MMCacheSim. HC-Sim is also based on Pin that generate traces during runtime and simulates multiple cache configurations in one run [2]. An on-line cache simulation using a retargetable application specific instruction set simulator is provided in [13]. CMPSched$im evaluates the interaction of operating system and chip multiprocessor architectures [11].

Simulators can be also used in the teaching process. Hardware courses in software oriented curriculum require a lot of effort, both from instructors and students [16]. The authors in [15] using visual simulators achieved significant improvements in grade distribution and computer science student interest in hardware. Visual EduMIPS64 helps teachers to better present the specific topics of computer architecture and also help students to learn easier [12].

In this paper we present our MMCacheSim simulator and analyze if a successful prediction of cache performance can be achieved by simulating the execution of an algorithm and measuring the number of misses on different levels of CPU cache. We build a model that can be easily configured to represent different types of cache architectures with different replacement policies. A series of experiments were performed for execution of dense matrix multiplication algorithm varying matrix sizes on real world implementations and simulation with same parameters for the CPU cache architecture.

3 MMCacheSim Simulator

This section presents the MMCacheSimulator architecture, design and class diagram, and briefly describes its inputs and outputs. The MMCacheSim simulator is implemented as a set of Java classes, each for different CPU cache parameter:

- *Cache Line* - Represents a single cache line. It is initialized with the size of the cache line, the size of the elements saved inside it, and the address of the first element saved inside. Contains methods for writing new elements in the cache line and checking if an element is in the cache line;
- *Cache Set* - Represents a collection of cache lines available for both LRU and FIFO implementations as cache replacement policies. It is initialized with the associativity and line size. Contains methods for writing an address inside the cache and with it replacing the obsolete one according to the chosen replacing policy, checking whether an address is inside the given set;
- *L1, L2, L3 Cache Levels* - The actual cache memory, also available as LRU and FIFO implementations initialized with the size, associativity and the cache line size. Contains the cache sets, the data about misses and hits made on that particular level and methods for reading from and writing to the level;
- *Processor Core* - As a real processor core would have access to the cache. Several cores may share same cache structures. The simulated model of a core is initialized with instances of cache levels, by giving different cores the same instance of a cache level we simulate sharing. A cache core has only method to read a data element. If the element is not found in the cache levels a cache miss is recorded;

Figure 1 depicts the class diagram of the LRUCore class. The *CPU Core* is the class that contains the three Cache Levels. It contains the fields to measure the number of cache misses and hits on each level for the memory calls that go through particular core. The *Cache Levels* classes also contain fields about the

number of hits and misses they generated. Since a cache can be shared among several CPU cores (mostly L3 cache), *Cache Level* also posses the information about cache misses and hits per particular CPU core.

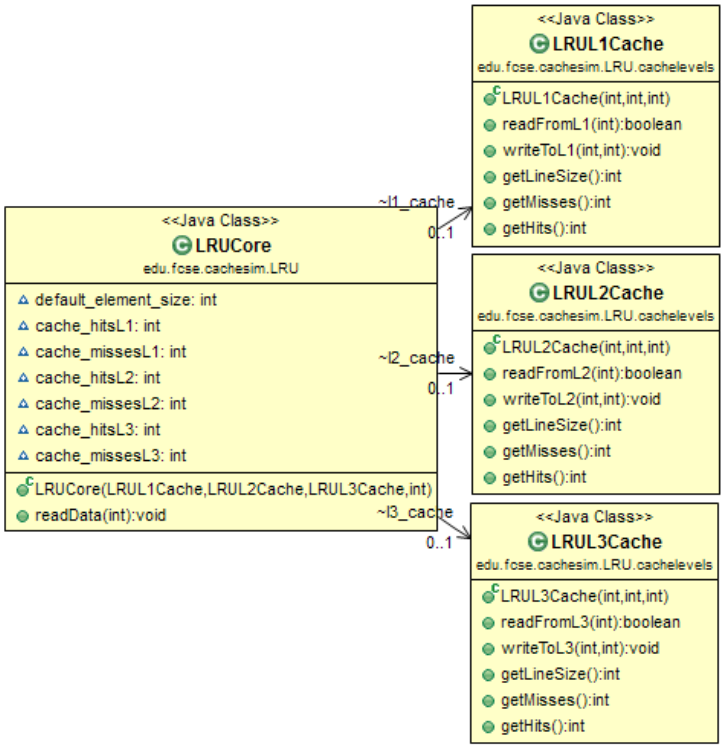


Fig. 1. MMCacheSim class diagram of a LRU Core

readData(memoryAddress) is the only method that is used during the simulation which goes through the *Cache Levels* searching if the required memory address is present in some of them going from the lowest to the highest. A sample code for inclusive caches is given in Appendix. This method also contains the logic that specifies the cache inclusivity. The *readFromLx* methods (where x denotes the cache level) in the *Cache Level* classes return a Boolean indicating whether the element is already stored in that *Cache Level*.

Figure 2 depicts the class diagram of the particular level LRUCache class. The other two classes *CacheSetLRU* and *CacheLine* contains the necessary information about particular cache level associativity.

The MMCacheSim simulates execution of the simple dense matrix multiplication algorithm. The simulation does not take into account the time required for arithmetic operations and memory writes because we are looking for the

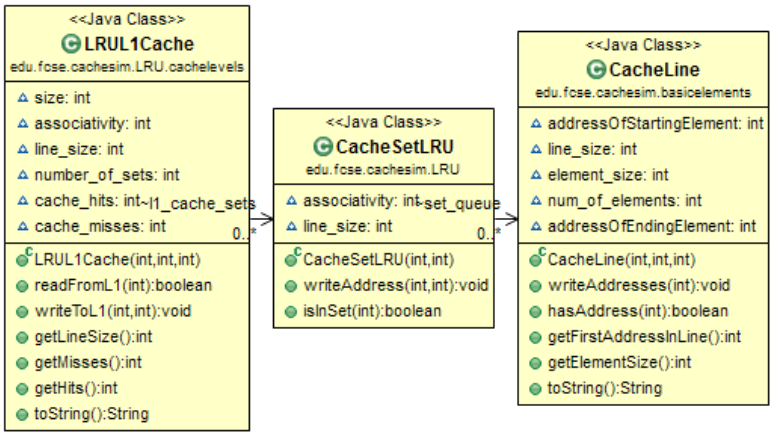


Fig. 2. MMCacheSim class diagram of the particular level LRUCache Class

effect that the cache produces when the same data is accessed multiple times and the speedup that can be gained when parallelizing the execution. The input for MMCacheSim is number of cores, cache levels, shared / dedicated cache per core, cache line, cache size, and cache replacement policy for each cache as input parameters. It returns the average clock cycles for cache hit per each cache level and cache miss for last level cache. It also measures the total clock cycles.

4 Experiment Environment

The experiments are performed on the real multiprocessors with totally different cache architectures. The first multiprocessor consists of 2 chips Intel(tm) Xeon(tm) CPU X5680 @ 3.33GHz and 24GB RAM. Each chip has 6 cores, each with 32 KB 8-way set associative L1 data cache dedicated per core and 256 KB 8-way set associative L2 cache dedicated per core. All 6 cores share 12 MB 16-way set associative L3 cache. The second server has one chip AMD Phenom(tm) 9950 Quad-Core Processor @ 2.6 GHz and 8 GB RAM. The multiprocessor has 4 cores, each with 64 KB 2-way set associative L1 data cache dedicated per core, and 512 KB 16-way set associative L2 cache dedicated per core. All 4 cores share 2 MB 32-way set associative L3 cache.

5 The Results of the Experiments

The first performed test is to determine the number of CPU cycles needed to access different levels of the cache in the simulated architectures. The same experimental tests are executed on both servers.

Figure 3 depicts the comparison of the simulation of matrix multiplication on a cache with FIFO replacement policy and cache parameters as Intel CPU.

The vertical axis represents the average number of memory accesses MA to each element of a matrix calculated as defined in (1). The values for total memory access cycles from the simulator are calculated as defined in [9]. The results prove the accuracy even for performance drawbacks due to cache associativity.

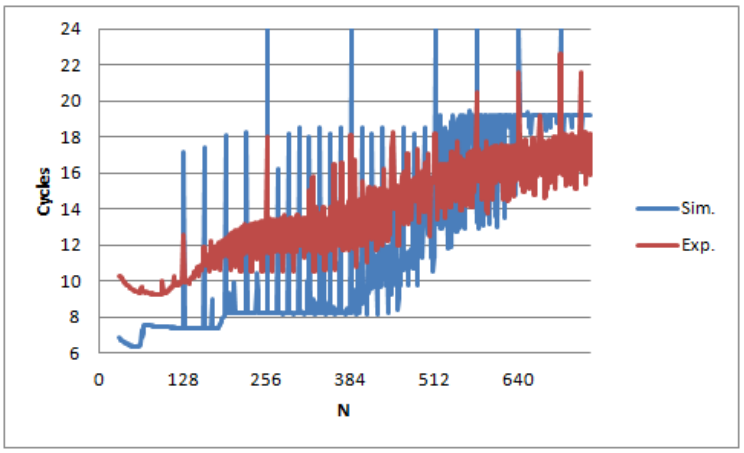


Fig. 3. Comparison CPU cycles for memory access for MMCacheSim simulation and sequential execution on Xeon server with FIFO replacement policy

$$MA = \frac{TotalMemoryAccessCycles}{N^3} \tag{1}$$

We run the simulation with LRU replacement policy for Phenom(tm) as it is AMD CPU and Opteron has LRU / PLRU replacement policy [3]. Figure 4 depicts the comparison of the number of CPU cycles used for memory access for Phenom server, LRU replacement policy, sequential execution. Because this simulation did not fit the experimental results we made two other simulations changing the replacement policy, since it was the only variable in the process. The simulation does not match neither for FIFO replacement policy as depicted in Figure 5.

The final experiment was to simulate with a new replacement policy Bit-Pseudo-LRU. Each cache line is associated with a MRU bit (most recently used) in this cache replacement policy. When the line is read the MRU bit is set to 1. When all lines in a cache set have their MRU bits set to 1, they are reset to 0. If some cache line should be replaced then the cache line in a set with the largest index that has a MRU bit 0 is replaced. Figure 6 depicts that the simulation is much closer to the experimental values. The simulation is still stepping away as the sizes of the matrices exceed the size of the cache memory. A possible explanation to the differences are: the authors in [3] show that the L3 cache

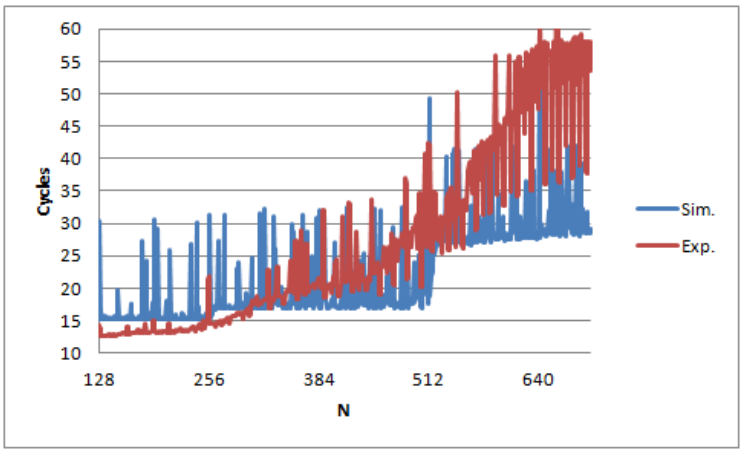


Fig. 4. Comparison of CPU cycles used for memory access for sequential execution on Phenom CPU and MMCacheSim simulation with LRU cache replacement policy

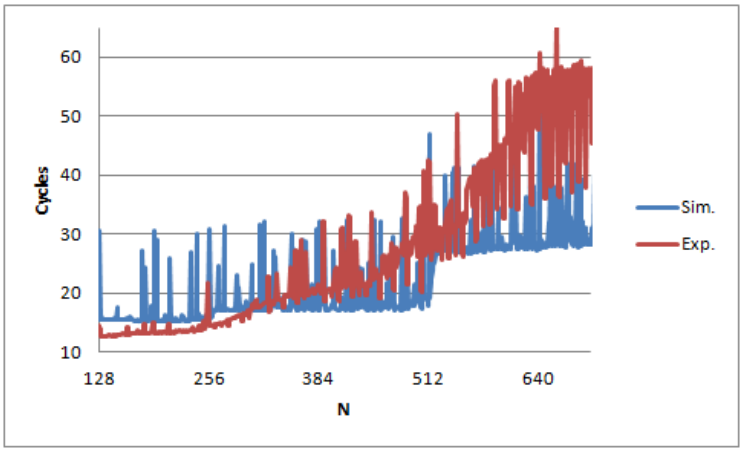


Fig. 5. Comparison of CPU cycles used for memory access for sequential execution on Phenom CPU and MMCacheSim simulation with FIFO replacement policy

at the Opteron processors uses some kind of pseudo-LRU cache replacement policy. The way of choosing the line inside the set seems to be different than the proposed Bit-PLRU policy. This is a logical explanation of the differences between simulated and experimental results, with the assumption that the cache replacement policy described in [3] is used in Phenom processors too. However this shows the ability to use the simulator not just to find favorable configurations for a certain algorithm but to research the configuration of a computer system when data for it are not available.

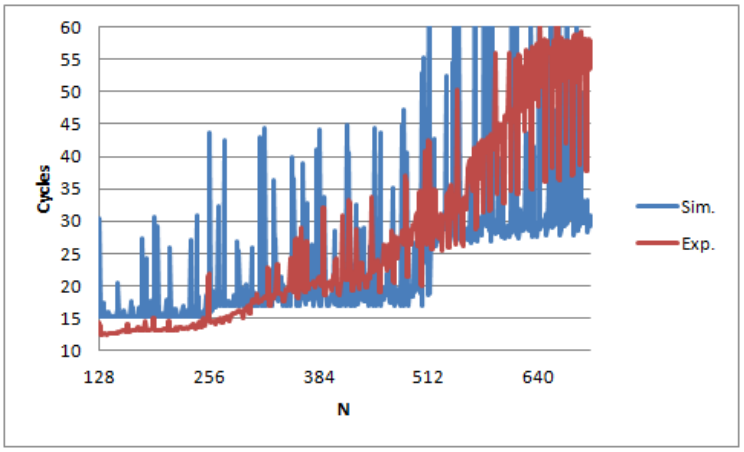


Fig. 6. Comparison of CPU cycles used for memory access for sequential execution on Phenom CPU and simulate with Bit-PLRU replacement policy

6 Conclusion and Future Work

Our MMCacheSim simulator simulates both FIFO and LRU cache replacement policies and the authors are working on implementation of other policies. All levels of cache hierarchy can be simulated. It is platform independent since the cache parameters are input parameters in the simulator.

This paper presents the simple implementation of our MMCacheSim simulator with its' features to simulate not only Matrix Multiplication algorithm execution, but any algorithm by giving a trace of memory accesses. MMCacheSim allows to change:

- The hierarchy between cache levels, to be shared between cores or dedicated;
- The inclusivity between different cache levels;
- The size of the cache memory, the associativity, cache line sizes
- Replacement policy, with ability to have different cache replacement policies per different cache levels.

The main contribution of MMCacheSim is to determine the most appropriate CPU cache architecture to achieve the best performance.

We will continue to improve MMCacheSim in order to decrease the time required for the results of simulation. Also we plan to implement additional utility classes to automate the process of building the required configurations to make the computer architecture teaching and learning process most appropriate.

References

1. An, B.S., Yum, K.H., Kim, E.J.: Scalable and efficient bounds checking for large-scale cmp environments. In: Proc. of the Int. Conf. on Par. Arch. and Compilation Techniq. pp. 193–194. PACT '11, IEEE Comp. Soc. (2011)
2. Chen, Y.T., Cong, J., Reinman, G.: Hc-sim: a fast and exact l1 cache simulator with scratchpad memory co-simulation support. In: Proc. of the 7-th IEEE/ACM/IFIP Int. conf. on HW/SW codesign and system synthesis. pp. 295–304. CODES+ISSS '11, ACM, New York, NY, USA (2011)
3. Conway, P., Kalyanasundharam, N., Donley, G., Lepak, K., Hughes, B.: Cache hierarchy and memory subsystem of the amd opteron processor. IEEE Micro 30(2), 16–29 (Mar 2010)
4. Edler, J., Hill, M.D.: Dinero iv trace-driven uniprocessor cache simulator (2012), `http://pages.cs.wisc.edu/~markhill/DineroIV/`
5. Eklov, D., Hagersten, E.: Statstack: Efficient modeling of lru caches. In: Performance Analysis of Systems Software (ISPASS), 2010 IEEE International Symposium on. pp. 55 –65 (march 2010)
6. Fraguela, B.B., Doallo, R., Zapata, E.L.: Automatic analytical modeling for the estimation of cache misses. In: Proc. of the Int. Conf. on Par. Arch. and Compilation Techniq. pp. 221–. PACT '99, IEEE Comp. Society (1999)
7. Gusev, M., Ristov, S.: Performance gains and drawbacks using set associative cache. Journal of Next Generation Information Technology (JNIT) 3(3), 87–98 (31 Aug 2012)
8. Haque, M.S., Peddersen, J., Janapsatya, A., Parameswaran, S.: Dew: a fast level 1 cache simulation approach for embedded processors with fifo replacement policy. In: Proc. of the Conf. on Design, Automation and Test in Europe. pp. 496–501. DATE '10 (2010)
9. Hennessy, J.L., Patterson, D.A.: Computer Architecture, Fifth Edition: A Quantitative Approach (2012)
10. Jaleel, A., Cohn, R.S., Luk, C.K., Jacob, B.: Cmpsim: A pin-based on-the-fly multi-core cache simulator. In: The Fourth Annual Workshop on Modeling, Benchmarking and Simulation (MoBS), co-located with ISCA'2008 (2008)
11. Moses, J., Aisopos, K., Jaleel, A., Iyer, R., Illikkal, R., Newell, D., Makineni, S.: Cmpsched$im: Evaluating os/cmp interaction on shared cache management. In: Performance Analysis of Systems and Software, 2009. ISPASS 2009. IEEE International Symposium on. pp. 113 –122 (april 2009)
12. Patti, D., Spadaccini, A., Palesi, M., Fazzino, F., Catania, V.: Supporting undergraduate computer architecture students using a visual mips64 cpu simulator. Education, IEEE Transactions on 55(3), 406 –411 (aug 2012)

13. Ravindran, R., Moona, R.: Retargetable cache simulation using high level processor models. Aust. Comput. Sci. Commun. 23(4), 114–121 (Jan 2001)
14. Ristov, S., Gusev, M.: Superlinear speedup for matrix multiplication. In: Information Technology Interfaces, Proceedings of the ITI 2012 34th International Conference on. pp. 499–504 (2012)
15. Ristov, S., Stolikj, M., Ackovska, N.: Awakening curiosity - hardware education for computer science students. In: MIPRO, 2011 Proc. of the 34th Int. Convention, IEEE Conference Publications. pp. 1275 –1280 (may 2011)
16. Stolikj, M., Ristov, S., Ackovska, N.: Challenging students software skills to learn hardware based courses. In: Information Technology Interfaces (ITI), Proceedings of the ITI 2011 33rd Int. Conf. on. pp. 339 –344 (june 2011)
17. Tao, J., Kunze, M., Nowak, F., Buchty, R., Karl, W.: Performance advantage of reconfigurable cache design on multicore processor systems. International Journal of Parallel Programming 36(3), 347–360 (Jun 2008)
18. Xu, C., Chen, X., Dick, R.P., Mao, Z.M.: Cache contention and application performance prediction for multi-core systems. In: ISPASS'10. pp. 76–86 (2010)
19. Zwick, M., Durkovic, M., Obermeier, F., Bamberger, W., Diepold, K.: Mcccsim - a highly configurable multi core cache contention simulator. Tech. rep., Lehrstuhl fr Datenverarbeitung, TU Mnchen (2009)

Appendix: Sample Code for *readData(memoryAddress)*

```
if (l1_cache.readFromL1(addressInMemory)) {
        cache_hitsL1++;
}
else {
        cache_missesL1++;
        if (l2_cache.readFromL2(addressInMemory)) {
                cache_hitsL2++;
        }
        else {
                cache_missesL2++;
                if (l3_cache.readFromL3(addressInMemory)) {
                        cache_hitsL3++;
                }
                else {
                        cache_missesL3++;
                        L3_cache.writeToL3(addressInMemory,
                                element_size);
                }
                l2_cache.writeToL2(addressInMemory,
                        element_size);
        }
        l1_cache.writeToL1(addressInMemory, element_size);
}
```

Improving the Traditional Testing Methods in Learning Foreign Languages

Veno Pachovski[1], Slobodanka Dimova[2], Marjana Vaneva[1]

[1] University American College Skopje, III Makedonska brigada br. 60,
1000 Skopje, Republic of Macedonia
{pacovski, vaneva}@uacs.edu.mk

[2] East Carolina University, USA/University of Copenhagen, Denmark
dimovas@ecu.edu

Abstract. A model for gathering oral answers as part of testing the speaker skills (i.e. command of language, native or foreign) is presented, as well as the software used in the experimentation.
The research presented here is a result of more than six (6) months' work with TESOL experts, based on 60 test subjects who gave 240 answers.

Keywords: Natural language processing, audio signals, voice analysis.

1 Introduction

Reliability and validity are critical language test qualities used to minimize measurement error and construct-irrelevant variance (Messick, 1991). However, these two qualities are not sufficient to ensure successful test application and usefulness. Practicality is another language test quality which may is sometimes overlooked even though it contributes to the test effectiveness and usefulness. While the other test qualities are associated with test score uses, practicality is a quality associated with the ways the test will be implemented and administered in a particular educational context (Bachman and Palmer, 1996). In particular, the test will be practical if it is designed to maximize the use of available resources, which may be material, human, or temporal.

Practicality becomes an issue when it comes to oral language testing and assessment. Administration of traditional, direct, interview-based oral tests can be unreliable and very time-consuming, so many teachers and institutions try to avoid including a speaking part in their language tests. For instance, the foreign language part of the Macedonian Matura Exam, like many other national and institutional

S. Markovski, M. Gusev (Editors): ICT Innovations 2012, Web Proceedings, ISSN 1857-7288

exams around the world, has only a reading and a writing section because the administration of an oral test is considered impractical.

Language testing companies and some educational institutions have developed semi-direct oral language tests which considerably reduce the oral test administration and rating time. For example, the oral section on the TOEFL iBT test is computer-administrated to allow for a more objective response elicitation and skill-integration (TOEFL). Purdue's Oral English Proficiency Test (OEPT) is another semi-direct, computer-based test that was developed to ensure a fair and reliable measurement of international teaching assistant oral proficiency (Ginther, Dimova, & Yang, 2010). The test format allowed for a dramatic reduction of administration time, i.e. for over 75%.

Despite the existence of computer-based oral tests, the computer testing format is not used widely in language classrooms and language teaching institutions because of software unavailability and because of existing software inapplicability for the particular testing/assessment needs of each educational context.

1.1 The Motivation and Problem Approach

In this paper, we will present the software developed for an oral language test administration at a Macedonian university. A four-item oral language test was designed to measure English oral proficiency among undergraduate and post-graduate students. The software development had four main goals: (1) administering the test to many students at the same time, (2) voice recording of test responses, (3) obtaining .wav files with test responses for subsequent rating, and (4) creating a test-response database.

We did not use existing sound recording software (e.g., Microsoft Office Sound Recorder or Garage Band on the Macintosh systems) because of several reasons. First, the use of the mentioned software would require training students how to use them. Second, a test administrator would have to assist each student with creation, naming, and storing the sound files. Third, the ready-made software could not limit the recording time to two minutes as required by the test response specifications.

2. Technical Description of the System in Relation to the Language Testing Procedure

As mentioned earlier, the main purpose of the software was to record student responses during an English oral language test. However, special attention needs to be paid to the interface design because the interface should allow students to navigate through the test easily following simple and straightforward instructions. At this time, the test item rubrics were given to students in a paper-based format while the recording was completed electronically. In the future, we plan to include the rubrics in the software interface. The following are the procedural test administration steps.

1. Students were given a booklet with information about the test, the recording procedures, and the test items. They were also given headsets with a microphone.
2. Students read the instructions to familiarize themselves with the ways they can navigate through the software interface.
3. Students read the item prompts, prepare their responses, and they click on the START button to record their response to the particular item. The program gives them only 120 seconds for the response. The system automatically cuts off the recording if it exceeds the allotted time. Students can stop the recording clicking on the STOP button if they finish the response time before the expiration of the 120 seconds.
4. The students repeat step 3 for all four questions.
5. The saved sound files were named using the student's ID numbers to ensure blind-rating. All four sound files and a text file with timing information were saved in a special folder for easy retrieval.

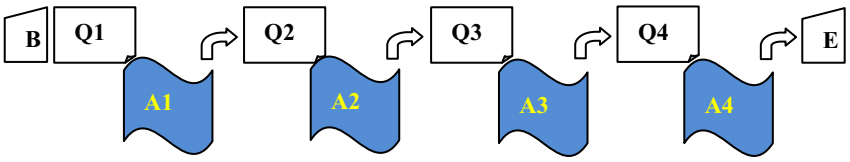


Fig. 1. The test activities – sequence of events

Participants - Participants were 60 undergraduate and post-graduate students at a Macedonian university (F=42; M=18). Twenty-eight of the students were low proficiency, and 32 2343 considered high proficiency. They participated in the study voluntarily.

2.2 The Interface

To avoid construct-irrelevant variance in measurement based on students' interaction with the interface, the interface must be simple and easy to navigate. In other words, students should be able to find the buttons and see the timing easily, which would minimize their anxiety level associated with the technology.

The first version of the interface was intended to test software's functionality meaning testing the background processes of the application (recording procedure, saving the recording, timer control, logging activities, etc.) . (Fig.2)

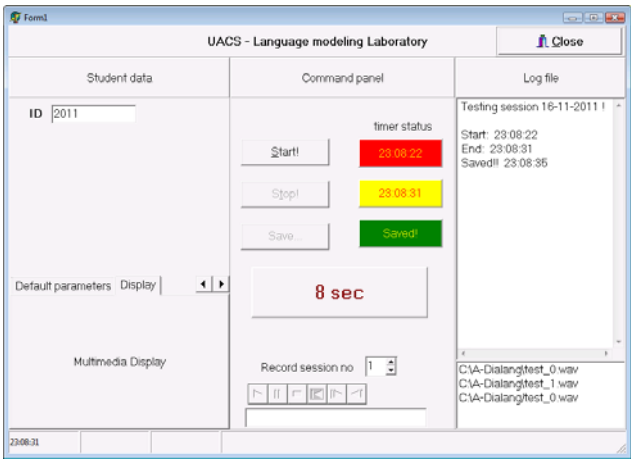


Fig. 2. The first version of the testing interface – everything is visible

After the functionality was ensured, it was decided that the interface should be remodeled according to requirements mentioned in the introductory part of this article. [list the previously mentioned requirements]

After the successful piloting of the new version of the interface, the following solution in Figure 3 was adopted. Figure 3a represents what the student initially sees on the screen, and Figure 3b represents what they see while answering the questions.

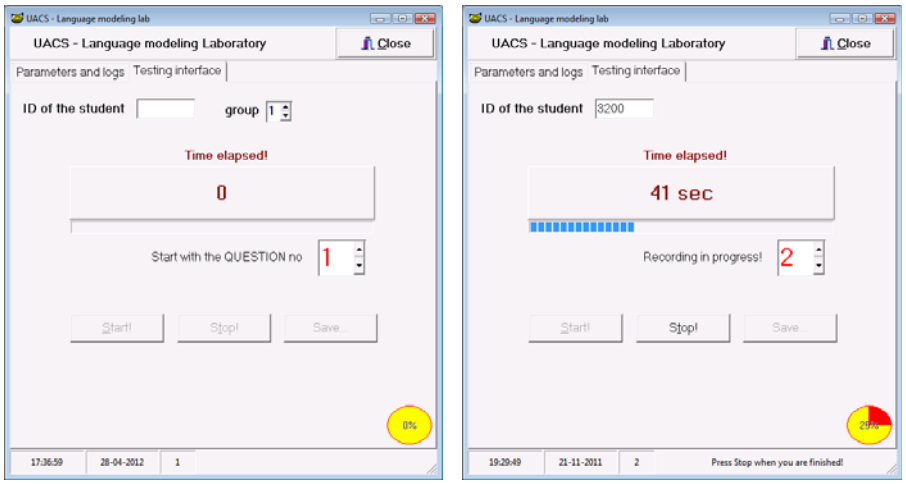


Fig. 3. (a) The initial state of the application (b) Answering question no. 2 – 41 sec elapsed

The system should also have some configuration parameters, considering that this was a pilot so there were parts of the software dedicated to that purpose as well. (Fig.4 – left picture)

Not knowing the full potentials of this kind of testing, the log file was created so that the quality of testing process can also be analyzed. (Fig.4 – right picture)

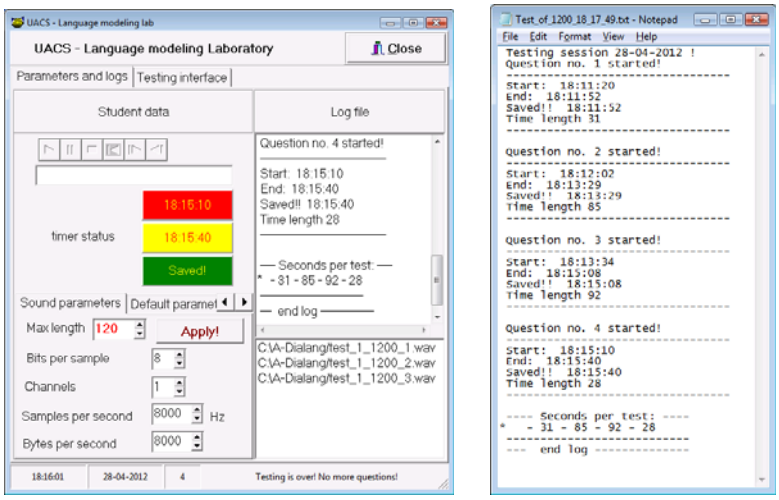


Fig. 4. Configuration interface and the log file– the user (testing subject) should not see this

After finishing the testing cycle (in this case, four answers), the candidate exits the program. As a result, four wave files are stored in a testing computer. (Fig. 5)

Names of the result files were designed so that the system can group them within a common folder, so that they could be retrieve for rating more easily.

Test_of_1200_18_17_49....	28.04.2012 18:17	Text Document	1 KB
test_1_1200_4.wav	28.04.2012 18:15	VLC media file (.w...	315 KB
test_1_1200_3.wav	28.04.2012 18:15	VLC media file (.w...	1.011 KB
test_1_1200_2.wav	28.04.2012 18:13	VLC media file (.w...	935 KB
test_1_1200_1.wav	28.04.2012 18:11	VLC media file (.w...	342 KB

Fig. 5. Stored answers - (testing subject should not see this either)

Considering that there were a lot of testing locations (the test was administered on different machines in groups of various sizes), a cloud approach was used to gather the resources for storage and analysis. This allowed for an easier access to the data by the team members.

2.3 Software Behind Commands

The work started using ready-made Delphi procedures. [5]. Afterwards, the interface was programmed in the usual manner using appropriate components for organization of content on the form (Panels, Page controls), entering text (Edit fields), measuring time (Timers), signaling (Panels), keeping the log of activities (Memo), controlling the input of question numbers and testing group (SpinEdit), activating processes (Buttons), following progress of the test (Gauge), etc.

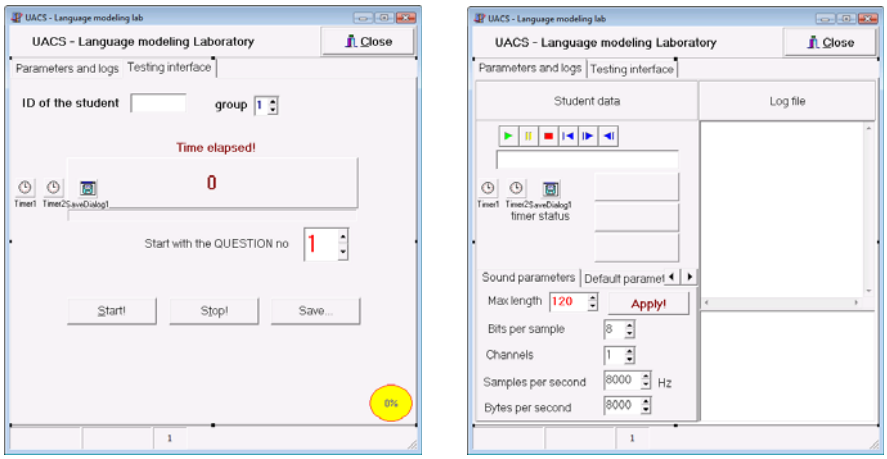


Fig. 6. The interface – designing phase

The initial recording parameters (8,1,8000, 8000) gave small sound files, but were found to require enhancement. So, later (before the main experiment started) on the request of the research supervisor, they were changed to (16,2,16000,16000). That resulted in much bigger files (4 to 5 times), easier to listen.

2.4 Testing Results

Although the recordings were, in general, of good quality, they can be grouped as normal, noisy, silent and (some of them were even) blank.

Some of it is the result of using common computer labs and standard issue head-phones with microphones, but also (the low quality of) some of them can be accredited to in-experience (and stress) of the students as well as simple lack of experience with this kind of testing.

The last probably accounts for frequent breaks or low voice.

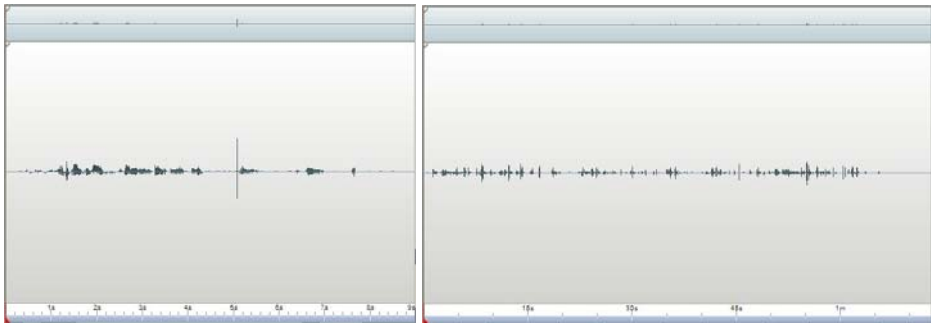


Fig. 7. Sound recordings of two answers (short one and a long one)

For listening and enhancing (where needed) the NCH Wavepad software was used, and it performed quite good. [7]. Figure 7 represent sound signatures of our recordings.

3 Conclusion and Future Work

The system had an excellent performance considering that it was used by many students in groups between 4 and 15 students, on various computers configurations or versions of Windows (XP, Vista, Win 7).

For the next series of tests, one of the improvements should be to represent the item rubrics on screen, so that the testing process can be completed without the use of booklets. In addition, this would allow for easier navigation from one item to another.

The screen should also present some type of feedback that the recording is on (some sinusoid wave) which will reassure the student and the test administrator that the speech is being recorded properly. This could also minimize the stress rather than draw unnecessary attention.

Recording of the answers, on the other hand, should be automated on a larger scale, meaning that the recording must be stored on server in a designated folder, or in a database.

Finally, the optimal protocol for communication between this software (a desktop application) and the web server (the cloud) should be devised, in order to enable following the progress of the testing process on the local machine from the beginning to end. In that way, the test administrator can follow candidate's progress and intervene if necessary.

References

1. Bachman, L. F., & Palmer, A. (1996). Language testing in practice. New York: Oxford University Press.
2. Dimova, S., Vaneva M., Pacovski V., Examining validity of explicit contextual clues in oral tasks. 9-th annual EALTA conference 2012 - Validity in Language Testing and Assessment, University of Innsbruck, Austria, May 31 - June 3, (2012)
3. Ginther, A., Dimova, S., and Yang, R. Conceptual and empirical relationships between temporal measures of fluency and oral English proficiency with implications for automated scoring. Language Testing Jul 01, 2010; 27: 379-399.
4. Messick, S. (1989). Validity. In R. L. Linn (Ed.), Educational measurement (pp. 13-103). New York: Macmillan.
5. SwissDelphiCenter – Developers Knowledge Base http://swissdelphicenter.com/en/showcode.php?id=132
6. Test of English to Speakers of Other Languages (TOEFL) http://www.ets.org/toefl/ibt/about/content/
7. NCH software - http://www.nch.com.au/wavepad/index.html

S. Markovski, M. Gusev (Editors): ICT Innovations 2012, Web Proceedings, ISSN 1857-7288

Using Highly Structured Document Collection for Simulating Multimedia Object Retrieval

Vesna Gega[1], Ilija Kumbaroski[2], Ivan Chorbev[3], Dancho Davchev[1]

[1]University for Information Science and Technology "St. Paul the Apostle" Ohrid
{vesna.gega, dancho.davchev}@uist.edu.mk
[2]European University Skopje
ilija.kumbaroski@gmail.com
[3]Faculty of Computer Science and Engineering "Ss. Cyril and Methodius" Skopje
ivan.chorbev@finki.ukim.mk

Abstract. This paper describes the use of a highly structured document collection in XML (IMDb) for simulating retrieval of non-overlapped XML multimedia objects. In our experiments conducted in Data Centric Ad Hoc Task, we inspect two types of dynamically formed units of retrieval, combining them with different models for searching and ranking. Our purpose is to find the combination that leads to most effective retrieval of XML multimedia objects from very complex documents. We concluded that retrieval is more effective using the Dirichlet Smoothing as Language Model rather than the Pivoted Cosine as Vector Space Model. Also, better scores are achieved using the textual content of the whole document because it supplies more terms and better identification of the multimedia object compared to the textual content contained in smaller document's parts as the root's first level descendants.

Keywords: XML, INEX, IMDb, Data Centric Track, Ad Hoc Task, relevant, retrieval

1 Introduction

Documents of interest in our research are logically structured documents with the standardized eXtensible Mark-up Language (XML)[1] as a wide accepted standard of World Wide Web Consortium (W3C)[2]. The structure of the document is represented like a tree, based on opened and closed tags. Each pair of an opened and closed tag is known as an XML element. If the XML element contains an attribute that refers to an external multimedia object, it is called XML multimedia element. The term multimedia object is usually connected to images, videos and audios. The textual content of the XML multimedia element in an XML structured document collection can be used to describe the multimedia object. Also, the textual content of its ancestors going up

[1]http://www.w3.org/XML/, Accessed April 2012
[2]http://www.w3.org/, Accessed April 2011

to the root, and its descendants going down to the leaves, at different levels of granularity, give wider description of that multimedia object. Therefore, if the relevance of the retrieved multimedia objects is based on the textual content, the retrieval of XML multimedia elements is the same as the retrieval of pure textual XML elements [12], [19], [23]. An interesting and challenging aspect of our research is to investigate how the complex logical structure of the documents influences the XML multimedia object retrieval. Regardless of the structure, we constructed the results using two strategies of grouping enclosing elements of the multimedia element from a single document, without overlapping. In order to discover the most effective method of retrieval from a strongly XML structured document collection, we experimented with two models for searching and ranking.

This paper is organized as follows. In Section 2, we present related work. In Section 3, we describe the document collection, and our approach for indexing and retrieval. In Section 4, we show the experiments we carried out, the results and the evaluation of our approach. The conclusion and future work are presented in Section 5.

2 Related Work

There are different aspects of multimedia object retrieval. The textual content from an XML structured document or its parts have the main role in determining the relevant multimedia objects. In [9], experiments were conducted on a small data set in order to investigate how element hierarchy influences multimedia object retrieval effectiveness. They showed that elements higher in the document's hierarchical structure are more effective in determining relevant documents containing the multimedia objects, whereas elements lower in the structure are better in identification of relevant multimedia objects. This research was continued in [10], where the logically structured regions at different levels of granularity were more effective in retrieval multimedia objects than the whole document as representation. Also, it was shown that the lower level regions lead to bigger precision, whereas higher level regions contain more terms and lead to improved recall. In [11] XPath[3] and XLink[4] were used to explore how the hierarchical and linking structural information combined with the content information works in the retrieval of multimedia objects in digital libraries. The paper [22] presents and compares the effectiveness of two context based methods for multimedia object retrieval. The first one is based on an implicit use of textual and structural context of multimedia objects, whereas the second one is based on an explicit use of both sources. They have experimented in order to show which sources of evidence: children nodes, brothers or ancestors offer the best multimedia object representation. In [7] a mixture of evidence from a content-oriented XML retrieval system is investigated. Content-based image retrieval (CBIR) system using a linear combination of evidence is presented. The authors showed that CBIR with text search leads to better retrieval performance. The paper [18] proposed a general framework for index-

[3]http://www.w3schools.com/xpath/default.asp, Accessed April 2011
[4]http://www.w3schools.com/xlink/xlink_intro.asp, Accessed April 2011

ing images with associated text. Their experiments showed that the both textual and visual information contribute for significantly better results, than using visual only or textual only information. The research in integrating text retrieval and image retrieval in XML Document Searching was implemented in [21]. They used different search engines, and the results generated by the two search engines were merged together and were evaluated as one result set. The [6] presents a model for retrieval of images from a large World Wide Web based collection. Rather than considering complex visual recognition algorithms, their model is based on combining evidence of the text content and hypertext structure of the Web. A lot of researches were conducted in [3] and [4] in direction to find the effective way for text retrieval from a big and highly XML structured document collection. But also, it is interesting to explore and find how the complex document structure affects multimedia object retrieval effectiveness. However, in our case XML elements were used for simulating multimedia object retrieval.

3 Document Collection, Indexing and Ranking

Our research was conducted in Ad Hoc Data Centric Task[5], where a ranked list of results is returned. The term result is related to an answer to a query, defined as a set of elements that are "collectively relevant" to the query. We used the newly created IMDb document collection, based on the www.imdb.com web site. The collection has 4418102 XML documents, very rich logical structure and contains the following objects: movies and persons (actors, directors, producers, and etc.) [3]. The XML document structure, shown in **Fig. 1**, is generated according to the appropriate complex DTD [3]. Each XML document refers exactly to one movie or to one person. There are 30 queries in the CO and CAS version, in the standard INEX format, shown in **Fig. 1**, but only 28 of them are assessed and used in the process of searching.

```
−<movie>
    <title>Exposição de 22 Anos de Tapeçaria (1970) </title>
  −<url>
      http://www.imdb.com/Title?Exposição de 22 Anos de Tapeçaria (1970)
    </url>
  −<overview>
      <rating>4.4 85votes </rating>
    −<genres>
        <genre>Documentary </genre>
      </genres>
    </overview>
    <cast> </cast>
  −<additional_details>
    −<countries>
        <country>Portugal </country>
      </countries>
    </additional_details>
    <fun_stuff> </fun_stuff>
</movie>
```

```
−<topic id="2010027" ct_no="24">
    <title>tom cruise movies</title>
    <castitle>//movie[about(., Tom Cruise)]</castitle>
  −<description>
      movies where tom cruise works as actor, producer or director
    </description>
  −<narrative>
      The user wants all movies where Tom Cruise have worked
    </narrative>
</topic>
```

Fig. 1. XML document (left), INEX topic (right)

[5] http://www.inex.otago.ac.nz/tracks/strong/strong.asp, Accessed April 2012

We have indexed only the movie object inspired by papers [16] and [17], because there the best results were achieved with movie object. Also, we have used only CO version of the queries, based on the same research where CO queries performed much better results than CAS queries. Before describing the indexing techniques used, we briefly describe the manner in which we have preprocessed the documents. This step was necessary for unique identification of each element and better navigation through the ordered list of returned relevant results, as shown in **Fig. 2**. We have determined the ordinal number of each element, the path from the root to that element using XPath and the position of its textual content expressed through the position of its initial and final character, according to the whole document. For this purpose, we have developed an intelligent algorithm which recursively loops trough the document tree.

```
1. 1400964:/movie[1]/overview[1]:98-485 (score 16.688488, docid
2227531)
2. 1036353:/movie[1]/overview[1]:236-2785 (score 15.784092, docid
202733)
3. 1270402:/movie[1]/cast[1]:218-927 (score 14.700630, docid
1500723)
4. 1249187:/movie[1]/cast[1]:451-748 (score 14.325662, docid
1382850)
5. 1582178:/movie[1]/fun_stuff[1]:5316-5707 (score 14.045104,
docid 3237973)
6. 650248:/movie[1]/cast[1]:286-1179 (score 13.748998, docid
5597009)
...
```

Fig. 2. A list of non-overlapped results obtained using the preprocessed documents

We have inspected two types of dynamically formed non-overlapping [2] units of retrieval that enclose, describe and identify the multimedia object, as shown in **Fig. 3**.

Fig. 3. Two types of indices (marked with dashes)

Considering the fact that the whole document is the most natural unit of retrieval [1] and it is an often researched approach, we have decided this to be one of the alternatives for our experiments. We have created an algorithm that groups all the elements from the document in a single element that will be returned as a response, taking in consideration the length and the structure of the document. In other words, the entire textual content of the document that surrounds the multimedia object located in the *movie* element can be viewed as a unit of retrieval. In this case only one element from a document was indexed. In our second approach, the elements were dynamically grouped up to the *movie's* first level descendants. This means that we have considered each of them as a unit of retrieval. In this case, usually there were from 4 to 6 indexed elements from each document. If some of the elements were empty, they weren't indexed.

We have combined our developed indices with the following methods for searching and ranking: Dirichlet Smoothing as Language Model [13], [14], [15], [24] and Pivoted Cosine as Vector Space Model [13], [14], [15], [20]. In the Vector space model, the query and the document are presented as two *n*-dimensional vectors, where *n* is a number of unique terms in the document. The main functionality of the Vector space model is to determine the similarity of these two vectors. The most famous and usually implemented Vector space technique is Cosine measure, where similarity is based on the cosine of the angle between the two vectors. When the angle is smaller, the cosine is bigger, so the two vectors are more similar. The Language model determines the probability of generating the query and the document from the same language model [24]. This is based on probability distributions, and Dirichlet smoothing is the commonly used technique.

4 Experiments and Results Analysis

The indexing, ranking and the retrieving were implemented using the advanced system for that purpose, Zettair[6]. In order to find the optimal parameters that lead to most effective retrieval of XML multimedia objects from very complex documents a training phase was performed. We used the two types of indices, and we applied Vector Space Model in combination with the Pivot Normalization (Pivoted Cosine) [13], [14], [15], [20], with different values for the slope parameter *s*. We also used the Language model (Dirichlet Smoothing) with different values for the smoothing parameter μ [13], [14], [15], [24]. We experimented in two directions: focused retrieval, where a list of ordered non-overlapped results in descending order is returned and relevant in context retrieval where a list of non-overlapped results is returned, but ordered according to the relevancy of the document where the results came from. The INEX MAiP (Mean Average interpolated Precision) and MAgP T2I (300) (Mean Average generalized Precision) measures were used to calculate effectiveness of focused retrieval and relevant in context retrieval, respectively [5], [8]. We used 14 randomly chosen queries. In the training process we discovered the values for the parameters

[6]http://www.seg.rmit.edu.au/zettair/, Accessed April 2012

that contributed for best result, **Table 1** and **Table 2**. So, in the focused retrieval using the Pivoted Cosine the high scores are performed with slope value of 0.2 for the *movie* index and slope value of 0.1 when indexing the *movie's* first level descendants. Also, using the Dirichlet Smoothing best results are achieved with the smoothing parameter value of 3000 for the *movie* index and smoothing parameter value of 2500 when indexing the *movie's* first level descendants, shown in **Fig. 4** and **Fig. 5**.

Table 1. The three best training results in the focused retrieval

Movie index			
μ	MAiP	*S*	MAiP
2500	0.244653	**0.2**	**0.240940**
3000	**0.245467**	0.3	0.196735
3500	0.241746	0.4	0.196735
Movie's first level descendants index			
2000	0.172256	0	0.120047
2500	**0.172764**	**0.1**	**0.162401**
3000	0.172278	0.2	0.125616

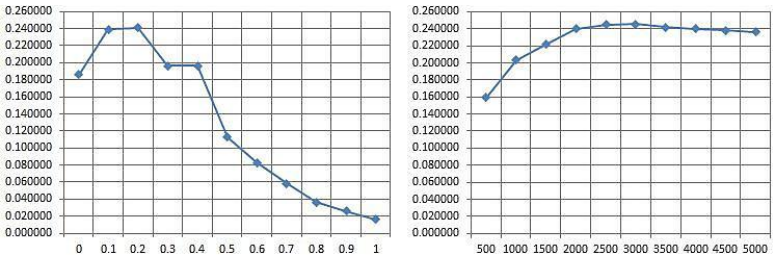


Fig. 4. MAiP, *movie* index

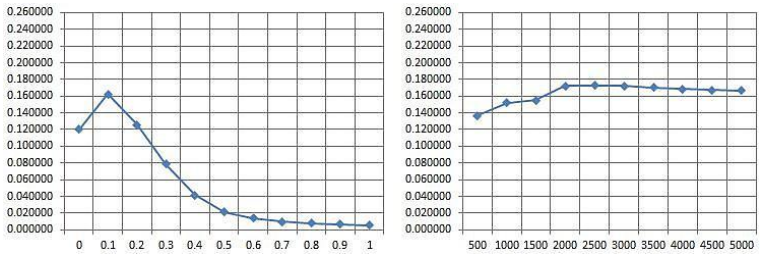


Fig. 5. MAiP, indexing *movie's* first level descendants

In the relevant in context retrieval using the Pivoted Cosine, the highest scores are performed with slope value of 0.2 for both indices. Also, using the Dirichlet Smoothing best results are achieved with the smoothing parameter value of 2000 for the *movie* index and smoothing parameter value of 500 when indexing the *movie's* first level descendants, shown in **Fig. 6** and **Fig. 7**.

Table 2. The three best training results in the relevant in context retrieval

Movie index			
μ	MAgP	*S*	MAgP
500	0.218528	0.1	0.152834
2000	**0.221105**	**0.2**	**0.172198**
2500	0.215092	0.3	0.146029
Movie's first level descendants index			
500	**0.215836**	0.1	0.145367
2000	0.214749	**0.2**	**0.169724**
2500	0.208271	0.3	0.142283

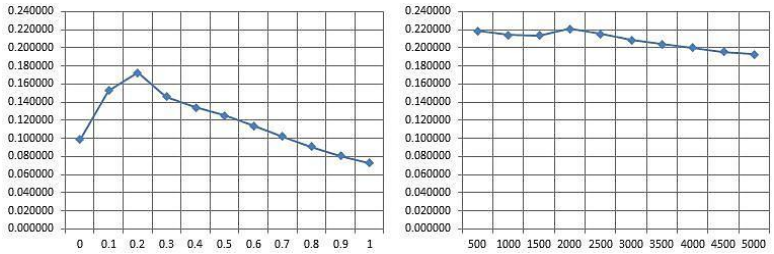


Fig. 6. MAgP, *movie* index

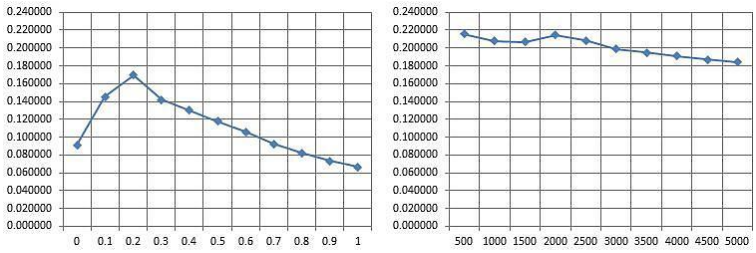


Fig. 7. MAgP, indexing *movie's* first level descendants

We can see in both cases of retrieval, when using the Pivoted Cosine, the effectiveness continuously decreases for slop values bigger than 0.2. In the focused retrieval using the Dirichlet Smoothing the effectiveness continuously decreases for smoothing parameter values bigger that 3000 using the *movie* index and bigger than 2500 indexing the *movie's* first level descendants. In the relevant in context retrieval using the Dirichlet Smoothing for both indices, the effectiveness continuously decreases for smoothing parameter values bigger than 2000. But in contrary to Pivoted Cosine, here the differences between values are not very drastic; they lie in a very small interval. Therefore, in the testing phase we applied the three optimal values for the slope and the smoothing parameters that contributed for best results in the training phase, shown in **Table 1** and **Table 2**. We made experiments with all 28 queries. In addition, we present the MAiP and MAgP values. In the focused retrieval:

- The MAiP values obtained using the *movie* index, for the slope parameter values **0.1**, 0.2 and 0.3 are **0.244884**, 0.238959, and 0.215998, respectively. Also, for the smoothing parameter values 2500, **3000**, and 3500 the MAiP values are 0.269152, **0.269813**, and 0.263837, respectively.
- The MAiP values obtained indexing the *movie's* first level descendants, for the slope parameter values 0.0, **0.1**, and 0.2 are 0.117725, **0.139662**, and 0.118115, respectively. Also, for the smoothing parameter values 2000, 2500, and **3000** the MAiP values are 0.158184, 0.157982, and 0.**158273**, respectively.

In the relevant in context retrieval:

- The MAgP values obtained using the *movie* index, for the slope parameter values **0.1**, 0.2 and 0.3 are **0.217642**, 0.210920, and 0.192393, respectively. Also, for the smoothing parameter values 500, **2000**, and 2500 the MAgP values are 0.26761, **0.277213**, and 0.269519, respectively.
- The MAgP values obtained indexing the *movie's* first level descendants, for the slope parameter values **0.1**, 0.2, and 0.3 are **0.212869**, 0.206800, and 0.188573, respectively. Also, for the smoothing parameter values **2000**, 2500, and 3000 the MAgP values are **0.262132**, 0.254090, and 0.257827, respectively.

These results lead to the followings: Through all the experiments, the Dirichlet Smoothing is more superior compared to the Pivoted Cosine. Also, the retrieval of the most relevant multimedia objects from highly structured collection is more effective using the *movie* index. That was expected, because the textual content of the whole document supplies more terms and better identification of the multimedia object than the smaller textual content contained in the *movie's* first level descendants.

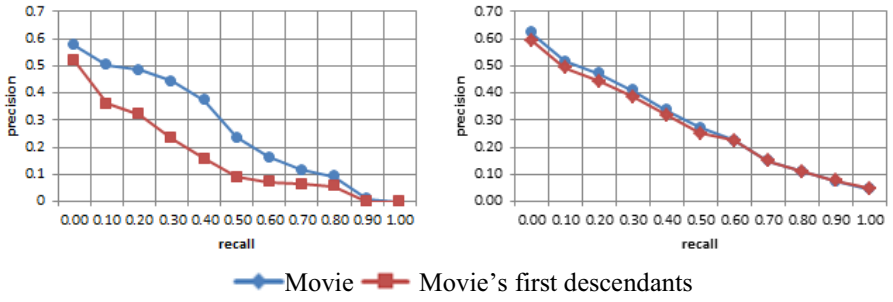


Fig. 8. Focused retrieval (left), Relevant in context retrieval (right)

From visual inspection in **Fig. 8** we can see that in the focused retrieval the *movie* index always performs high precision at every level of recall. On the other hand, in the relevant in context retrieval, the curves are closer to each other. It means that the *movie* index is just a little bit more effective than indexing *movie's* first level descendants because the second index fails to return all the relevant document parts. The situation where the two curves are much closer to each other (overlapped) is confirmed with the well-known definition for the recall. With higher recall values the

number of relevant retrieved *movie's* first level descendants is bigger. In this scenario, the curves would be fully overlapped when all the movie's first level descendants from the document are considered as relevant and they are returned as a response.

The official best ranked INEX 2010 runs are ufam2010Run2 (MAiP = 0.1965) and OTAGO-210-DC-BM25 (MAgP = 0.2491). Our best ranked results are higher than these and we suppose that it comes from the fact that we have indexed only the movie object of the IMDb collection.

5 Conclusion and Future Work

This paper described the use of strongly XML structured document collection (IMDb) for simulating retrieval of non-overlapped XML multimedia objects. We investigated how the context influences the multimedia object retrieval effectiveness. In order to discover the most effective way of retrieval, we experimented with two models for searching and ranking and two strategies of indexing and grouping different elements from a single document without overlapping. We showed that the Dirichlet Smoothing is more superior compared to the Pivoted Cosine. Also, as we expected, the *movie* index contributed for better results, because the textual content of the whole document supplies more terms and better identification of the multimedia object than the smaller textual content contained in the *movie's* first level descendants. The index for the *movie's* first level descendants always fails to return all the relevant document parts, therefore in both cases of retrieval its effectiveness curve is under the *movie* index curve.

We are planning to experiment with INEX 2011 Data Centric topics and make fusion with our present results. Also, we plan to extend this work with some non-English document collections, like a Macedonian document collection. A challenging approach is to make a mixture of context and content retrieval from documents with a very rich XML structure.

References

1. Betsi, S., Lalmas, M., Tombros, A., Tsikrika, T.: User Expectations from XML Element Retrieval. In: ACM SIGIR Conference on Research and Development in Information Retrieval (Poster), pp. 611-612. Seattle, USA (2006).
2. Clarke, C.: Controlling overlap in content-oriented XML retrieval. In: Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 314-321. New York, USA (2005).
3. Geva S., Kamps J., Schenkel R., Trotman A.: INEX 2010 Workshop Pre-proceedings. In: IR Publications, Amsterdam. Huize Bergen, Vught, the Netherlands (2010).
4. Geva S., Kamps J., Schenkel R.: INEX 2011 Workshop Pre-proceedings. In: IR Publications, Amsterdam. Hofgut Imsbach, Saarbrücken, Germany (2011).
5. G¨overt, N., Kazai, G., Fuhr, N., Lalmas, M.: Evaluating the effectiveness of content oriented XML retrieval. In: Information Retrieval, Vol. 9, Nr. 6, pp. 699-722. (2006).
6. Harmandas V., Sanderson M., and Dunlop M.: Image Retrieval by Hypertext Links. In: SIGIR ACM, pp. 296-303. (1997).

7. Iskandar A., Pehcevski J., Thom J., Tahaghoghi S.: Combining Image and Structured Text Retrieval. In: INEX, Vol. 3977 Springer, pp. 525-539. (2005).
8. Kamps, J., Lalmas, M., Pehcevski, J.: Evaluating Relevant in Context: Document Retrieval with a Twist. In: ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 749-750. Amsterdam, Netherlands (2007).
9. Kong. Z., Lalmas M.: XML Multimedia Retrieval. In: Symposium on String Processing and Information Retrieval (SPIRE 2005), pp. 218-223. Buenos Aires, Argentina (2005) (Short Paper).
10. Kong Z., Lalmas M.: Using XML Logical Structure to Retrieve (Multimedia) Objects. In: 11th European Conference on Research and Advanced Technology for Digital Libraries (ECDL 2007), pp. 100-111. Budapest, Hungary (2007).
11. Kong Z., Lalmas M.: Integrating XLink and XPath to Retrieve Structured Multimedia Documents in Digital Libraries. In: RIAO 2004 Conference on Coupling approaches, coupling media and coupling languages for information retrieval, pp. 571-581. University of Avignon (Vaucluse), France (2004).
12. Lalmas, M., Trotman, A.: XML Retrieval. In: Encyclopedia of Database Systems Springer, pp. 3616-3621. US (2009).
13. Pehcevski, J.: Evaluation of Effective XML Information Retrieval. PhD thesis RMIT University. Melbourne, Victoria, Australia (2006).
14. Pehcevski, J., Thom, J.: HiXEval: Highlighting XML retrieval evaluation. In: INEX, Vol. 3977 Springer, pp. 43-57. Dagstuhl Castle, Germany (2005).
15. Pehcevski, J., Thom, J.: Evaluating focused retrieval tasks. In: Proceedings of the SIGIR Workshop on Focused Retrieval. Amsterdam, the Netherlands (2007).
16. Ramírez G.: UPF at INEX 2010: Towards Query-Type Based Focused Retrieval. In: INEX, Vol. 6932 Springer, pp. 206-218. (2010).
17. Ramirez G.: UPF at INEX 2011: Data Centric and Books and Social Search tracks. In: INEX 2011. Workshop. Pre-proceedings., pp. 117-123. Hofgut Imsbach, Saarbrücken, Germany (2011).
18. Sclaroff S., Cascia M., Sethi S., Taycher L.: Unifying Textual and Visual Cues for Content-Based Image Retrieval on the World Wide Web. In: Computer Vision and Image Understanding, Vol. 75, Nr. 1-2, pp. 86-98. (1999).
19. Sigurbjörnsson, B., Kamps, J., de Rijke, M.: An element-based approach to XML retrieval. In: INEX 2003 Workshop Proceedings. (2004).
20. Singhal, A., Buckley, C., Mitra, M.: Pivoted Document Length Normalization. In: ACM SIGIR, pp. 21-29. (1996).
21. Tjondronegoro D., Zhang J., Gu J., Nguyen A., Geva S.: Integrating Text Retrieval and Image Retrieval in XML Document Searching. In: INEX, Vol. 3977 Springer, pp. 511-524. (2005)
22. Torjmen M., Pinel-Sauvagnat K., Boughanem M.: Using textual and structural context for searching Multimedia Elements. In: IJBIDM, Vol. 5, Nr. 4 (2010), pp. 323-352.
23. Trotman, A., O'Keefe, R.A.: Identifying and ranking relevant document elements. In: INEX 2003 Workshop Proceedings. Dagstuhl, Germany (2004).
24. Zhai, C., Lafferty, J.: A Study of Smoothing Methods for Language Models Applied to Information Retrieval. In: ACM Transactions on Information Systems (TOIS), Volume 22, Issue 2, pp. 179--214. New York, USA (2004).

An Analytical View on the Software Reuse

Florinda Imeri[1], Ljupcho Antovski[2]

[1]State University of Tetovo
Ilindenska bb, 1200 Tetovo, Macedonia
florindaimeri@yahoo.com
[2]University Ss. Cyril and Methodius
Rugjer Boshkovikj 16, 1000 Skopje, Macedonia
ljupcho.antovski@finki.ukim.mk

Abstract: High-quality software, delivered on time and budget, constitutes a critical part of most products and services. Today software engineers are faced with a demand for complex and powerful software systems. To be competitive in the market software engineers are forced to create software as quickly as possible. Software reuse or component-based development (CBD) is regarded as one of the most potent software technologies in order to reduce lead times, increase functionality, and reduce costs. In the region, CBD is still a process with lot of problems, not well defined either from theoretical or practical points of view. The lack of knowledge is probably the biggest problem. Our aim is to systemize the knowledge and understanding of CBD. This research is a literature review on the software reuse and the concepts behind it. It starts with an overview of software components and CBD, continues describing benefits and obstacles to software reuse. Nontechnical aspects like legal, economic and measurement issues are covered as well. Finally examples of successful software reuse and the state of practice are summarized.

Keywords: software components, components based software development, reuse benefits, obstacles and metrics.

1 Introduction

From the time that software development started to be discussed within the IT industry, researchers and practitioners have been searching for methods, techniques and tools that would allow for improvements in costs, time-to-market and quality. Thus, an envisioned scenario was that managers, analysts, architects, developers and testers would avoid performing the same activities over and over [1].

Software reuse or component-based software development is regarded as one of the most potent software technologies in order to reduce lead times, increase functionality, and reduce costs. Software reuse and component-based development is a promising way to promote the productivity of large systems development. It is proved that designing a system that supports this approach requires more efforts and the time to market might be longer, but in the long run, the reusable approach will prove profitable[2].

Component-based software development (CBSD) involves both, the development of software components and building of software systems through integration of pre-

S. Markovski, M. Gusev (Editors): ICT Innovations 2012, Web Proceedings, ISSN 1857-7288

existing software components, developed in-house or procured from the component market [3].

In this paper we describe the key characteristics of component based development - software reuse. This research is a literature survey on the software reuse.

The paper is organized as follows. Section 2 provides some background on the concepts of software components and on reuse; section 3 reviews the advantages and obstacles to reusability; section 4 describes the key factors that influence the success and/or failure of software reuse in software development. Section 5 defines some metrics to measure software reusability; sections 6 describe the industrial best practices of software reuse; section 7 describe the state of practice of reuse, and finally discuss future directions.

2 Component-based Software Development

Component-based Software Development (CBSD) is an emerging discipline that promises to take Software Engineering into a new era. The aim of CBSD is to deliver Software Engineering from a '*cottage industry'* into an '*industrial age for Information Technology*', whereby software can be assembled from components, in the same manner that hardware systems are constructed, from kits of parts [4].

2.1 Origins of Software Reuse

According to Pareto-Diaz[5] notion of reusability began when humans start finding solutions to problems. To solve a problem, we try to apply the solution to similar new problems. If only some elements of the solution apply, than we adapt it to fit to the new problem. Proven solutions, used over and over to solve the same type of problem, become accepted, generalized, and standardized.

Software reuse was first envisioned by McIllroy[4], at a NATO Software Engineering Conference 1968, where he predicted that mass-produced components would end the software crisis. He proposed an industry of off-the-shelf, standard source-code components and envisioned the construction of complex systems from small building blocks available through catalogs. The final objective was very clear: to make something once and to reuse it several times. Morisio et al. [6] define software reuse as: the systematic practice of developing software from a stock of building blocks, so that similarities in requirements and/or architecture between applications can be exploited to achieve substantial benefits in productivity, quality and business performance. CBSD advocates the use of prefabricated pieces, perhaps developed at different times, by different people, and possibly with different uses in mind [7]. The idea involves reusing experience, such as requirements specification, design, architecture, test data and documentation. A widely growing approach for the development of information systems is component-based engineering discipline which deals with both, developing components and developing with components [5].

2.2 Software Components

A software component may be thought of as an independent module that provides information about what it does and how to use it through a public interface, while hiding its inner workings[8]. Components are seen as black-box and are binary unit of composition, whose internals cannot be viewed or accessed. Their quality characteristics can be evaluated through externally observable elements[9].

Software components may be any coherent unit of design effort that can be packaged, sold, kept in a library, assigned to one person or team to develop and maintain, and re-used. Components can be classes or frameworks; or objects that can be dynamically plugged at run-time; high-level designs; specifications; patterns; extensions to existing components; or even project plans [10].

Several studies into reuse have shown that 40% to 60% of code is reusable from one application to another, 60% of design and code are reusable in business applications, 75% of program functions are common to more than one program, and only 15% of the code found in most systems is unique and new to a specific application [11]. According to Mili et al.[3] rates of actual and potential reuse range from 15% to 85%.

According to Kim [12], there are two types of component-based reuse: with and without change to an existing component.

Reuse without change means simply selecting a component from a software component database, and dropping it into new software being developed. The cost of developing the component anew is zero! These components are called commercial components -COTS components.

Components-based reuse with change can be found in at least three types of use:

- The first and the most common is reuse of most of existing software when developing the next version of the software. Typically, some 60-80 percent of the existing software gets to be reused in this situation. However, developers do not go through the formality of "registering" components in a common software component database in this case.
- Reuse of thirty-party software, such as a sorting package, a database loader, etc. on the market or on the Internet as open source code. Again, such software is not "registered" in an organization's common software component database.
- The third type of reuse is the use of common functions available in programming language libraries, such as the math functions in the C Programming Library.

3 The Benefits and Obstacles of Software Reuse

The argument that software components will improve programmers' productivity is an old one with roots in the study of software reuse [13]. Thinking of effective software reuse as a problem-solving reuse provides a good general heuristic for judging a work product's reuse potential. For example, modules that solve difficult or complex problems (like hardware driver modules in an operating system) are excellent reuse candidates because they incorporate a high level of problem-solving expertise that is very expensive to replicate. Software reuse can have major, and possibly unfore-

seen, positive effects on the software development process. Components are standardized building blocks that can be used to assemble, rather that develop, information systems.

According to Bollinger [14] there are a number of benefits to software reuse:

- It increases the software productivity and decreases the time required for the development of software.
- By using the technique of software reuse, a company can improve software system interoperability and needs less people for software development; this provides a competitive advantage for the company and helps to produce better quality software and standardized software.
- Software reuse helps the company to reduce the costs involved in software development and maintenance; by using it the software developers can be moved from one project to the other project easily.
- Using well-tested components increase the reliability of a software system. Moreover, the use of a component in several systems increases the chance of errors being detected and strengthens confidence in that component.
- Software reuse reduces the risk involved in software development process.
- Since the documentation is very important for the maintenance of a system, reusing software components reduces the amount of documentation to be written

Even good components can corrupt a good product if they are managed in the wrong way. In some domains, such as industrial automation, this risk is unacceptable, and additional measures are required to minimize the risk [15]. There are some factors that directly or indirectly influence its adoption. These factors can be managerial, organizational, economical, conceptual and technical [14].

- Managerial and Organizational Obstacles; reuse is not just a technical problem that has to be solved by software engineers. The most common reuse obstacles are: lack of management support, project management, inadequate organizational structures and management incentives.
- Economic Obstacles**;** reuse can save money in the long run, but it is not for free. Cost associated with reuse can be: costs of making something reusable, costs of reusing it, and costs of defining and implementing a reuse process. Reuse requires up-front investments in infrastructure, methodology, training, tools and archives, with payoffs being realized only years later. Higher levels of quality, reliability, portability, maintainability, generality and more extensive documentation are necessary.
- Conceptual and Technical Obstacles; the technical obstacles for software reuse include issues related to search and recovery components, legacy components and aspects involving adaptation. In order to reuse software components there should exist efficient ways to search and recover them, this mean it is important to have a well-organized repository containing components with some means of accessing it.
- Lastly, the component must be integrated into the system under development, and thoroughly tested.

4 Success and Failure Factors

Reuse principles place high demands on the reusable components. According to de Almeida et al. [1] developing a reusable component requires three to four times more resources than developing a component for particular use. The more reusable a component is, the more demands are placed upon from products using that component. According to Poulin[16] to recover development costs, software components-assets must be reused more than dozen times. A successful program of software reuse provides benefits in three areas: increased productivity and timeliness in the software development process, improved quality of the software product and an increase in the overall effectiveness of the software development process [17].

In general there are six factors that are critical for systematic software reuse: management, measurement, legal issues, economics, design for reuse and libraries [18][6].

As with any engineering activity, measurement is vital for systematic reuse. In general, reuse benefits (improved productivity and quality) are a function of the reuse level- the ratio of reused to total components- which, in turn, is a function of reuse factors, the set of issues that can be manipulated to increase reuse, either of managerial, legal, economic as technical background [18].

As regarding to legal issues, many of which are still to be resolved, are also important, like, what are the rights and responsibilities of providers and consumers of reusable assets? If a purchased component fails in a critical application should the provider of reusable assets be able to recover damages?

Once an organization acquires reusable assets, it must have a way to store, search, and retrieve them– a reuse library. Although libraries are a critical factor in systematic software reuse, they are not a necessary condition for success with reuse. An example to this is Agora, a software prototype being developed by the Commercial Off-the-Shelf (COTS)-Based Systems Initiative at the Software Engineering Institute (SEI)[19].

5 The Metrics of Software Reuse

What makes a software component reusable? The reusability can be seen as a combination of two attributes, (re)usefulness, which means that the component addresses a common need, or provides an often requested service, and usability, which means that the component is of good enough quality and easy enough to understand and use for new software developments.

Building a reusable asset represents a more or less major investment, depending on the reuse approach used. The concept such as reuse and reusability naturally has led to questions of how to measure them, and of how to run experiments to establish their impact on quality and productivity [20].

The universally accepted truth that what cannot be measured cannot be managed also holds for the software engineering field and particularly to the area of software reuse. These lead us to another important software engineering area closely related to software reuse: software metrics[21].

Existing software reuse metrics are divided into two main categories: Economics Oriented Reuse Metrics and Models (EORM), Software Structure Oriented Reuse Metrics (SORM) and Reuse Repository Metrics (RRM)[23][22][24].

Economics oriented reuse metrics are mainly concerned with the economical aspects related to reuse programs in organizations and are the basic instruments for organization-wide return on investment models (ROI) [24]. Models or software reuse economics try to help us answer the question as when is worth to incorporate reusable components into a software development process and when custom developments without reuse are preferable. The reuse benefit corresponds to how much was saved by reusing existing reusable components. The ratio of reuse benefits to reuse investments determines if the reuse effort resulted in profit or loss to the organization[14].

Software Structure Oriented Reuse Metrics are concerned on what is being reused and how it is being reused from a strictly technical standpoint. Such metrics can be used as a supporting tool to help organizations determine where should direct their effort regarding maintenance and evolution of reusable assets available in the repository and potential new reusable assets to be developed. Such metrics are concerned on how much was reused versus how much was developed from scratch, but fail to help on the analysis of what was reused and how it was reused..

Reuse Repository Metrics (RRM) is another software reuse metrics category related to reuse repositories whose target is the assessment of reuse repository. It covers the aspects such as availability of the repository, search engine performance, quality of available assets and number of times assets were successfully reused. The efficiency of reuse repositories in aspects such as availability and quality of search results may be a decisive factor for a better reuse activity and can have a greater positive impact on the quality and the cost of the produced software.

6 Industry Projects Best Practices

"Software crisis" was first used to describe the frustration that software development and maintenance have placed on otherwise happy and productive organizations. The systematic application of software reuse to prototyping, development, and maintenance is one of the most effective ways to significantly improve the software process, shorten time-to-market, improve software quality and application consistency, and reduce development and maintenance costs. By building systems from pre-tested components, one will save the cost of designing, writing and testing new code. There are significant corporate reuse programs at AT&T, HP, IBM, GTE, NEC, Toshiba, Hitachi, Ltd, Motorola, Inc., National Aeronautics and Space Administration (NASA), and others [1][25].

The United States approach to software reuse is concentrated on tools, technology, and feature sets. There is significant work at research consortia such as MCC, SPC, SEI, and STARS program founded by DoD of US.

Manufacturers of computer systems and instruments, such as Hewlett-Packard Co. and IBM, whose businesses relied mostly on hardware and mechanical engineers, today find that over 70 percent of their research and development engineers are working in the areas of software and firmware. Maintenance and rework account for about 60 to 80 percent of the total software costs. It appears that product development costs, factoring in the cost of producing, supporting, and integrating reusable software com-

ponents, can decrease by a sustainable 10 to 12 percent; defect rates in delivered products can drop drastically to 10 percent of their former levels; and long-term maintenance costs can drop to 20 to 50 percent of their former values when several products share the same, high-quality components [26][25]. IBM has a corporate program with several reuse support centers, a large library, and a multisite Corporate Reuse Council, key aspect of which is formal identification of reuse "champions" and agents [27].

Starting from 1988, AT&T developed a domain-specific, large-scale software-bus system for on-line transaction processing and network management. With a support staff of about 30, their reuse program has reduced development costs by about 12 percent and time-to-market from 18-24 months to 6-9 months[28] [29].

According to Joos, to increase quality and productivity, Motorola around the 90's had software reuse as one of its pillars [30]. To implement reuse approach several things were needed. As first the support from top management and reuse training for engineers is crucial; second, to provide incentives as support to the program until reuse becomes part of the culture; and third, provide tools to help engineers. Motorola started a 3-phase software reuse process. The first phase included creation of a reuse task which had two main activities: to educate how to spread the software reuse knowledge; and to emphasize metrics, rewards, career paths and job descriptions for producers and users of reusable components. At the second phase, senior management accepted the responsibility of initiating reuse by getting the upper management more involved with advancing the state-of-the-practice in software. And finally, at the last phase, groups of developers realized that software reuse promises more than internal savings since it provides new markets for reusable components and support tools.

Most of the European work consists of industrial or industrial-academic consortia. The ESPRIT initiative has funded several industrial-academic collaborations, such as KNOSOS, PRACTITIONER, ITHACA, SCALE, REDO, and REBOOT [31]. The objectives of the REBOOT project are to study, develop, evaluate and disseminate advanced methodologies and environment for reuse-driven and object-oriented software development with the emphasis on planned reuse.

According to Griss [26] the Japanese approach to reuse is concentrated on core functionality, design, productivity, and quality. They were able to produce much higher quality systems at a faster rate. Hitachi reduced the number of late projects from 72 percent in 1970 to 12 percent in 1984, and reduced the number of defects to 13 percent of the 1978 level. Toshiba increased productivity by a factor of 250 percent between 1976 and 1985, and by 1985 had reduced defect rates to 16 to 33 percent of the 1976 level. NEC improved productivity by 126 to 191 percent and reduced defect rates to 33 percent of prior levels. Nippon Telegraph & Telephone Corp. (NIT) has a comprehensive program including a reuse-specific organization, printed catalogs, guidelines, and certification for reuse, leading to reuse levels of 15 percent or more, with several hundred small components[32].

7 State of Practice of Software Reuse in the Cloud

To support scientific research, the development of technological infrastructure have include cloud computing, which offers capabilities for research communities to utilize scientific data and services in order to conduct analyses that otherwise would be more

costly or prohibitive to complete [33]. Cloud providers install and operate software applications in the cloud and cloud users access it from clients. Cloud computing is a recent example of a technology that can benefit from software reuse. Employing software reuse techniques enables the adoption of new technological approaches to support new models for conducting scientific research. The software reuse methods and tools also can contribute to the evaluation of systems being developed to support scientific research. Software reuse instruments, such as the Reuse Readiness Levels (RRLs) [34], offer the potential to improve capabilities for software development and evaluation when developing and adopting software assets for scientific support systems [35].

OOS development as a special instance of software reuse development, typically takes place in informal collaborations of globally distributed teams communicating over the internet. Software repositories on the Internet provide a tremendous amount of freely reusable source code, frameworks and libraries for many recurring problems. The Internet itself has become an interesting reuse repository for software projects [36]. Search engines like Google Code Search provide powerful search capabilities and direct access to millions of source code, written in a multitude of programming languages. It allows search of publicly available source code by language, package name, file name, and types of license. The extraordinary success of some of the resulting software products (such as GNU/Linux, Apache, Bind DNS server, OpenOffice, Mailman) has drawn attention from the public and both software creating and software-using organizations to this way of developing software.

8 Conclusion and Future Work

Software reuse has been practiced since programming began. Reuse as a distinct field of study in software engineering, however, is often traced to Doug Mcilroy's paper that proposed to base the software industry on reusable components.

Clearly it is seen that software reuse is an inevitable solution that has potential to improve time –to-market and man power/cost trends that have been ongoing. Currently seem to be one of the most active and creative research areas in Computer Science. Software reuse has a significant impact on software industry. It helps organize large-scale development and what is more important; it makes system building less expensive.

Software reuse has been a buzz word in large companies for some time now, with its potential for achieving good quality systems in short time scales by the reuse of currently available components. Many success stories have been quoted, but what about the small, less structured companies, whose livelihood depends on the ability to produce their product as quickly as possible, while trying to keep standards high enough to keep their customers happy and their maintenance costs low. To them, the benefits of software reuse could be invaluable. One important issue is how to make best use of reusable components at the companies of small size.

Our future work will focus on the small software development companies in Macedonia and the SEE region to reevaluate the issues surrounding software reuse from the perspective of developers involved in a software development. In particular, we want to explore their experience with software reuse and possibly try to increase the

knowledge and understanding of CBD. We would like to look at the possible benefits, disadvantages and contributors towards successful reuse of software components

REFERENCES

[1] Eduardo Santana de Almeida, A. Alvaro, V. C. Garcia, J. C. C. P. Mascena, V. A. de A. Burégio, L. M. Nascimento, D. Lucrédio, and S. L. Meira, *Component Reuse in Software Engineering*. .

[2] I. Crnkovic and M. Larsson, "Challenges of Component-based Development Challenges of Component-based Development," *Computer Engineering*, vol. 61, no. 3, pp. 201–212, 2002.

[3] H. Mili, F. Mili, and A. Mili, "Reusing software- Issues and research directions," *IEEE Transactions on Software Engineering*, no. 6, pp. 528 – 562, 1995.

[4] *Component-Based Software Development-case studies*, vol. 1. World Scientific Publishing Co. Ptc. Ltd. 5 Toh Tuck Link, Singapore, 2004.

[5] R. Prieto-Diaz, W. Schafer, and M. Matsumoto, "Status report: Software reusability," *Software, IEEE*, vol. 10, no. 3, pp. 61–66, 1993.

[6] M. Morisio, M. Ezran, and C. Tully, "Success and failure factors in software reuse," *IEEE Transactions on Software Engineering*, vol. 28, no. 4, pp. 340–357, Apr. 2002.

[7] A. Cechich and M. Piattini, *Component-based software quality: methods and techniques*. 2003.

[8] T. N. Form, "Chapter 1. Introduction to IEEE Std. 1517— Software Reuse Processes 1.1," *Components*.

[9] M. F. Bertoa, J. M. Troya, and A. Vallecillo, "Measuring the usability of software components," *Journal of Systems and Software*, vol. 79, no. 3, pp. 427–439, Mar. 2005.

[10] A. C. Wills, D. D. Souza, and I. Computing, "Rigorous Component-Based Development," *Components*, pp. 1–28, 1997.

[11] M. Ezran, M. Morisio, and C. Tully, *Practical software reuse*. 2002.

[12] W. Kim, "On Issues with Component-Based Software Reuse," *Journal of object technology*, vol. 4, no. 7, pp. 45–50, 2005.

[13] C. Buhman and F. Long, "Volume I : Market Assessment of Component-Based," *Internal Research and Development*, vol. I, 2000.

[14] B. Bollinger and B. Barnes, "Making Reuse Cost -Effective," *IEEE Software*, 1991.

[15] I. Crnkovic and M. Larsson, "Overview Component-based Software Engineering State of the Art Report," *Computer*.

[16] J. S. Poulin, "The Business Case for Software Reuse: Reuse Metrics, Economic Models, Organizational Issues, and Case Studies," *9th International Conference on Software Reuse*, vol. 4039/2006, p. 439, 2006.

[17] Y. Kim, "Software reuse: survey and research directions," *Journal of Management Information Systems*, 1998.

[18] W. B. Frakes and S. Isoda, "Success factors of systematic reuse," *IEEE Software*, vol. 11, no. 5, pp. 14–19, Sep. 1994.

[19] R. C. Seacord, S. A. Hissam, and K. C. Wallnau, "Agora : A Search Engine for Software Components Agora : A Search Engine for Software Components," no. August, 1998.

[20] W. Frakes and C. Terry, "Software reuse: metrics and models," *ACM Computing Surveys*, vol. 28, no. 2, pp. 415–435, Jun. 1996.

[21] K. Jasmine, "Cost estimation model for reuse based software products," *IMECS 2008: International Multiconference of*, vol. I, pp. 19–21, 2008.

[22] J. Mascena and E. de Almeida, "A comparative study on software reuse metrics and economic models from a traceability perspective," *Information Reuse and Integration*, pp. 72–77, 2005.

[23] W. Lim, "Effects of reuse on quality, productivity, and economics," *Software, IEEE*, 1994.

[24] J. Poulin, "An agenda for software Reuse Economics," *International Conference on Software Reuse*, no. April, 2002.

[25] K. Wentzel, "Software reuse-facts and myths," *Software Engineering, 1994. Proceedings. ICSE-*, 1994.

[26] M. L. Griss, "Software reuse: From library to factory," *IBM Systems Journal*, vol. 32, no. 4, pp. 548–566, 1993.

[27] J. Tirso, "The IBM reuse program," *of the 4th Annual Workshop on Software Reuse*, 1991.

[28] R. Martin and G. Jackoway, "Software Reuse Across Continents," *Workshop in Reuse*, pp. 1–6, 1991.

[29] J. Tirso, "Championing the cause: making reuse stick," *... of the 5th Annual Workshop on Software Reuse*, 1992.

[30] R. Joos, "Software reuse at Motorola," *IEEE Software*, vol. 11, no. 5, pp. 42–47, Sep. 1994.

[31] J. Faget and J. Morel, "The REBOOT approach to the concept of a reusable component," *5th Workshop Institutionalizing*, 1992.

[32] M. Aoyama, "CBSE in Japan and Asia," Addison-Wesley Longman Publishing Co., Inc. Boston, MA, USA, 2001, pp. 213–225.

[33] J. Marshall, R. R. Downs, and C. A. Mattmann, "Software reuse methods to improve technological infrastructure for e-Science," *Information Reuse and Integration (IRI), 2011 IEEE International Conference*, pp. 528–532, 2011.

[34] J. Marshall, S. Berrick, and A. Bertolli, "Reuse Readiness Levels (RRLs)," *sciencedatasystems.org*, 2010.

[35] P. Vitharana, J. King, and H. S. Chapman, "Impact of Internal Open Source Development on Reuse: Participatory Reuse in Action," *Journal of Management Information Systems*, vol. 27, no. 2, pp. 277–304, Oct. 2010.

[36] W. Frakes, "Software reuse research: Status and future," *Software Engineering, IEEE Transactions*, vol. 31, no. 7, pp. 529–536, 2005.

Model of a Generic Classification System based on a Multiple Kernel Data Fusion

Kristina Spirovska [1], Ana Madevska Bogdanova[1], Ph.D[1]
[1] Faculty of Computer Science and Engineering
1000 Skopje, Republic of Macedonia
{kristina.spirovska, ana.madevska.bogdanova}@finki.ukim.mk

Abstract. Classification, as a machine learning technique, can be applied for problem solving in almost every aspect of our lives. That is the reason why today there are lot of specialized classification systems and tools, based on different methods, whose purpose is to solve a specific problem. In this paper, we propose a model of a generic classification system based on a multiple kernel data fusion. The model describes a system that does a data pre-processing, intelligent choice of a classification method, proposes a solution and automatically tunes the parameters. This model is also parallel and scalable and it integrates multiple heterogeneous data sources. The main goal is to create a simple system that will focus the user on a method that he/she would further tune to obtain the final solution of a given problem.

Keywords: models of classification systems, generic classification, multiple kernel data fusion

1. Introduction

The design of systems that can learn from provided input data and use the gained knowledge to improve the processing of similar inputs in the future, has always been a challenging problem in the machine learning domain.

In this paper we propose a model of a generic classification system based on multiple kernel data fusion. To begin with, this section is devoted to a brief introduction to the main method for classification used in this paper - multiple kernel learning method. Section 2 introduces other existing models of classification systems. In section 3 we present our model. Section 4 discusses the feasibility of the model. Finally, the last section concludes this work.

S. Markovski, M. Gusev (Editors): ICT Innovations 2012, Web Proceedings, ISSN 1857-7288

1.1 Kernel methods

Kernel methods represent a class of algorithms for pattern analysis. These methods approach the problem by mapping data into a high dimensional feature space where every coordinate represents one attribute of the data entities. In that space we can use a variety of methods for finding relations between data.

Formally speaking, if we have two data instances x_i, $x_j \in X$ and a mapping ϕ: $X \rightarrow \Re^N$, we can define kernel function as

$$k(x_i, x_j) = < \phi(x_i), \phi(x_j) > \quad (1)$$

The data instances, x_i and x_j, could be elements from any set. On the other hand, their mappings $\phi(x_i)$ and $\phi(x_j)$ represents vector in $\Re^N$. The matrix $K_{ij}=(x_i,x_j)$, where x_i, x_j are data points, is called "kernel matrix" or "Gram matrix". The kernel matrix is a symmetric positive definite matrix. Because it specifies the inner products between all pairs of data points, it completely determines the relative positions between those points in the embedding space [6] [16].

In other words, a kernel matrix is a representation of the input data space through defined similarities between pairs of its instances, which is achieved through the kernel function [19]. Each kernel function provides a partial description or view of the data and extracts a specific type of information, i.e. the kernel function defines generalized similarity relationships between the pairs of instances.

1.2 Multiple kernel learning

Multiple kernel learning (MKL) has been pioneered by Lanckriet et al. [7] as an extension of the single kernel SVM. The idea behind MKL is the fact that multiple sources could be seen as multiple partial descriptions of the data instances. The benefit of using them comes from the fact that different data sources are likely to contain different and partly independent information about the task. If we combine those complementary pieces of information we might enhance the total information about the problem at hand [14].

For example, consider a major movie data repository such as the Internet Movie Database (IMDb) and the task of determining whether movie would be interesting for a certain user. That is a classification task, the movie has to be classified as interesting or not interesting for that particular user. The movies which are classified as interesting would be suggested to the user. For that purpose multiple data sources could be used, like the movie storyline, the genres of the movie, the actors, the

director and the writers of the movie, the other user reviews, the ratings and meta-score, the budget of the movie, the language of the movie.

Other example could be predicting the function of a protein using multiple sources like variable length amino acid strings, real-valued gene expression data and a graph of protein-protein interactions.

During the past decade, it has been shown that classifiers that use combinations of multiple kernels instead of classical single kernel-based ones attain better results in certain problems [19] [20] [21].

This mathematical problem could be solved with combination of those partial descriptions using a convex optimization method known as semi-definite programming (SDP) [15] [16]. This SDP-based approach gives us a general methodology for combining many partial descriptions of data that is statistically sound, as well as computationally efficient and robust.

The solutions found by kernel-based algorithms such as the support vector machine (SVM) are linear functions in the feature space:

$$f(x) = w^T \Phi(x) \tag{2}$$

for the optimal weight vector w. The kernel can be exploited whenever the weight vector can be expressed as a linear combination of the training points $w = \sum_{i=1}^{m} \alpha_i \Phi(x_i)$, implying that we can express *f* as follows:

$$f(x) = \sum_{i=1}^{m} \alpha_i k(x_i, x) \tag{3}$$

An important issue in the kernel applications is the choice of a kernel k for a given learning task. Intuitively, we try to choose a kernel that induces the "right" metric in the space [16].

2. Related work

Today there are lots of models of classification systems. Describing a generic classification system model and creating this kind of a system has been ambitious job for many researchers in the field of data mining and machine learning.

Over the years, many of the proposed approaches aim at the construction of an (approximately) accurate classifier. That is the case in paper [11], where a generic framework for rule-based classification is proposed.

Nevertheless, researchers had turned their focus to creating models and systems specialized for a single or a class of problems because, as we mentioned earlier, a great deal of the choice of the classifier depends on the nature of the problem.

G. Lanckriet et al. propose a statistical framework for genomic data fusion in [12]. Endeavour is a web resource for the prioritization of candidate genes based on genomic data fusion. It is based on genomic data fusion techniques, whose principles are described in [3]. The model of the system, as well as details of its implementation, are described in [4].

The Weka workbench, as stated in [8], is a collection of state-of-the-art machine learning algorithms and data preprocessing tools. Another similar machine learning toolbox is Shogun, whose capabilities are described in [9]. One characteristic about this toolbox is that it offers a multiple kernel classification algorithm and currently is the only toolbox that provides this option.

Almost all of the existing frameworks and software tools for classification focus on a particular class of problems instead of developing general principles and techniques.

Having that idea in mind, Luc De Raedt in [22] proposed to "alleviate these problems by applying the constraint programming methodology to machine learning and data mining and to specify machine learning and data mining problems as constraint satisfaction and optimization problems. What is essential is that the user to be provided with a way to declaratively specify what the machine learning or data mining problem is rather than having to outline how that solution needs to be computed." [22].

3. The model description

In this section, we give a description of our model of a generic classification system based on a Multiple Kernel Data Fusion.

The system is generic, which means that it could be used for different kind of classification problems with heterogeneous data sources. The classification system is composed of four main components: Preprocessor, Training unit, Data Fusion Unit, Classification Unit. The data flow diagram of this system is shown on Fig. 1.

3.1 Preprocessor

Preprocessing is the first phase of the classification process and contains different preprocessing processes:

1. dealing with irregular or inconsistent data,
2. data transformations.

When data are given to the Preprocessor, missing values, outliers and data noise are being handled by standard preprocessing techniques [17]. Also, sometimes data needs to be normalized to prevent features with a large range to out-weight features with smaller ranges. In order to implement the classification algorithm the nominal input data is required to be transformed into an algebraic model, so then it can be processed. The standard practice in information retrieval on textual data is to represent input data as vectors in t-dimensional Euclidean space where every dimension corresponds to a word (term) of the vocabulary. Several text tokenization methods are used, such as Bag of words and N-Grams.

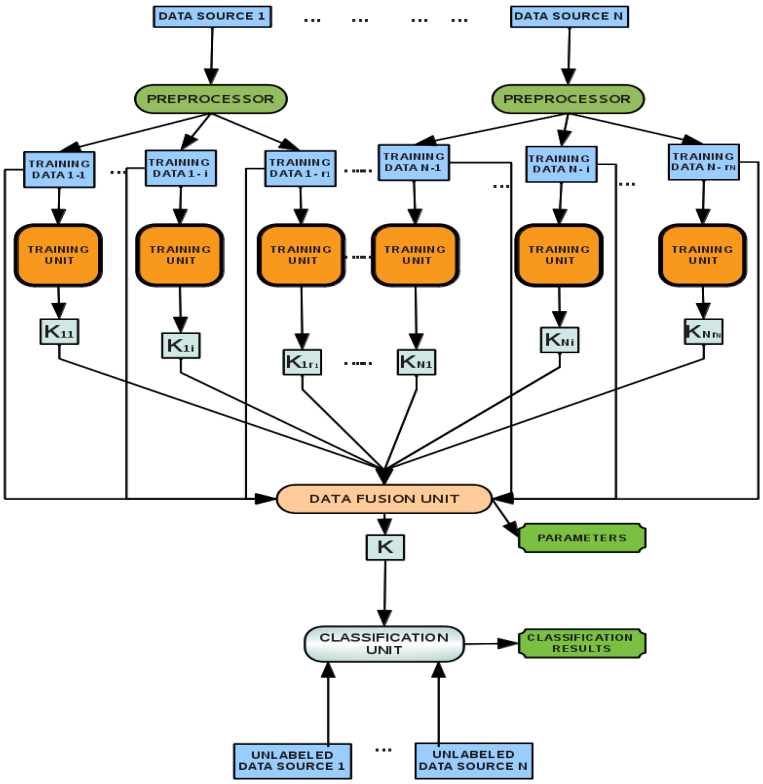


Fig.1. Data flow diagram of the generic classification system based on MKDF

We use the fact that there could be different representations of the data space from one data source. For example if the users have data sets with many missing values, the system has to handle this problem. There are several ways to conduct this: replacement with a descriptive statistic, deletion, attempt to predict their values etc. The model applies several different techniques in order to encapsulate all the possible information that we can gain from the data.

Suppose that the user has N data sources, which serve as an input to the Preprocessor. For every data source, the Preprocessor creates r_i different transformations, i=1..N. At the end, we will get M data sources, M >= N, where $M = r_1 + ... + r_n - N$.

3.2 Training Unit

The training unit is the most complicated component of the system. Its role is to create kernel matrix representations from every training data source retrieved from the Preprocessor unit.

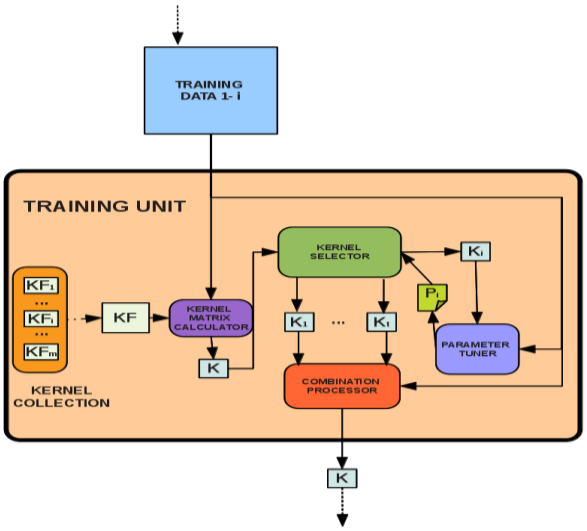


Fig.2. Data flow diagram of the Training Unit

The training unit is composed of Kernel Collection (KC), Kernel Matrix Calculator (KMC), Kernel Selector (KS), Parameter Tuner (PT) and Combination Processor (CP) components.

The goal of the training unit is to find *m* best fitting kernel functions to build kernel matrices that best represent the problem space in order to combine them further.

The Kernel Collector is a collection of kernel functions, i.e. that is the place where a set of kernel functions are stored, for example: Linear, Polynomial, Guassian, Exponential, Laplacian, ANOVA etc.

The Kernel Matrix Calculator (KMC) is a component which calculates the kernel matrix from the data. It has the training data and the kernel function as input. Its output is a kernel matrix which is given further to the KS.

Kernel Selector unit selects the *m* best suited kernel matrices. Firstly, it sends the received kernel matrix from the KMC to the PT. Afterward, when the PT has finished its work, the KS receives the calculated precision value *p* from the PT. The precision

value is used as a measure of the feasibility of a kernel matrix. When precision values for all kernel matrices have been sent, the KS unit selects only the *m* best matrices. Those *m* best kernel matrices are at the end send to the Combination Processor.

The Parameter Tuner serves as a calibrator for the best possible parameters of the used kernel function for the input matrix. Every kernel function has its own parameters which have to be calibrated in order to achieve satisfactory classification precision *p* for the given problem. It uses cross-validation techniques for that purpose. In order to use cross-validation techniques, the PT needs the training data as its second input. Those data are divided into training and testing data which the cross-validating method will use to tune the parameters. The output from the PT is the precision *p* of classification gained with the cross-validation process.

The Combination Processor is the component that creates a data fusion of the homogeneous data. For now we have presented single data source and its *m* kernel matrices. The next step is to combine them into one kernel matrix which is the output from this unit and the training unit in general. The process of this combining kernel matrices is described in section 1.2 of this paper.

3.3 Data Fusion Unit

The Data Fusion Unit (DFU) is the most important component of the system. Its task is to apply the data fusion process on heterogeneous sources represented with kernel matrices. From every data source we get one kernel matrix which also is a combination of multiple kernel matrices representing the view of particular data source. So if we have M training data sources, the training unit will output M kernel matrices which will be combined in the DFU. The combination of different views from different sources is accomplished with data fusion algorithm using Eq. (3). We have M kernel matrices that gave best precision in the previous fusion on single sources. The Data Fusion unit will find the best combination from the given best kernel matrices from every training unit. It does not necessary has to have all the matrices in the final combination. The result is a single kernel matrix K, which goes into the Classification Unit, and the parameters of the kernel function are used to create this classification model. Completing this section, the process of training is finished and a model of the classifier has been build.

3.4 Classification Unit

After the system has been trained and the classification model has been created, the system is ready to handle the unlabeled data with the Classification Unit (CU). CU is

composed of two components, the Preprocessor, and the MKL Classifier (MKLC). Its inputs are the unlabeled data sources, and its outputs are the classification results.

Before the unlabeled data enter the MKL Classifier, they have to be preprocessed in the same way as in the training phase. This is done with the same Preprocessor unit described in section 3.1. After the preprocessing, the unlabeled data are sent to the MKL Classifier. It is based on MKL algorithm techniques described in [19][20][21]. MKLC also receives another input – the combined kernel matrix that is of key importance to this system and the classification process. It represents a combination of the different views from the different sources.

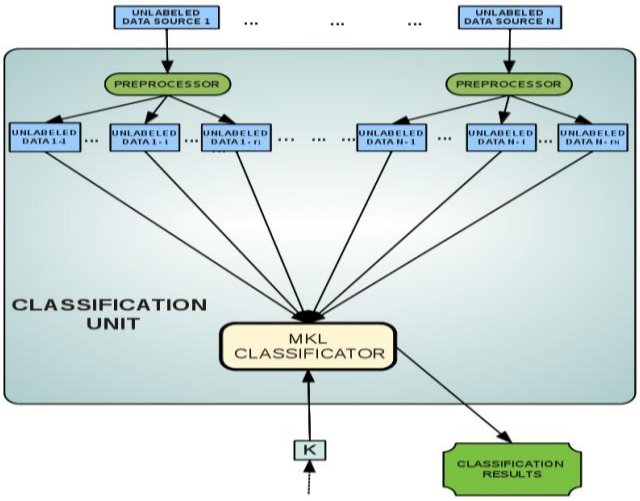


Fig.3. Data flow diagram of the Classification Unit

The output from the CU is the output from the whole system. It gives the classification results or the labels or classes of the unlabeled data that we wanted to classify / label.

4 Discussion

Experience shows that no single machine learning method is appropriate for all possible learning problems. Although there are such attempts, all of them have some advantages and disadvantages.

Weka [8] and Shogun [9] both represent large collection of learning algorithms that the user can combine according to his needs. The user has to make some decisions: he has to choose the method according to his background and knowledge of the nature of the problem and data and he has to make a lot of experiments to find the best parameters for the chosen method. The user is the key player of these systems.

Also there are systems where the user uses declarative language to instruct the system about the problem that needs to be solved and the system automatically transforms such inputs into a format that can be used by a solver to efficiently generate a solution. The disadvantage is that the user will have to learn how to express his problem on the declarative language.

The main advantage of the proposed model is the simplicity of its use. The user does not have to have deep understanding of the learning algorithms for the classification process. The only thing he has to provide is the data, and the system will give him the preliminary results. The accuracy of the system is based on the accuracy of the methods it uses in the different stages of the classification process.

5 Conclusion

In order to get accurate models for the classification problems, we must make the right choice of the learning algorithm for the problem at hand. Afterward, the right parameters of the method must be chosen, which is also of great importance for the classification process.

The model presented in this paper does not depend on the user or the structure of the data or the nature of the problem. It can classify data coming from different sources with homogeneous or heterogeneous nature. As a result of the classification process, the system will propose a method and the best choice of parameters. Those results are very useful because they will focus the user on a method that he/she would further tune to obtain the optimal solution of the given problem.

The next step is implementation and testing of this proposed model. Further, it can be improved with automatic data sources separation process. The user would only provide a database containing all the sources for the data, and the system will separate the data sources.

References

1. Kotsiantis, S.B.: Supervised Machine Learning: A Review of Classification Techniques.; Informatica (Slovenia)(2007)249-268
2. Tan, P., Steinbach, M., Kumar, V.: Introduction to Data Mining. ; (2005) 145-198
3. Aerts S, Lambrechts D, Maity S, Van Loo P, Coessens B, De Smet F, Tranchevent LC, De Moor B, Marynen P, Hassan B, Carmeliet P, Moreau Y. ; Gene prioritization through genomic data fusion; Nature Biotechnology 24, (2006) 537 – 544
4. Tranchevent LC, Barriot R, Yu S, Van Vooren S, Van Loo P, Coessens B, De Moor B, Aerts S, Moreau Y; ENDEAVOUR update: a web resource for gene prioritization in multiple species;

5. Ding, Q., Ding, Q., Perrizo, W.: Decision tree classification of spatial data streams using Peano Count Trees. ;In SAC(2002)413-417
6. Yu, S., Tranchevent, L., Moor, B.D., Moreau, Y.: Kernel-based Data Fusion for Machine Learning - Methods and Applications in Bioinformatics and Text Mining. ;Studies in Computational Intelligence(2011)1-208
7. Lanckriet, G.R.G., Cristianini, N., Bartlett, P.L., Ghaoui, L.E., Jordan, M.I.: L6. ;Journal of Machine Learning Research(2004)27-72
8. Frank, E., Hall, M., Holmes, G., Kirkby, R., Pfahringer, B., Witten, I.H., Trigg, L.: Weka-A Machine Learning Workbench for Data Mining. ;Computer software engineering series(2010)1269-1277
9. Sonnenburg, S., Rätsch, G., Henschel, S., Widmer, C., Behr, J., Zien, A., Bona, F.D., Binder, A., Gehl, C., Franc, V.: The SHOGUN Machine Learning Toolbox. ;Journal of Machine Learning Research(2010)1799-1802
10. Classification Software for Data Mining and Analytics, http://www.kdnuggets.com/software/classification.html
11. Arnaud Giacometti , Eynollah Khanjari Miyaneh , Patrick Marcel , Arnaud Soulet ; A Generic Framework for Rule-Based Classification
12. Lanckriet, G.R.G., Bie, T.D., Cristianini, N., Jordan, M.I., Noble, W.S.: A statistical framework for genomic data fusion. ;Bioinformatics(2004)2626-2635
13. Konstantin Tretyakov; Methods of Genomic Data Fusion: An Overview; (2006)
14. Gert R. G. Lanckriet, Tijl De Bie, Nello Cristianini, Michael I. Jordan, William Stafford Noble; A Framework for Genomic Data Fusion and its Application to Membrane Protein Prediction (2003)
15. L. Vandenberghe and S. Boyd. Semidefinite programming. SIAM Review, 38(1):49-95, 1996.
16. Lanckriet, G.R.G., Cristianini, N., Bartlett, P.L., Ghaoui, L.E., Jordan, M.I.: Learning the Kernel Matrix with Semi-Definite Programming. ; In ICML(2002)323-330
17. Zhang, S., Zhang, C., Yang, Q.: Data Preparation for Data Mining. ;Applied Artificial Intelligence(2003)375-381
18. Souza, César R.: Kernel Functions for Machine Learning Applications. ; (2010) http://crsouza.blogspot.com/2010/03/kernel-functions-for-machine-learning.html
19. Bach, F.R., Lanckriet, G.R.G., Jordan, M.I.: Multiple kernel learning, conic duality, and the SMO algorithm. ;In ICML(2004)
20. Rakotomamonjy, A., Bach, F., Canu, S., Grandvalet, Y.: More efficiency in multiple kernel learning. ;In ICML(2007)775-782
21. Sonnenburg, S., Rätsch, G., Schäfer, C., Schölkopf, B.: Large Scale Multiple Kernel Learning. ;Journal of Machine Learning Research(2006)1531-1565
22. Luc De Raedt:Declarative modeling for machine learning and data mining; In Lecture Notes in Computer Science, Volume 7278, Formal Concept Analysis; (2012) p.2
23. Ana Madevska Bogdanova, Nevena Ackovska "Data Driven Intelligent Systems", ICT Innovations 2010 Proceedings ISSN 1857-7288
24. Kristina Spirovska, Ana Madevska Bogdanova; Multiple kernel learning Methods and their application in Yeast Protein Subcellular Localization Prediction ; CIIT 2012 Proceedings

Mobile phone applications for motivating physical activity

Ilina Kareva

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

Jovan Kostovski

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

Andrea Kulakov

Faculty of Computer Science and Engineering, Rugjer Boshkovikj 16, Skopje Macedonia
andrea.kulakov@finku.ukim.mk

Abstract. The goal of the research described in this article was to determine the possibilities of using the widely spread every day used smartphones to motivate physical activity among the youngest users. Another aim was to find out to which degree these devices and applications can replace the expensive exergaming equipment offered on the market. Three prototype applications were developed and tested with the target group. The applications can be used indoor and outdoor, they offer mental training and the possibility of user generated content which makes them highly customizable to the user's needs and ages. The applications are entertaining, educational and at the same time motivate physical activity.
Applications were tested in a kinder-garden with a group of 30 children at the age of 5 and 75 primary school students at the age of 10-11 and they were well accepted by both groups. Teachers and parents present during the testing found the applications very helpful for educational purposes as well as for motivating physical activities and activities that can be used in the free-time.
The principal conclusion was that the smart phones with these applications are decent replacement for the expensive equipment offered for edutainment and exergaming. On the other hand, the possibility to create new levels boosts the involved parties' (both parents and children) creativity.

Keywords: mobile development, physical activity, exergaming, edutainment

adfa, p. 1, 2011.

S. Markovski, M. Gusev (Editors): ICT Innovations 2012, Web Proceedings, ISSN 1857-7288

1 Introduction

In the recent past the computers left the research labs and moved into the everyday life of people. Recently they are in our pockets in the form of so-called "smart phones". This migration certainly influenced the lifestyle, behavior of people and the whole society. Although in many areas they justify their existence, their usage has serious draw-backs in children's world. The advancement of computer hardware became foundation for development of better video games. Game consoles equipped with powerful processors enabling games with excellent 3D graphics and features to be played immediately became widely accepted in homes. These are another reason for children's inactivity.

The percentage of overweight children is growing at an alarming rate with each 1 in 3 children are considered obese [1]. Many children spend less time exercising and more time in front of the gaming devices. Once this problem was spotted some companies offered products on the market with a purpose of solving it. The game consoles got wireless controllers with embedded sensors that can reflect the human's position at every point in time and thus determining the movements [2,3]. So seating was suddenly replaced by a physical activity. That is defined as exergaming. The aim of this paper is to bring exergaming to a device that is part of the everyday life of many people nowadays - smartphones.

2 Related Work

2.1 Existing solutions for mobile devices that motivate physical activity

With the decreasing prices of sensors like small size multi-pixel cameras, accelerometers, GPS devices and Bluetooth devices they became widely used for research in laboratories. Usually the goal is to turn the phone as a game controller or create applications that measure the amount of physical activity and spent calories. When it comes to motivating children's activity the situation is different. Children would never keep logs of their activities, measure how many kilometers they have run or how long they have been exercising, instead they should be engaged to exercise by the fun the game offers. We will mention some good examples where this has been achieved. MarioFit[4] is system for playing the Nintendo game SuperMarioBros using a handheld computer and natural human movements as a mechanism for entering data. In this implementation MultiSensorBoard is used for collecting data from the accelerometer and compass. Then based on these data six different movements are used as input for the game: jumping, low walking, turning, walking, running and throwing. This project was developed in 2005 when smart phones were still not widespread, and are very expensive so it never gained any significant popularity. Another project in which augmented reality comes to the streets of Singapore is physically interactive version of the famous arcade game - Pacman. Human Pacman [5] is a real mobile entertainment system built on the concepts of ubiquitous computing, human-computer interaction and networks entertainment from a wide range. Players interact with each

other and with digital 3D PacWorld placed in their range of view using a portable computer, headset and goggles. One player has the role of Pacman and others are ghosts and their playground are the streets from the real world augmented through the goggles. GPS is used for locating and WLAN for communication. These projects are example that mobile and portable devices can be used to create a new genre in computer games that has the potential to help solving obesity problem.

Today smartphones are equipped with powerful processors, large memory and different types of sensors - camera, GPS, compass, light sensor, temperature sensors and they offer good basis for development of various applications.

3 Applications' Overview

For purposes of this research three applications were developed: RunGame, ColorGame and MapGame. For playing RunGame a smartphone and a personal computer or laptop with a monitor or projector is required and the phone is used as a controller. ColorGame is a game designed to replace interactive walls[6]. MapGame is an application designed for playing outdoors. In this application GPS location of the user is read and he needs to reach certain points given on a map which hold a different tasks. Augmented reality view is used to guide the user to the next point of interest.

Important aspects in the development of the games were:

- Easy installation
- Easy connection
- Possibility to work without an internet connection
- Possibility to create new challenges – levels
- The phone as sufficient device for playung without additional hardware or server for data processing

4 Applications' Implementation

4.1 RunGame

The purpose of this application is to transform the phone into game controller like the ones used with consoles. The application sends accelerometer readings while their processing and determination of the player's movement is executed by the desktop application. The desktop application is a game from the book by David Brackeen - Developing game in Java[7]. This game is chosen for this research because it has a part for editing levels. Additionally a module for communication and movement classifier were added to the implementation.

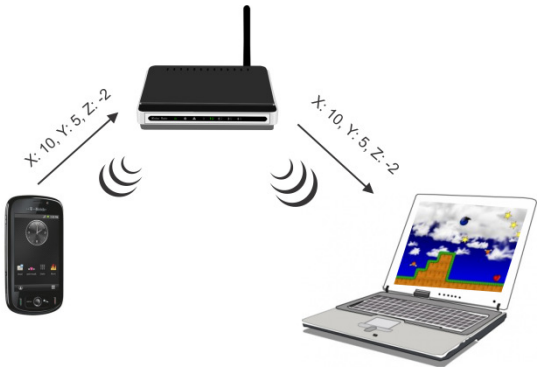
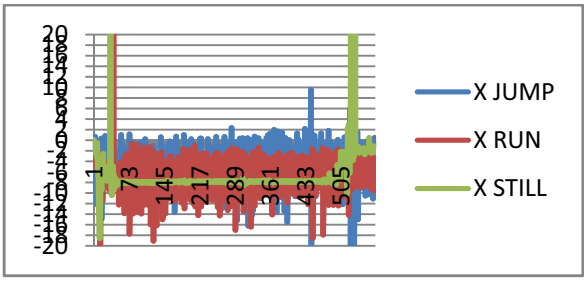


Fig. 1. Setting connection

The built-in accelerometer detects movement in X- Y- and Z-axes. The most challenging task was the determination of the user's activity in real-time: running, standing still or jumping. For the purpose of this application an accelerometer logger application was developed. With the help of this application few testing data files were created with different activities: standing in place, running and jumping. From their plots one can come to several conclusions. Since the application needs to use only three types of motion the classification can be done only by using the values of two axes. When the phone is placed in the pocket the Z-axis always gives the same values because it is parallel to the ground. Depending on whether the phone is set with X- or Y- axis pointing toward the earth the data from one of these two is relevant for making conclusions.



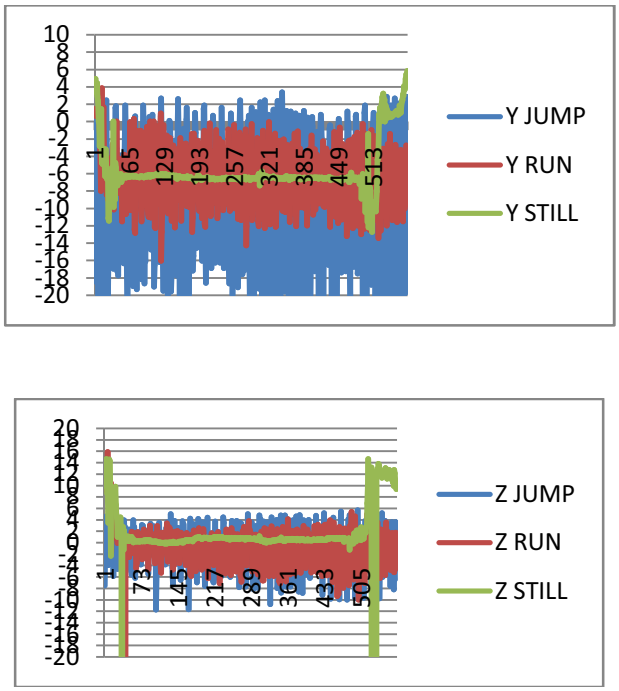


Fig. 2. X-, Y- and Z- axes logged values from different activities

The algorithm that was used in the preparation of the classifier determines the difference between maximum and minimum value of the sensor readings on the axes in a window of 30 readings.

$$\Delta x = \max(x) - \min(x) \mid x = x_k .. x_{K+30} \quad (1)$$

$$\Delta y = \max(y) - \min(y) \mid y = y_k .. y_{K+30} \quad (2)$$

In other approaches[8,9,10,11,12,13] windows with width three times greater than the frequency of the accelerometer readings is used. In our approach a window with a size one third from the accelerometer readings is used because it gave best results when changing from one activity to other. This was determined after several tests with windows of varying width. This simple approach is sufficient for the application to feel like real-time. According to test this approach has proven well in cases when the user holds the device in his hand. Another important feature is that users can create their own levels with different length and weight, with more running or jumping. The game elements are displayed on the next images.

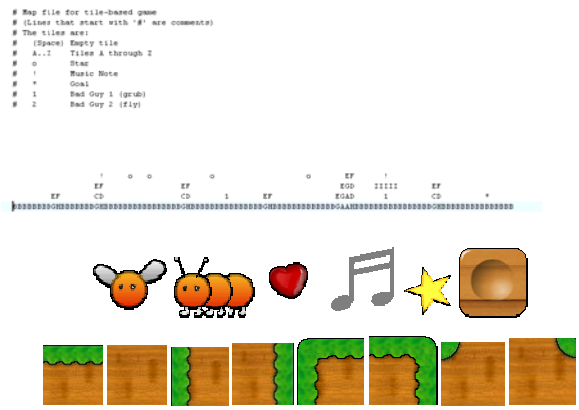


Fig. 3. Game levels design and game tiles

4.2 ColorGame

The purpose of the second application is transformation of any wall with papers in different shapes and colors into interactive [6]. The player must point the figure with a shape and color that is displayed on the screen. This is a good exercise for younger children who have just learned the shapes and colors. It can also be played in "Memory" mode that offers mental training. When making this application it was necessary to take certain approaches from computer vision[23]. Image processing in mobile phones is new and exciting field with many challenges because of limited hardware. The Hue component from the HSV color space is used for determining the color of the pointed figure. For optimization purposes only the central pixels were processed. Next challenge was the determination of the geometric shape that is being pointed by the player. First a threshold is applied to get a binary image and then the geometric properties of the region are analyzed. For determining the shape of the region we use the surface of the region - counting the pixels that compose it. The number of pixels in the area defines the shape. In the case of a square that would be from 85-100% of total area, circle - 65-85% and triangle less than 65%. One can have the same results when comparing the formulas for the area of geometric shapes within the same region.

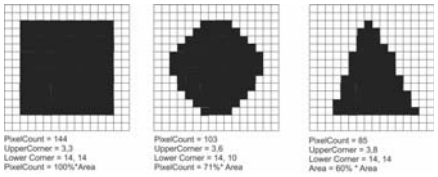


Fig. 4. Figures areas

$$P_{square} = a^2 \quad P_{circle} = a^2 * \pi / 4 \quad P_{triangle} = a^2 \sqrt{3}/4 \quad (3)$$

This method for determining the shape proved most convenient and with real-time response. For this implementation we also tested several well-known libraries like OpenCV [14], silhouette [15] and jjil [16] (John's image processing library) which didn't satisfy the real-time demands.

4.3 MapGame

This application is a combination of several new and modern technologies and approaches in developing applications for mobile phones. It consists of two parts:

- Web interface - a map where users can create games by adding points of interest and questions.
- Mobile application - application that displays games created in an Augmented reality view and guide the player.

This application is primarily designed for playing outdoors, such as amusement parks, zoos and other green areas. Web application is developed using Openstreet-Map[17] . The OpenLayers[18] JavaSctipt library was used for maps manipulation and display and the web interface of the application is made using jQuery[19] library.

The mobile application is an augmented reality browser where the user has two views:

- Forward - Augmented reality browser and markers display
- Down - map display with markers

Fig. 5. Mobile and web application

The built-in GPS and compass were used for determining the position of the user. On the market there are several augmented reality browser like Wikitude [20], Layar [21], Junaio [22], but they were not used because it was not possible to customize them for our needs. Creating the game through the web interface is simple and intuitive. First the user must choose the region in which they would like to create a game. Points are added by clicking on the map. Then through a dialogue the user sets the questions and possible answers. At the end the game should be downloaded and played on the mobile device.

Following is a table where the three games are compared according the criteria set at the beginning.

	RunGame	ColorGame	MapGame
Installation	Easy	Easy	Easy
Connecting	Wireless router needed (Android devices can't create ad-hoc connection)	No connections needed	No additional connections needed
Internet connection in play-time	Not needed	Not needed	Not needed (except when creating new games)
Possibility to create new levels	Yes	No	Yes
Intended for	Parents, teachers and instructors that would like to create new challenges. Could be played indoor on a big screen.	Indoor and outdoor on a white wall with geometrical figures.	Outdoor play in amusement parks, zoos, open areas
Types of activities that this applications motivates	Running and jumping	Aerobic activities and stretching, mental activities	Movement and coordination in space, mental activity

Fig. 6. Applications comparison table

5 Testing and discussion

In order to test the effectiveness of the developed games testing was carried out in the kindergarten "25 Maj" and elementary school "Blaze Konevski". Totaly: 30 children at the age of 5 were present during testing and 75 children at the age of 10-11 tested RunGame and ColorGame.

The game evaluation was done by the principles of the Structured Expert Evaluation Method[24]. A short list of questions was created in order to test if the game goals would be understood and can be achieved by making certain actions and whether the goal would be fun. As evaluation criteria was analyzed the opinion of children compared to the list of questions. Children said that they were able to understand the goal of each game, the actions that they were supposed to perform and they had fun playing the games. They even set their own rules like: being better than the previous player, opening a new level, discovering a new trick in the level, forgetting about the fact that they are actually exercising. This was is accordance to our assumptions about how a fun game should engage the children into exercising without them being aware of it.

Fig. 7. Children testing the applications

Additionally a challenge for the Zoo - Zoogame was set for the Mapgame. Different locations from the zoo were chosen close to the animal cages and different questions regarding the animals were set. The challenge was tested one day at the Zoo by a group of 10 children. Even though the tests were made with an early prototype user-unfriendly version of the mobile application the children quickly understood how to use it. They were able to find the answers from the panels near the animal cages and to finish the challenges successfully. This test gave us an overview of the current state of the game regarding the gameplay and the players' reactions and adoption. The children showed great interest to play the game again and learned many new facts about the animals while exercising. ZooGame won the 2nd place on a competition for an Android application organized by the Agency for Electronic communication of Macedonia where the jury was impressed by the fact that it was intended for the youngest audience.

6 Conclusion

We have developed a prototypes of a games which are entertaining, educational and motivate physical activity. Since smartphones nowadays are necessity many mobile operators have packages that offer free devices which is a precondition for the acceptance and spread of these applications by many users. People do not carry game consoles wherever they go and in case one has a smartphone, it can easily turn any wall into an interactive or any park into fun. There is no need for special game rooms or additional expenses, but the games can be played at home or outside. Children love things related to technology and this is a way to unobtrusively encourage physical activity. This system is recommended for parents or teachers that would like to spare some time for developing applications for their children's wellness.
We hope that in this era in which people get attached by modern technologies like internet, gaming, social networking and spend huge amount of time sitting in front of the screens with our gaming system we can make them use the same technologies in a way that it would make them more physically active and healthy.

7 References

1. http://www.letsmove.gov/learn-facts/epidemic-childhood-obesity
2. http://www.xbox.com
3. http://www.nintendo.com/wii
4. MarioFit: Exercise Through Mobile Entertainment - Chandrika Jayant and T. Scott Saponas, December, 2005
5. Human Pacman: a mobile, wide-area entertainment system based on physical, social, and ubiquitous computing - Adrian David Cheok,Journal: Personal and Ubiquitous Computing, Volume 8 Issue 2, May 2004, Pages 71 - 81,Springer-Verlag London, UK
6. http://en.wikipedia.org/wiki/Interactive_whiteboard
7. Developing game in Java-David Brackeen, Publication Date: August 31, 2003 | ISBN-10: 1592730051 | ISBN-13: 978-1592730056 | Edition: 1
8. Activity Recognition Using Optical Sensors on Mobile Phones - Michael Wittke, Uwe Janen, Aret Duraslan, Emre Cakar, Monika Steinberg, and Jurgen Brehm, Leibniz Universitat Hannover, Institut fur Systems Engineering, 2009
9. Single-Accelerometer-Based Daily Physical Activity Classification - Xi Long, Student Member, IEEE, Bin Yin, and Ronald M. Aarts, Fellow, IEEE 31st Annual International Conference of the IEEE EMBS Minneapolis, Minnesota, USA, September 2-6, 2009
10. Activity Recognition from Accelerometer Data - Nishkam Ravi and Nikhil Dandekar and Preetham Mysore and Michael L. Littman, Department of Computer Science, Rutgers University Piscataway, NJ 08854, 2005, American Association for Artificial Intelligence
11. Real time human activity recognition using tri-axial accelerometers - Narayanan C. Krishnan,Dirk Colbry,Colin Juillard,Sethuraman Panchanathan, Center for Cognitive and Ubiquitous Computing, Dept of Computer Science and Engineering, School of Computing and Informatics, Arizona State University, Tempe, AZ. 85281, 2006
12. Classification of basic daily movements using a triaxial accelerometer - M. J. Mathie, B.G. Celler, N.H. Lovell, A.C.F. Coster, Centre for Health Informatics, University of New South Wales, Sydney, Australia, Med Biol Eng Comput. 2004 Sep;42(5):679-87.
13. Activity Recognition using Cell Phone Accelerometers - Jennifer R. Kwapisz, Gary M. Weiss, Samuel A. Moore, Department of Computer and Information Science, Fordham University, 441 East Fordham Road, Bronx, NY 10458, ACM SIGKDD Explorations Newsletter, Volume 12 Issue 2, December 2010, Pages 74-82
14. OpenCV - www.openCV.org
15. Silhouette - http://mark.koli.ch/silhouette-project.html
16. Jon's Java Imaging Library, for mobile image processing - http://code.google.com/p/jjil/
17. http://www.openstreetmap.org/
18. http://openlayers.org/
19. www.jquery.com
20. http://www.wikitude.com/
21. http://www.layar.com/
22. http://www.junaio.com/
23. Digital Image Processing - An Algorithmic Introduction using Java, Mark J. Burge, ISBN: 978-1-84628-379-6, Textbook with 560 pages, 271 figures and 17 tables © Springer 2008
24. A Structured Expert Evaluation Method for the Evaluation of Children's Computer Games, Ester Baauw, Mathilde M. Bekker and Wolmet Barendregt, Lecture Notes in Computer Science, 2005, Volume 3585, Human-Computer Interaction - INTERACT 2005, Pages 457-469

Determination of Protein Functional Groups Using the Bond Energy Algorithm

Cvetanka Atanasova, Kire Trivodaliev, Slobodan Kalajdziski

Ss. Cyril and Methodius University, Faculty of Computer Science and Engineering, Ruger Boskovic 16, 1000, Skopje, Macedonia
cvetanka_atanasova@yahoo.com, {kire.trivodaliev, slobodan.kalajdziski}@finki.ukim.mk

Abstract. Nowadays it is possible to understand the basic components and organization of cell machinery from the network level due to increased availability of large-scale protein-protein interaction (PPI) data. Many studies have shown that clustering of the protein interaction network (PIN) can be found as an effective approach for identifying protein complexes or functional modules. A significant number of proteins in such PIN remain uncharacterized and predicting their function remains a major challenge in system biology. We propose a protein annotation method based on clustering according to Bond Energy Algorithm (BEA), which first transforms the PIN into matrix form suitable for BEA, and after generating the resulting matrix of the BEA the AUTOCLASS algorithm is performed to obtain the PIN clusters. Protein functions are assigned based on cluster information. Experiments were performed on PPI data from the bakers' yeast and since the network is noisy and still incomplete, we use pre-processing and purifying. Results reveal improvement over previous techniques and the most prominent characteristic of the BEA clustering is that the clustering result is not dependent of the initial number of clusters.

Keywords: Bond Energy Algorithm (BEA), protein interactions network, clustering methods, protein function prediction.

1 Introduction

Proteins seldom act as single isolated units to perform their functions within the cells. It has been observed that proteins involved in the same cellular processes often interact with each other [1]. Protein-protein interactions are thus fundamental to almost all-biological processes [2]. With the advancement of the high-throughput technologies, such as yeast-two-hybrid, mass spectrometry, and protein chip technologies, huge data sets of protein-protein interactions become available [3]. Such protein-protein interaction data can be naturally represented in the form of networks, which not only give us the initial global picture of protein interactions on a genomic scale but also help us to understand the basic components and organization of cell machinery from the network level.

S. Markovski, M. Gusev (Editors): ICT Innovations 2012, Web Proceedings, ISSN 1857-7288

An important challenge for system biology is to understand the relationship between the organization of a network and its function. It has been shown that clustering protein interaction networks is an effective approach to achieve this goal [4]. Clustering in protein interaction networks is to group the proteins into sets (clusters), which demonstrate greater similarity among proteins in the same cluster than in different clusters. Since biological functions can be carried out by particular groups of genes and proteins, dividing networks into naturally grouped parts (clusters or communities) is an essential way to investigate some relationships between the function and topology of networks or to reveal hidden knowledge behind them.

Classical graph-based agglomerative methods employ a variety of similarity measures between nodes to partition PPI networks, but they often result in a poor clustering arrangement that contains one or a few giant core clusters with many tiny ones [5]. To improve the clustering results, PPI networks were weighted based on topological properties such as shortest path length [6], [7], clustering coefficients [8], node degree, or the degree of experimental validity [9]. As a new type of clustering algorithms, the edge-betweenness was defined as a global measure to separate PPI networks into subgraphs in a divisive manner [10], [11], [12]. Edge-betweenness is the number of shortest paths between all pairs of nodes that run through the edge. It is able to identify biologically significant modular structures, but it requires lots of computational resources. As an approach to coordination of typical clustering algorithms, an ensemble method was proposed to combine multiple, independent clustering arrangements to deduce a single consensus cluster structure [13].

Heuristic rule-based algorithms were proposed to reveal the structure of PPI networks [14], [15]. A layered clustering algorithm was presented, which groups proteins by the similarity of their direct neighborhoods to identify locally significant proteins that links different clusters, called mediators [16]. Power graph analysis transforms biological networks into a compact, less redundant representation, exploiting the abundance of cliques and bicliques as elementary topological motifs [17]. Spectral clustering analysis, which is an appealing simple and theoretically sound method [17], [18], has hardly been studied to partition PPI networks, while it is used for detecting protein complexes [19].

The quality of the obtained clusters can be evaluated in couple of ways. One of the criteria rates the clustering as good if the proteins in a cluster are densely connected between themselves, but sparsely connected with the proteins in the rest of the network [20]. Some systems provide tools for generation of graphs with known clusters, which is modelled with the parameters of the explored network [21]. Then the clusters obtained with the clustering algorithm are compared to the known ones. The clustering method can also be evaluated by its ability to reconstruct the experimentally and biologically confirmed protein complexes or functional modules [12], [19], [20].

In this paper we set up a framework for predicting protein functional groups by using clustering in PIN. We use Bond Energy Algorithm, which first transforms the PIN into matrix form suitable for BEA, and after generating the resulting matrix of the BEA the AUTOCLASS algorithm is performed to obtain the PIN clusters. Protein

functions are assigned based on cluster information. Experiments were performed on PPI data from the bakers' yeast and since the network is noisy and still incomplete, we use pre-processing and purifying. We also performed network weighting based on the annotation correlation between nodes.

2 Research Methods

The methods for protein function prediction by clustering of PIN generally consist of three phases. The first phase is dividing the network in clusters, using its topology or some other information for the nodes or the edges, if such information is available. In this paper we use the Bond Energy Algorithm in combination with the AUTOCLASS algorithm in order to obtain the clusters of the PIN. The compactness and the characteristics of the obtained clusters are then evaluated in the second phase. From physical aspect the clusters can be assessed by the ratio of the number of edges within and between the clusters, and from biological aspect they can be assessed by the functional and biological similarities of the proteins in the clusters. This second phase is not mandatory, but might be useful because it can point out what to expect from the function prediction itself. The prediction of the protein annotations for the proteins in the clusters is the task of the third phase.

2.1 Preprocessing and Transforming the PIN Data

The protein interactions network represents an annotated graph in which the nodes are the proteins themselves, whereas the links between the proteins in the graph are the interactions among the proteins. The graph can be seen as an annotated graph since every node is assigned with one or more terms corresponding to the functions that are performed by the protein represented by that node. Because there is a symmetry in the nodes (proteins) of the graph, this means that if the protein P_1 is in an interaction with the protein P_2, than the protein P_2 will also be in an interaction with the protein P_1. If we mark the multitude of nodes with J and the multitude of connections among the nodes with V, than the notation of the graph G would be the following: $G = (J, V)$.

Of great importance for the research is the neighborhood matrix by which the graph $M = (a_{ij})_{i, j \in J}$ is represented. For the neighborhood matrix applies that $a_{ij} = 1$ if the protein i is in an interaction with the protein j, namely $a_{ij} = 0$ if the protein i is not in an interaction with the protein j. This applies for a non-weighted graph. There are algorithms according to which such non-weighted graphs are prescribed appropriate measures which represent the probabilities of a protein's interaction with the other protein and in such a case the graph converts into a weighted graph for which the formula $a_{ij} = w_{ij}$ applies.

If we mark the multitude of all the protein's functions with F, then, besides the neighborhood matrix which represents the interactions network, the annotations matrix $Z = (z_{ij})_{i \in J, j = F}$ is also important where the rows are marked with $| J |$ and the

columns are marked with | F |. In addition, $z_{ij} = 1$ if the protein i is annotated with the function j and $z_{ij} = 0$ if the protein i is not annotated with the function j.

After obtaining the matrix representation M of the graph, we can apply the Bond Energy Algorithm in order to reorder the matrix according to the affinity between the proteins from the data set. The matrix F is used during the phase of determining the functions of the proteins after the clustering is performed.

2.2 Clustering by Using the Bond Energy Algorithm

Bond Energy Algorithm (BEA) has been widely used for vertical fragmentation in distributed databases. The algorithm was originally proposed by McCormick and Hoffer and Severande [22]. BEA creates clusters by using a non-trivial similarity metric (attribute affinity measures) defined on the elements of the data set. BEA is comprised by two algorithms, the first one is used for ordering the data set to locate the most related elements close together (and to separate the unrelated elements) and the second one is used for creating the groups to determine the best cut of the ordered data set (i.e. create a cluster).

```
input: Attribute Affinity Matrix
output: CA: Clustered Affinity matrix and order list array.
begin
  [initialize; the AA matrix is created]
  CA(*, 1) ← AA(*, 1)
  CA(*, 2) ← AA(*, 2)
  index ← 3
  while index ≤ n do [choose the "best" location for profile AA_index]
  begin
    for i from 1 to index - 1 by 1 do
      calculate cont(AA_i-1, AA_index, AA_i)
    end-for
    calculate cont(AA_index-1, AA_index, AA_index+1) [boundary condition]
    loc ← placement given by maximum cont value
    for j from index to loc by -1 do [shuffle the two matrices]
      calculate CA(*, j) ← CA(*, j - 1)
    end-for
    CA(*, loc) ← AA(*, index)
    index ← index + 1
  end-while
  order the rows according to the relative ordering of the columns
end.
Note: * means for each element in the data set
```

Fig. 1. Pseudo-code of the Bond Energy Algorithm

The fundamental task in the first part of the algorithm is to find some means of grouping the attributes in a data set based on the attribute affinity values in AA (Attribute Affinity Matrix). Bond Energy Algorithm takes as input the attribute affinity matrix, permutes its rows and columns, and generates a clustered affinity matrix (CA). Once this measure is maximized the permutations are considered to be done. The pseudo-code of the BEA is presented on the Figure 1.

After obtaining the permutation of the initial matrix, we can apply the AUTOCLASS method [23], which is an unsupervised clustering method that seeks for a maximally good clustering. The structure of this classifying method is shown on Figure 2. AUTOCLASS represents a computer implementation of the Bayesian unsupervised data classification technique. By assigning real or discreet values, AUTOCLASS determines the probable number of classes in which the data and probabilities with which a given object belongs to the assigned class are distributed. As can be seen from the Figure 2, the AUTOCLASS algorithm iterates the number of clusters until it converges. For each element of the dataset, it computes the probabilities for belonging to some of the classes, and in each iteration it reestimates these values according to the current cluster distribution. One of the best features of this method is that the number of classes is automatically determined and there is no need of any previous setting of the number of resulting clusters.

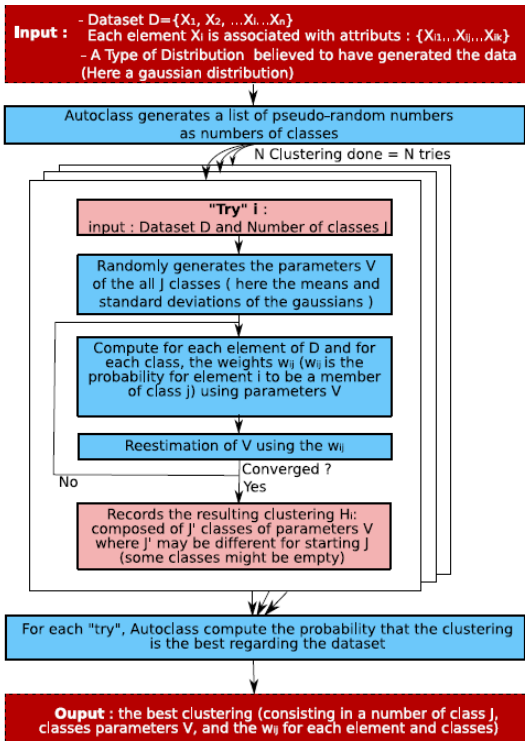


Fig. 2. AUTOCLASS Structure

After applying the AUTOCLASS method, the resulting clusters can be further evaluated in order to obtain the efficiency parameters of the clustering method.

2.3 Functional Annotation Using Clusters

As we previously mentioned, the whole process consists of several phases. As an underlying data set is a non-weighted protein interactions network.

The first phase includes preprocessing of the data in order to get them in a format which is compatible with the BEA Algorithm. The second phase includes execution of the BEA Algorithm, which provides an ordered matrix which is an input to the AUTOCLASS method. This phase provides the data set clusters and performs an evaluation of the features and compatibility of the clusters found, from a biological aspect, namely, how similar the proteins are within the cluster according to their functional and biological features. Thus, a global functional mark is received for the cluster. This phase is very useful because it indicates what can be expected for the results from the prediction. This phase also calculates the enrichment ratio of the cluster by a certain function, a specification which directly indicates the possible function of the interrogative protein. The prediction of the functions of the non-annotated protein is elaborated in the third phase. Every function receives an appropriate rating depending on the frequency of appearances within the cluster *K* which contains the interrogative protein. The ratings of every function are received by the formula (1) and are then normalized within range from 0 to 1. All the proteins from the protein interaction network go through the process as non-annotated proteins so that they can get their functions according to the methods elaborated in this research.

$$f(j)_{j\in F} = \sum_{i\in K} z_{ij} \tag{1}$$

where F is the set of functions present in the cluster K, and

$$z_{ij} = \begin{cases} 1, \text{if protein i from K has function j from F} \\ 0, \text{otherwise} \end{cases} \tag{2}$$

The last phase includes evaluation of the results.

3 Results and Discussion

High-throughput techniques are prone to detecting many false positive interactions, leading to a lot of noise and non-existing interactions in the databases. Furthermore, some of the databases are supplemented with interactions computationally derived with a method for protein interaction prediction, adding additional noise to the databases. Therefore, none of the available databases are perfectly reliable and the choice of a suitable database should be made very carefully.

For the needs of this paper the PIN data are compiled, pre-processed and purified from a number of established datasets, like: DIP, MIPS, MINT, BIND and BioGRID.

The functional annotations of the proteins were taken from the SGD database [24]. It is important to note that the annotations are unified with Gene Ontology (GO) terminology [25].

The data for protein annotations are not used raw, but are preprocessed and purified. First, the trivial functional GO annotations, like 'unknown cellular compartment', 'unknown molecular function' and 'unknown biological process' are erased. Then, additional annotations are calculated for each protein by the policy of transitive closure derived from the GO. The extremely frequent functional labels (appearing as annotations to more than 300 proteins) are also excluded, because they are very general and do not carry significant information.

The multitude used for this experiment is highly confidential and is consisted of 2502 proteins, among witch 12708 interactions are noted, with a total of 888 annotated functional marks. The protein interactions network does not represent a linked graph, but is consisted of several components, the biggest of which contains 2146 proteins. This multitude of interactions of the yeast is used for testing and evaluation of the methods for prediction of protein function subjected in this paper. Each protein in the PIN is streamed through the prediction process one at a time as a query protein. The query protein is considered unannotated, that is we employ the leave-one out method. Each of the algorithms works in a fashion that ranks the "proximity" of the possible functions to the query protein. The ranks are scaled between 0 and 1. The query protein is annotated with all functions that have rank above a previously determined threshold ω. For example, for $\omega = 0$, the query protein is assigned with all the function present in its cluster. We change the threshold with step 0.1 and compute numbers of true-positives (TP), true-negatives (TN), false-positives (FP) and false-negatives (FN). For a single query protein we consider the TP to be the number of correctly predicted functions, and for the whole PIN and a given value of ω the TP number is the total sum of all single protein TPs.

To compare performance between different algorithms we use standard measures as sensitivity and specificity (3).

$$sensitivity = \frac{TP}{TP + FN} \qquad specificity = \frac{TN}{TN + FP} \tag{3}$$

We plot the values we compute for the sensitivity and specificity using a ROC curve (Receiver Operating Curve). The *x*-axe corresponds to the false positive rate, which is the number of false predictions that a wrong function is assigned to a single protein, scaled by the total number of functions that do not belong to that particular protein. This rate is calculated with (4).

$$fpr = \frac{FP}{FP + TN} = 1 - specificity \tag{4}$$

The *y*-axe corresponds to the rate of true predictions that is the sensitivity. At last we use the AUC (Area Under the ROC Curve) measure as a numeric evaluator of the ROC curve. The AUC is a number that is equal to the area under the curve and its

value should be above 0.5, which is the value that we get if the prediction process was random. The closer the value of AUC to 1, the better is the prediction method. The experimental results are shown on the Table 1, and on the Figure 3.

Table 1. Experimental Results. The values of the sensitivity and specificity are obtained for different values of the parameter ω.

ω =	**0**	**0,1**	**0,2**	**0,3**	**0,4**	**0,5**	**0,6**	**0,7**	**0,8**	**0,9**
sensitivity	0,82	0,82	0,77	0,71	0,64	0,57	0,49	0,42	0,34	0,26
specificity	0,35	0,35	0,20	0,11	0,07	0,04	0,03	0,02	0,01	0,01
AUC					**0.8**					

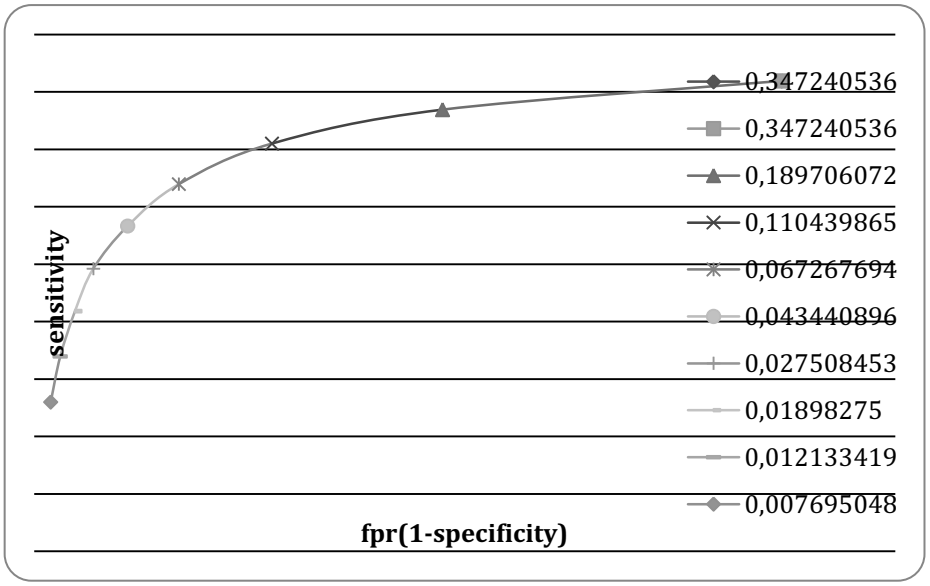


Fig. 3. The ROC curve for the obtained experimental results shown on table 1.

From the presented results, it can be noted that the sensitivity is highest in the first column reaching above 82%, which happens when the functions are falsely predicted with 35%. By increasing of ω for 0,2% it can be noted that the sensitivity decreases to 77% which, in fact, is not that great a margin, but the number of falsely predicted functions is decreased by 15%. Figure 3 shows the visual results of this experiment. It can be also seen from the Figure 3 that sensitivity values are 20% only when 1% of the functions are wrongly predicted. The value of AUC on this classificatory is 0,80.

4 Conclusion and Future Work

This paper exploits the possibilities for applying the Bond Energy Algorithm for clustering and detecting the functional modules and predicting protein functions from PIN. The method was tested over one of the richest interactomes: the interactome of

the baker's yeast. Database of Interacting Proteins (DIP) which was used, aimed to integrate the various experimental results from the biochemical analyses of the protein interactions. The protein interactions network is a non-weighted one, which means that we have information whether one protein is in an interaction with another one, and, as such, was used for prediction of the protein functions. We have used the matrix representation of the PIN and applied the BEA and AUTOCLASS in order to obtain the matrix permutation according to the affinity between the proteins. Due to the fact that the PIN data contain a lot of false positive interactions, the dataset needed to be preprocessed and purified prior to the functional annotation. The results show that our algorithm achieves high sensitivity and small false positive rate on PIN graphs and it has a high AUC value.

Our future work will be concentrated on the possibilities for adding weights to non-weighted graphs as well as to make a prediction of the protein functions on a weighted graph. It is also useful to add the additional settings within the Bond Energy Algorithm so that more data can be separated, coming from the weighted protein interactions network. The AUTOCLASS method which determines the clusters can also be modified i.e. improved by trying some other heuristics during its iteration process.

References

1. von Mering, C., Krause, R., Sne, B.: Comparative assessment of large-scale data sets of protein-protein interactions. Nature 417, (2002) 68-87
2. Hakes, L., Lovell, S. C., Oliver, S. G.: Specificity in protein interactions and its relationship with sequence diversity and coevolution. PNAS 104, (2007) 19
3. Harwell, L. H., Hopfield, J. J., Leibler, S., Murray, A. W.: From molecular to modular cell biology. Nature 402, (1999) c47- c52
4. Brohée, S., van Helden, J.: Evaluation of clustering algorithms for protein-protein interaction networks. BMC Bioinformatics 7, (2006) 48
5. Barabasi, L., Oltvai, Z. N.: Network biology: understanding the cell's functional organization. Nat Rev Genet 5, (2004) 101-113
6. Arnau, V., Mars, S., Marin, I.: Iterative cluster analysis of protein interaction data. Bioinformatics 21, (2005) 364- 378
7. Rives, W., Galitski, T.: Modular organization of cellular networks. PNAS 100, (2003) 1128-1133
8. Friedel, C., Zimmer, R.: Inferring topology from clustering coefficients in protein-protein interaction networks. BMC Bioinformatics 7, (2006) 519
9. Pereira-Leal, J. B., Enright, A. J., Ouzounis, C. A.: Detection of functional modules from protein interaction networks. Proteins 54, (2004) 49-57
10. Dunn, R., Dudbridge, F., Sanderson, C. M.: The use of edge-betweenness clustering to investigate biological function in PINs. BMC Bioinformatics 6, (2005) 39
11. Luo, F., Yang, Y., Chen, C. F., Chang, R., Zhou, J.: Modular organization of protein interaction networks. Bioinformatics 23, (2007) 207–214
12. Newman, M. E., Girvan, M.: Finding and evaluating community structure in networks. Phys Rev E Stat Nonlin Soft Matter Phys 69, (2004) 026113
13. Asur, S., Ucar, D., Parthasarathy, S.: An ensemble framework for clustering protein-protein interaction networks. Bioinformatics 23, (2007) i29–40

14. Gagneur, J., Krause, R., Bouwmeester, T., Casari, G.: Modular decomposition of protein-protein interaction networks. Genome Biol 5, (2004) R57
15. Morrison, J. L., Breitling, R., Higham, D. J., Gilbert, D. R.: A lock-and-key model for protein-protein interactions. Bioinformatics 22, (2006) 2012-2019
16. Andreopoulos, B., An, A., Wang, X., Faloutsos, M., Schroeder, M.: Clustering by common friends finds locally significant proteins mediating modules. Bioinformatics 23, (2007) 1124-1131
17. Royer, L., Reimann, M., Andreopoulos, B., Schroeder, M.: Unraveling protein networks with power graph analysis. PLoS Comput Biol 4, (2008) e1000108
18. Spirin, V., Mirny, L. A.: Protein complexes and functional modules in molecular networks. PNAS 100, (2003) 21
19. Belkin, M., Niyogi, P.: Laplacian Eigenmaps for Dimensionality Reduction and Data Representation. Neural Computation 15, (2003) 1373–1396
20. Chen, J., Yuan, B.: Detecting Functional Modules in the Yeast Protein-Protein Interaction Network. Bioinformatics 18, (2006) 22
21. Lancichinetti, A., Fortunato, S., Radicchi, F.: Benchmark Graphs for testing Community Detection Algorithms. Physical Review E78, (2008) 046110
22. Arabie, P., Hubert, L.: The Bond Energy Algorithm Revisited. IEEE Transaction on Systems, Man and Cybernetics, (1990)
23. Achcar, F., Camadro, J. M., Mestivier, D.: AutoClass@IJM: a powerful tool for Bayesian classification of heterogeneous data in biology. Nucleic Acids Research 37, (2009)
24. Dwight, S., Harris, M., Dolinski, K., Ball, C., Unkley, G. B., Christie, K., Fisk, D., Issel-Tarver, L., Schroeder, M., Sherlock, G., Sethuraman, A., Weng, S., Botstein, D., Cherry, J. M.: Saccharomyces Genome Database (SGD) provides secondary gene annotation using Gene Ontology (GO). Nucleic Acids Research 30, (2002)
25. The gene ontology consortium: Gene ontology: Tool for the unification of biology. Nature Genetics 25, (2000)

Application of BCI Technology for Color Prediction Using Brainwaves

Martin Angelovski[1], Petre Lameski[1], Eftim Zdravevski[1], Andrea Kulakov[1]
[1] Faculty of Computer Science and Engineering
mangelovski@gmail.com, {petre.lameski, eftim.zdravevski, andrea.kulakov}@finki.ukim.mk

Abstract. In this paper we present the results of the research regarding color recognition using brain-waves read from a consumer grade non-intrusive Brain Computer Interface device. The brainwaves obtained from the device are pre-processed and filtered using different algorithms and then several machine learning techniques are applied for prediction of three colors. At the end, the obtained results are presented and discussed.

Keywords: brain computer interface, brainwaves, color prediction

1 Introduction

At the beginning of the computer era people did not pay much attention on easy interaction to computer systems with humans. It was assumed that they will only be used from highly qualified users. However, computers are nowadays wide-spread and used in every aspect of people's lives. One of the many ways to communicate with a computer that is still under development is the computer interface with brain waves. Brain computer interface (BCI) is a system that interprets the thoughts of man to make commands that control a computer or other device. Human-computer interaction is bidirectional: many systems adapt to the user, which in turn should constantly adapt to new solutions of the system. The task of this feedback loop adaptation between system and user is important in BCI systems that seek to approximate the neuronal function [2].

There are several techniques for reading the electrical signals produced by the brain in humans. BCI can be divided the 3 groups: totally invasive, partially invasive and noninvasive. The first two demand complex procedures for setting up on the human body, while the third one is simpler for set-up, but is significantly more inaccurate. Obviously, the consumer grade BCI hardware belongs to the third group.

Much research is done with more complex noninvasive hardware (fMRI, MEG, EEG etc.) and many promising results are obtained in this area [1]. Consumer grade hardware however is rarely researched and its capabilities are limited to the manufacturers' software products that are usually limited to different kinds of games.

In this paper we are using a consumer grade BCI device that is easy to obtain and has a low price. The goal of this paper is to detect the capabilities of the device to be used as a control device for the color selection task. The basic idea is to train a

S. Markovski, M. Gusev (Editors): ICT Innovations 2012, Web Proceedings, ISSN 1857-7288

decision making model using several machine learning algorithms on real data. After the model is obtained, the model would be used to detect what the user is thinking about. For the purposes of this paper, the training data is obtained by measuring the brainwave activity while the user watches a certain color and then the trained model is used by the user for color selection. The bidirectional training of the BCI devices is not taken into consideration in this paper. Only the computer model of the way the human thought is generated. The human is not trained for recognizing colors with the device.

2 System Architecture

For the purposes of the experiment, we used a cheap and widely available device. The device used in this experiment is Mindwave from the company Neurosky [3]. This device measures the standardized EEG brain waves but it gives special measurements for blink detection, attention and meditation levels of the user. Although it is relatively new, there are already lot of programs and games that mostly use the specialized readings, and much less applications that use the whole spectrum of EEG waves. Most of the applications are games, but there are several that allow visual display of the readings from the device. The device is one of the best selling commercial EEG devices and is financially much more accessible than other devices on the market. Some primary schools in the United States have implemented it for various purposes. A lot of the pupils that used this device while studying and problem solving have shown better results by learning how to increase their concentration and attention. The Mindwave is also used in training for sports that need calmness and concentration like archery and billiard. There are several applications for educational purposes from which the ones that help with math and problem solving are used the most.

Despite the above mentioned advantages, this device has also some disadvantages. Its biggest disadvantage, that is also common with a lot of other consumer based BCI devices, is that because of not using conductive gel there is a significant amount of noise. Furthermore, this device has only 2 sensors for measuring brainwaves.

For the purposes of this experiment the Neurosky Mindwave SDK package is used [4]. With the help of this software development kit a program in Java was developed for connecting with the device, showing colors to the user and collecting the data from the Mindwave. The software developed had additional functionalities that are not in the scope of this paper. Fugure 1 shows the Mindwave device used for the purposes of our experiments.

Fig.1. Mindwave device

The used SDK provides access to the Mindwave sensor measurements as numeric values and several aggregated values. These values are high and low Alpha, Beta and Gamma waves, Delta, Theta, Attention and Meditation, in addition there is the blinking strength. The values are obtained with frequency of 1 Hz. This means that a new measurement is available every second. Figure 2 shows the process for obtaining the model for color recognition in this paper.

Fig.2. Model training workflow

The device is used for obtaining the data for each color. The data is obtained from the brainwaves of a person that looks at a certain color. For this experiment we used three colors – Red, Blue and Yellow. During the training phase the raw data is stored in a file. Afterwards with the addition of the header it is converted in the .ariff file format and inserted into the WEKA program [5]. WEKA (Waikato Environment for Knowledge Analysis) is popular program for machine learning written in Java developed by the University of Waikato. The data that is inserted in the system consists of several parameters: Attention, Meditation, Delta, Theta, Low Alpha, High Alpha, Low Beta, High Beta, Low Gama and High Gama. The colors used were Red, Blue and Yellow. WEKA software is used in both the preprocessing and the model training phase.

For the preprocessing phase, the attribute selection and the wavelet transformation were used as provided by WEKA. After that, models were built by several machine learning algorithms:

- Bayes Network
- Multilayer perceptron
- Decision trees
    - J48 (pruned and unpruned)
    - Random Forest

- Hybrid Algorithm: Rotation Forest.

The models can be used in the decision making process as described in the workflow on Figure 3:

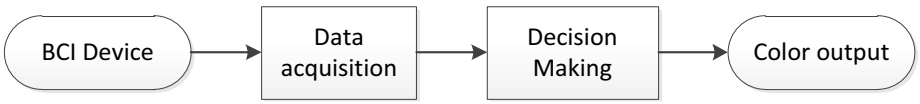


Fig.3. Using BCI for color selection

The built models can be used for selecting colors trough the BCI device or for other tasks (provided that a model is built for each specific task that the BCI device is intended for). As stated before, all of these algorithms were implemented in the WEKA software.

3 Experimental Results

The comparison of each approach was done by comparing the performance of the trained models with 10 fold cross validation.

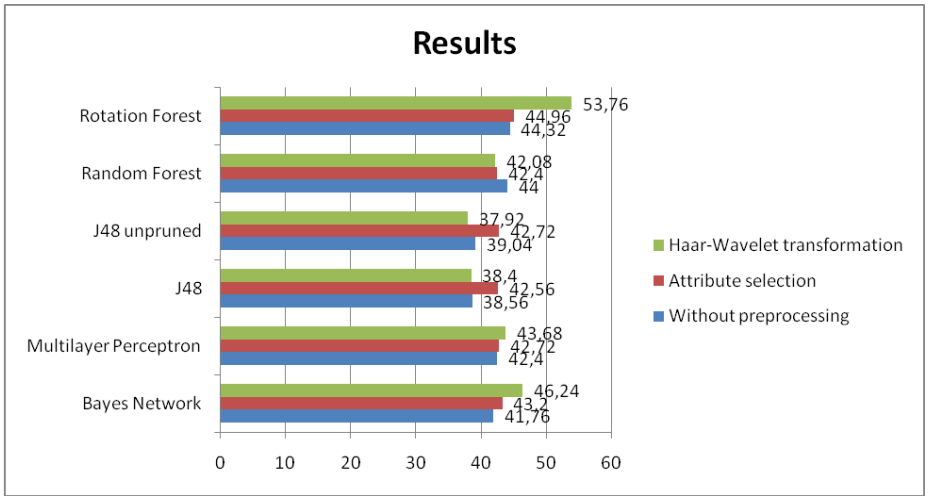


Fig.4. Precision results in %

The training was done with total of 675 samples. The obtained results for different preprocessing methods are given on Figure 4.

While building of the models the Multilayer perceptron requires noticeably more time for training then the other classifiers and gives almost the same results. The best

result for the Multilayer perceptron is the precision of 43,68% with preprocessing of the data with Wavelet function.

With the training of the Bayes classifier gives better precision than the multilayer perceptron when the data is preprocessed. The best result was obtained with wavelet filtering of the data with a precision of 46,24%. This classifier gave good classification precision for the yellow and red color. The precision for the blue color classification was much worse and contributed towards the lower total performance.

The worst results for the decision tree algorithms were obtained with the J48 classifier although they gave the fastest training time. Un pruned J48 tree with attribute selection showed a precision of 42,72% which is almost as good as the neural network, with the difference that the Decision tree is trained and executed much faster than the Multilayer perceptron.

The Random forest, gives a precision of 44% for the whole data set. Here the success rate with the red color is only 28% while with the other 2 colors it is substantially bigger.

The best precision with all learning algorithms is achieved with the Rotation forest. The best precision obtained with this algorithm was 53,76%. The algorithm is described in [6].

From the collected results it can be seen that the best classifier for our data set from the selected classifiers is the Rotational forest. The analysis showed that this algorithm also gave the best performance in terms of the ROC area. The ROC area results for each algorithm with each preprocessing are given in Figure 5.

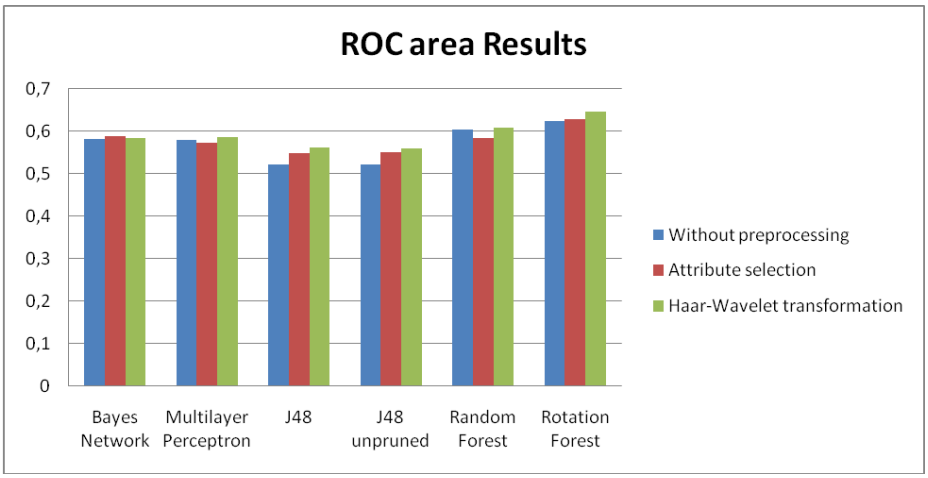


Fig.5. Results given in ROC area for each training algorithm

As it can be seen in both Figure 4 and Figure 5, the best results are obtained with the use of the Haar-Wavelet transformation of the dataset which is expected since the dataset is consisted of series of data. We must note that the results are influenced by

the significant noise that the used device introduces in its measurements. The Mindwave is reported to have some noise canceling techniques but due to the sensitivity of the sensor it is hard to implement a more successful method to cancel out the noise coming mostly from muscle movements and electronic devices in the vicinity.

4 Conclusion and Future Work

BCI has great potential, both in science and medicine and in the entertainment industry. Despite the current shortcomings of commercially available BCI devices, they can be used for entertainment and research purposes. It is expected in the future this technology would become greatly advanced and enable more reliable reading of brain waves.

For this paper a computer program was implemented for reading brain waves when visual stimuli were shown. As already mentioned, our main goal was to use the same training process with minimal modifications with other stimulus, like paintings or music. Also small modifications are needed if we want to use another similar BCI device, or want to process data using statistical and other classification techniques "on the fly".

The research on the same set of data gave best results using a Rotational Forest where the data were previously processed by wavelet transformation. The obtained results have shown a precision of around 40% with the best accuracy of about 53%. This precision is a promising result, because it is a lot better than randomly guessing the color – 33% accuracy, but is far from any practical applications.

In future the research could lead in two directions. The first is the use of other techniques like transformations of attributes and their classification, use of time series, temporal analysis, Fourier transformation and some of the newer algorithms for classification to obtain other parameters and information on the measurements.

The second direction is to use more advanced but more expensive equipment for which would have more sensors and less noise. As it is shown in the results, the used algorithms didn't differ much in performance, thus we can expect a better result from less noisy equipment.

We must point out, however, that brain wave interface devices also require some learning on the side of the user. So in future research we plan to include models that would learn from both the data obtained from the device and the feedback from the human. In our research case the human was just seeing the color which doesn't guarantee that the brainwaves are related solely to that visual stimulus. Other distractions must be taken into consideration and somehow included in the model.

References

1. Allison B. Z., Wolpaw E. W., Wolpaw J. R. Brain–computer interface systems: progress and prospects. Expert Rev. Med. Devices 4, 2007, 463–474. doi: 10.1586/17434440.4.4.463.
2. B. Graimann et al. (eds.), *Brain–Computer Interfaces*, The Frontiers Collection. DOI 10.1007/978-3-642-02091-9_1
3. Neurosky Official Website, http://www.neurosky.com/ (Retrieved at 20.05.2012).
4. Neurosky Development website, http://developer.neurosky.com/ (Retrieved at 20.05.2012).
5. Weka3 Official Website, http://www.cs.waikato.ac.nz/ml/weka/ (Retrieved at 21.05.2012)
6. Juan J. Rodriguez, Ludmila I. Kuncheva, Carlos J. Alonso. Rotation Forest: A New Classifier Ensemble Method, IEEE Transactions on Pattern Analysis and Machine Intelligence October 2006 (vol. 28 no. 10) pp. 1619-1630

S. Markovski, M. Gusev (Editors): ICT Innovations 2012, Web Proceedings, ISSN 1857-7288

WORLD (World Bank Random Linked Data)

Darko Bozinoski[1] and Solza Grceva[1]

[1] Faculty for ICT, FON University,
bul. Vojvodina b.b. 1000 Skopje, R. Macedonia
darko.bozinoski@fon.mk
solza.grceva@fon.edu.mk

Abstract. World Organizations all around provide access to variety sets of different data listing available datasets, tables, reports, and other resources making it available for the wider audience. A smart interface is needed to use, combine and present relevant information that will provide the best usability of the opened data. Developing a single solution that will work across different sectors using the opened data from those World Organizations will have a great impact on different sectors in the future, especially in education. Our solution WORLD (WORLD RANDOM LINKED DATA) that was awarded by the World Bank in 2011 is a cross-sector prototype of a web based service which provides information using the open data of the World Bank and generates random information in the form of sentences with a combination of pictures, videos and maps. Using the World Bank database as well as the United Nations Human Development Index, the service generates intelligent and relevant information, adapting to different sectors, like education, agriculture, poverty, etc. The current solution was tested among the students in one of its practical implementation of this web service: the Moodle e-learning system. It is a Moodle block that uses the data from the web service from areas that are within the scope of Millennium Development Goals. Analyzing the negative and positive aspects of the WORLD Block we've come to the conclusion that this solution presents a good approach for displaying a lot of new information from different areas of interest in a small section. Further development will include refining of the structure in the Block, expanding the media types of generated data from the web service like games, flash, custom maps etc. The final goal is to make a generator of intelligent information that will: personalize the information, a service fully adaptable to different sectors, provide valuable information and positive user experience, be used by any web application and provide efficiency and effectiveness in gathering and presenting the information using the open databases.

Keywords: open data, web service, web interactivity, World Bank open data

S. Markovski, M. Gusev (Editors): ICT Innovations 2012, Web Proceedings, ISSN 1857-7288

1 Introduction

World Organizations all around the world provide free access to a variety of data organized in datasets, tables, reports, and other resources making it available for the wider audience. Designing and developing a smart interface which combines and presents relevant information providing the best usability of the opened data is strongly needed. Single solution that will work across different sectors using the opened data from those World Organizations will have a great impact on different sectors in the future, especially in education.

The purpose of this paper is to present our solution WORLD and the results we obtained through testing and analysis on the solution. WORLD (WORLD RANDOM LINKED DATA) is a cross-sector prototype of a web based service which provides intelligent information using the open data of World Bank and generates random information in the form of sentences with a combination of pictures, videos and maps. Using the World Bank database as well as the United Nations Human Development Index, the service generates intelligent and relevant information as a step towards reaching the Millennium Development Goals and is adaptable to different sectors where these data can be implemented. It could be applied in education (which is now a case), in water supplies, in poverty issues, in state economies, in other words with some adjustments, actually in every field where the world organizations have their accurate and unique data. We designed this solution in order to be more accessible and available for the wider public and to be used by any application on the web like various e-learning systems, Non-Governmental Web sites, Facebook applications and many others.

2 Project architecture

The project contains the following parts:

- Web service
- Database
- Public web site
- Moodle block

Our web service is based on .NET technology and it is written in C# programming language. It has only one web method, called “worldservice”. This method takes six parameters from a query string as an input, and, depending on the input, it returns an XML file as a response. The six input variables are: language, field (area of interest, example: education), country (data for desired country, or all countries), media (media type, images, video) and resolution (width and height). Depending on the resolution, the web service generates appropriate videos and images.

S. Markovski, M. Gusev (Editors): ICT Innovations 2012, Web Proceedings, ISSN 1857-7288

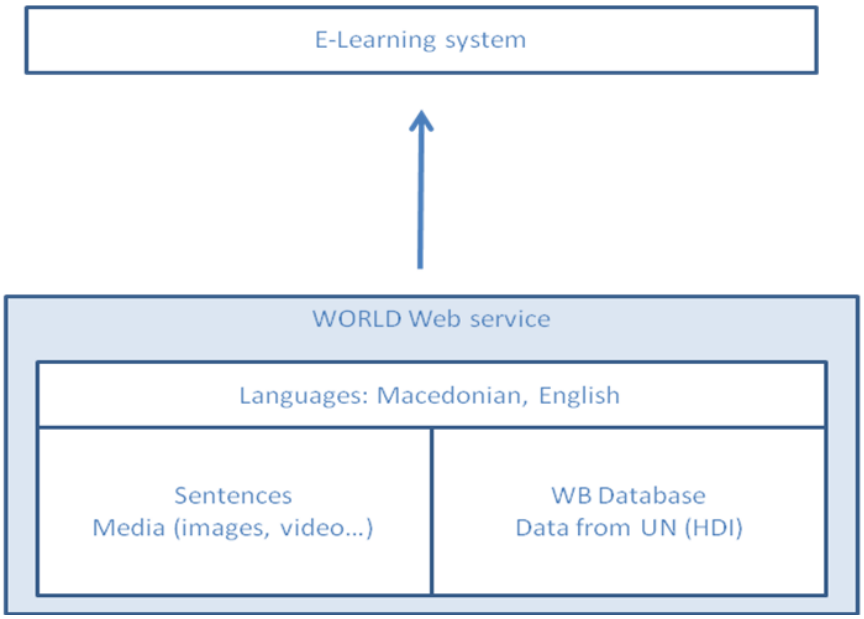


Figure1. Current architecture of the project

These parameters define the type and the content of the information that will be generated by the web service. Depending on these input parameters the web service checks what's needed to be extracted from the database and determines the display language, the area of interest, the targeted country , the type of media and resolution (there are two optional parameters, "width" and "height" and they are used to define the desired resolution of the image or video) . According to the type of media, the web service always generates a sentence. At this point, the web service user can choose what other type of media to be represented (picture, video or map).Then a connection to the database is made and the proper data is extracted. After all this, the service puts the random generated data in proper tags in XML format that is being returned as a response.

For example, if we want to display information in the field of education, for all countries, in English, and the desired media is text, we send the following query string request to the web service:

```
http://world.fon.edu.mk/world.asmx/worldservice?ln=en&f
ield=1&cnt=0&media=0&width=0&height=0
```

The service reads the input data and returns a XML response in the following format:

```
<?xml version="1.0" encoding="utf-8"?>
<world xmlns:xsi="http://www.w3.org/2001/XMLSchema-
instance" xmlns:xsd="http://www.w3.org/2001/XMLSchema"
xmlns="http://world.fon.edu.mk">
<ti-
tle>&lt;ahref='http://world.fon.edu.mk'&gt;WORLD&lt;/a&gt
;</title>
<content>&lt;p&gt;The youth literacy rate for the male
population in West Bank and Gaza, in the year 2006, was
99,053%. In comparison, the youth males literacy rate in
the year 2007, was 99,147 %.<content>
<footer>&lt;hr /&gt;&lt;p style='font-size:10px;line-
height:12px'&gt;The service has been called 153 times
from this site. Find more information &lt;a
href='http://world.fon.edu.mk'&gt;here&lt;/a&gt;.&lt;/p&g
t;</footer> </world>
```

3 Moodle Block

Practical implementation of this web service we have seen in the possibility this content to be displayed on the most used e-learning system in the world - Moodle .
This system is used by 50.000 universities around the world, with currently 37 million subscribed users. The idea is implementation of Moodle block that contains data from our web services from areas that are within the scope of MDG (Millennium Development Goals). The block in the background calls our web service which returns content that is displayed in the block. The institutions which have this e-learning system installed are able to install and use our block by choosing different service parameters. Then, on their home (or course) page is shown random content and data from the database of the World Bank, which is generated by our web service. These data inform the (mostly) students and other users of the system about the current situation of these areas, showing data in form of sentences, images or video, depending on the choice of the institution.

4 Experiment

The current status of our solution was tested at the Laboratory for Human Computer Interaction and Inclusive Design at Cyprus University of Technology as part of a Short Term Scientific Mission through the COST Project in one of its practical implementations: the Moodle e-learning system. Throughout this experiment two objectives were taken into consideration. The first objective was testing the location of the WORLD Block in a Moodle course. The results showed where the Block drew most attention of the users regarding the type of information displayed and which location

should be considered as a best possible location for displaying the WORLD Block. The second objective was to test the usability and learnability of the structure of the information provided in the Block. The results showed the participants attitudes about the Block and their negative and positive impressions about it.

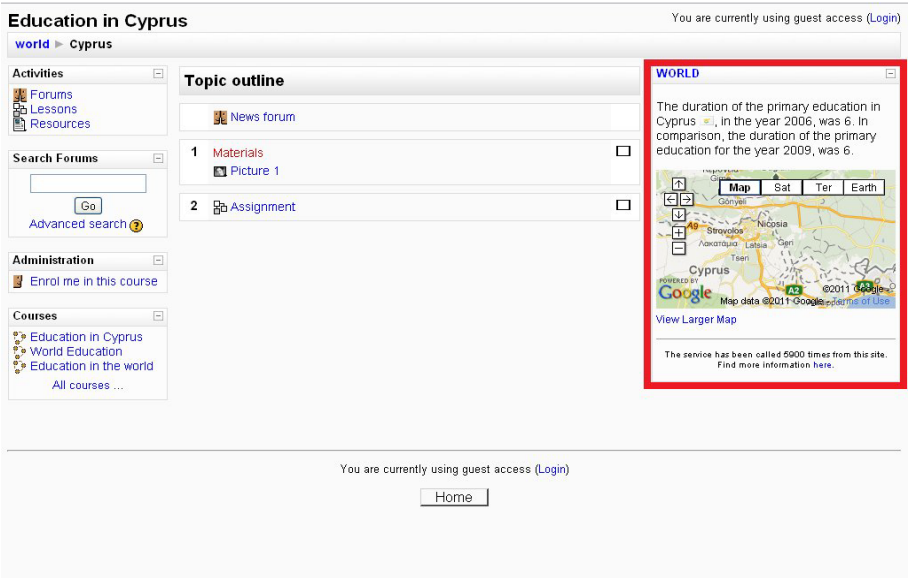
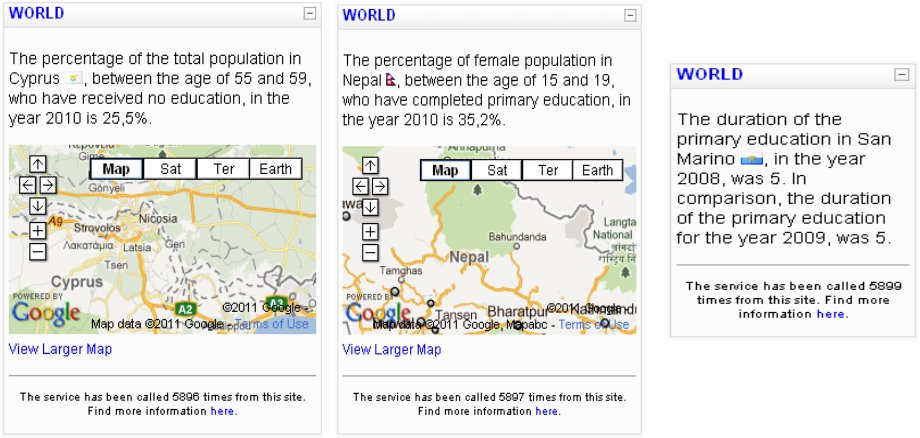


Figure2. Example of using a WORLD Block in a Moodle course
(The WORLD Block is outlined with a red color)

This experiment was structured to have several steps. First, three Moodle courses were created in the Moodle e-learning system. In all three the WORLD Block was installed in the right hand side section of the course screen. The WORLD Block displayed information related to education. We set up different features for the WORLD Block on every course. The first course was named Education in Cyprus. In this course our Block was set up to display information related to Cyprus education in a form of sentences and/or maps according to the type of the sentence. In the second course named WORLD Education the WORLD Block was displaying information for all world countries in the form of sentences and/or maps. And, in the third course information for all countries in the world were presented in a form of sentences without any media types.

The next step was setting up an experiment with the eye tracker (a RED 500 SMI eye tracker). For that purpose the Experiment Center TM 2.5 was used. This software program enabled to prepare and execute gaze tracking experiments. It is complemented by SMI iView XTM for gaze tracking data acquisition and SMI BeGazeTM 2.5 for gaze tracking data analysis. Experiment Center TM 2.5 was running on a SMI laptop

and was connected to the iView X system. The iView X system in turn operated an attached gaze tracking device.



a) b) c)

Figure3. WORLD Block in the course: a) Education in Cyprus,
b) World Education,
c) Education in the world

The experiment was conducted with twelve test subjects whose gaze positions and actions were monitored. At the beginning of every session we entered ID information for every participant and verified the calibration necessary to adapt the iView X eye tracking software to the participant's eye characteristics. Then, every participant entered one of the Moodle courses with a quest account. About 15 seconds the test subjects were not instructed with any command. Then after that time period, every participant was instructed to focus on the right hand side section of the website where the WORLD Block was installed. After reading the information in the WORLD Block the participants were asked to do an assignment in the course. The assignment consisted of answering a question related to the information that was presented in the WORLD Block. Every course was visited individually by 4 participants. In addition, as a further contribution to the subject, the participants were asked to fill a questionnaire. The main goal behind designing a questionnaire was providing an easy way for analyzing the attitude that the participants had about this Block.

5 Methods used for the analysis

After finishing the experiment, the phase of analyzing results begun. In order to see all the negative and positive aspects of this solution, we used three different methods for analysis: gaze tracking analysis, moodle block analysis and questionnaire analysis. To analyze gaze tracking data we used SMI BeGazeTM 2.5 software. SMI Be-

GazeTM 2.5. is a behavioral and gaze analysis software for eye tracking data. It displays, analyses and visualizes various kind of stimuli (pictures, web pages etc.) with integrated filter functions that allow analyzing subgroups of participants. It provides full spectrum of visualizations like:

- Gaze plots (scan path, bee swarm, gaze replay)
- Attention maps (focus map, heat map)
- Real time statistics (key performance indicators, gridded AOIs)

These are powerful methods for analyses. For better analysis, we divided every course in three areas of interest. The first one was the Navigation section in the left hand side of the course that helps users to navigate through the moodle e-learning system. The second was the Main section in the middle of the course where you can find the lectures, assignments and materials for the students. And the third section was the WORLD Block positioned on the right hand side of the course. For analyzing the assignment in the Moodle course the Reports section in the Moodle e-learning system was used where a detailed statistics for every user answer was made. Using a Google document form for the questionnaire analysis provided us with a good structure for analyzing the participants' attitude. We've created the questionnaire as a Google document for its simple and effective structure. The structure of the questionnaire was organized as follows: two required text questions, followed by twelve scaled questions (from 1 to 7) and two text questions that weren't required. The aim of the scaled questions was to determine respondents' attitudes about the features of the WORLD Block. If they really liked some of the WORLD Block features they selected maximum - 7 and if they disliked some of the features they selected 1.

6 Results

A large volume of data was collected during the experiment regarding several aspects of the WORLD Block. The experiment went as expected, without error or loosing data. The results were divided into three main parts according to the methods used for analysis. In the first part we displayed the results from the gaze tracking analysis. In this part, the results for each course were split into two time intervals. The first interval was when the participants were not instructed to do anything in the course and the second interval was when the participants were instructed to read and look into the information presented in the WORLD Block. Every participant saw different information in the Block, because the Block displays new information every time a course is visited.

The second part represented the results from the assignment the participants made in the Moodle course and the third section presented the results from the questionnaire. In this section we will present only the summary of the overall results for the three different courses that the participants visited without any further details.

Comparing the overall results in all courses we've noticed that the WORLD Block section is the second area of interest. Also, we noticed that in the course World Education there was lower visual activity in the WORLD Block section comparing to the other courses: Education in Cyprus and Education in the world where the results indicated a bigger level of visual activity in the first interval of participants' sessions.

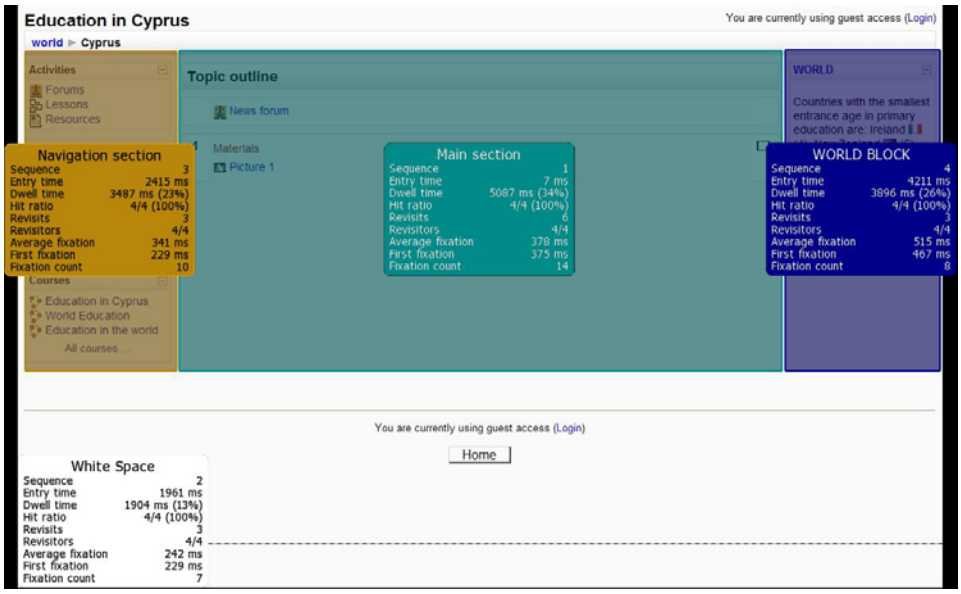


Figure4. Key Performance Indicator of all participants that visited the course Education in Cyprus in the first interval of their session

Analyzing several important statistical indicators associated to each Area of Interest (AOI) in the first interval when the participants were not instructed to do anything in the course enabled a sound comparison of different statistical information for the WORLD Block in the three different courses. According the entrance time, which represents the average duration for first fixation into specific AOI, the course WORLD Education (3192 ms) had best result in this section in the first interval, followed by Education in Cyprus(4211 ms). Education in the world had the biggest value for entrance time - 4472 ms, which meant participants needed a longer time for noticing the Block in the first interval of their sessions. According the Dwell Time, which represents the sum of all fixations and saccades within an AOI for the selected participants divided by the number of selected participants, the course Education in the world had biggest percentage for the WORLD Block section -31%, followed by WORLD Education – 29% and Education in Cyprus – 26 %. And about the information for HIT Ration (how many participants of the selected looked at least one time into a specific AIO) and revisit, the course Education in Cyprus had the best results. In the course mentioned above, every participant took a look into the Block and every

participant revisited our Block. Unlike this, the course World Education had the worse results with only 75% hit ration and two of three revisions in this section.

The results from the assignment that the participants performed showed which media types drew attention and distraction when reading the information in the Block. The results were divided in three tables according to the course that the participants visited. And the results from the questionnaire showed the attitude of the participants, and their negative and positive aspects of the WORLD Block.

7 Discussion and Conclusions

During the experiment a large volume of data was collected. The results were displayed and discussed in the previous section in three different parts. Analyzing the results from the Gaze Tracking we've seen that information with included multimedia types like flags, maps attract more attention than the information displayed only in a form of sentence. Expanding the media type of generated data from the web service i.e. games, flash, custom maps etc should be one of the next steps that need to be taken in improving this solution.

From the findings, we concluded that the location of the Block in the right hand side of the Moodle course is the appropriate place where the WORLD Block should be installed. Analyzing the other results, we've noticed that users are attracted by the flags and the maps of the countries displayed in the Block. The combination of a name of the country and the appropriate flag makes a really good impression for the users.

Analyzing the results from the questionnaire we've recognized that the users feel the potential of the WORLD Block and its possible implementation. Learning a lot of new information through a small area of the WORLD Block is the reason because most of the students assess the Block positively, believing that the Block would save them time.

Through analyzing the negative and positive aspects of the WORLD Block we've come to the conclusion that this solution presents a good approach for displaying a lot of new information from different areas of interest in a small section, but we should continue working on the structure on the Block adding new multimedia types and changing the structure of the sentence displayed, and even maybe changing the structure of the information like listing the information in points instead of using a full sentences. We also need to work on the interactivity with the user, personalization of the block and better design which will include more colors and multimedia information that will drew the users' attention and keep it there.

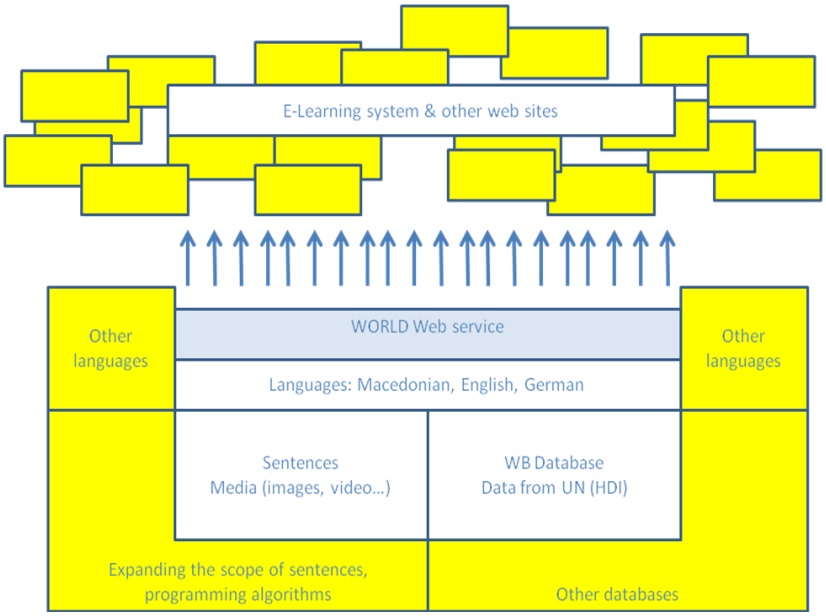


Figure5. Final goal of the project

References

1. R. J.K. Jacob, K. S. Karn, Ph.D, *Eye tracking in human-computer interaction and usability research: Ready to deliver the promises*. In R. Radach, J. Hyona, & H. Deubel, The mind's eye: cognitive and applied aspects of eye movement research (pp. 573–605). Boston: North-Holland/Elsevier.
2. M.A. Just, P.A. Carpenter, *A theory of reading: from eye fixation to comprehension,* 1980. Psychol Rev 87:329–354
3. K. Pernice, J. Nielsen, *Eyetracking Methodology: How to Conduct and Evaluate Usability Studies Using Eyetracking,* 2009
4. E. Cutrell, Z. Guan, *What are you looking for? An eye-tracking study of information usage in Web Search*. In Proceedings of CHI'07, Human Factors in Computing Systems, (San Jose, April 2007), ACM press, 407-416.
5. L. A. Granka, T. Joachims, G. Gay, *Eye-Tracking Analysis of User Behavior in WWW Search*, SIGIR '04 Proceedings of the 27th annual international ACMSIGR conference on Research and development in information retrieval Pages 478-479, 2004
6. B. Hogge, *Open Data Study – New Technologies*, May 2010
7. *Evaluation of public dialogue on open data*, Report to Research Councils UK, 2012
8. T. Davies, *Supporting open data use through active engagement,* Using open data: policy modeling, citizen empowerment, data journalism, June 2012
9. C. Guéret, *Decentralized Open Data*, Using open data: policy modeling, citizen empowerment, data journalism, June 2012

Risk Management Framework for IT-Centric Micro and Small Companies

Jasmina Trajkovski[1], Ljupcho Antovski[2]

[1]Trajkovski & Partners Management Consulting
Sveti Kliment Ohridski 24/2/1, 1000 Skopje, Macedonia
jasminat@tpconsulting.com.mk

[2]Faculty of Computer Science and Engineering
University Ss. Cyril and Methodius
Rugjer Boshkovikj 16, 1000 Skopje, Macedonia
ljupcho.antovski@finki.ukim.mk

Abstract. This paper proposes a new risk management framework tailored for IT-centric micro and small companies based on the analysis of the best practices in risk management concepts, specifically the risk management frameworks. The proposed framework for risk management is a synergy of various elements from the existing frameworks, tailored to the specifics of the IT-centric micro and small companies and deals with the identified challenges for the implementation of risk management frameworks. The framework focuses on 4 elements: people, policy, methodology and process, and tools.

Keywords: risk management methodology, risk management framework, ISO31000, ISO27005, enterprise risk management, IT-centric micro and small companies

1 Introduction

In this paper we review the risk management concepts, specifically the risk management frameworks, and based on them we propose a new risk management framework tailored for IT-centric micro and small companies. The proposed framework for risk management is a summary of various elements from the existing frameworks, but adapted to the specifics of the IT-centric companies and deals with the identified challenges for the implementation of risk management frameworks. The research is based on the direct experience of the leading author in the last 5 years with over 20 micro and small companies that are heavily IT-centric in their operations. In each of these companies, there has been a process of application of risk management frameworks, and specifically conducting the risk assessment exercises. The necessity of further work in testing the proposed framework in real-life IT-centric micro and small com-

S. Markovski, M. Gusev (Editors): ICT Innovations 2012, Web Proceedings, ISSN 1857-7288

panies is elaborated toward the end of the paper, and such work will be conducted as part of the PhD research on integrated risk management frameworks and the proposal of a usable model for valuation of risks for the micro and small companies.

This paper is structured in several segments. In Chapter 2, an overview of risk and risk management is provided together with reviews of the risk management frameworks and standards. In the following chapter, the related specifics and challenges for micro and small IT-centric companies are elaborated, while in the Chapter 4, the proposed risk management framework is presented.

2 Overview of Risks and Risk Management Frameworks

The main concepts of risks management in IT-centric micro and small companies are divided into 2 groups: (i) definition of risk, types of risks and risk management, and (ii) risk management frameworks and standards.

Based on the International standard for Risk Management – ISO31000, risk is defined as: "effect of uncertainty on objectives"(ISO, 2009), where the uncertainties include events (which may or not happen) and uncertainties caused by ambiguity or a lack of information, while the objectives can have different aspects (health and safety, financial, IT, environmental) and can apply at different levels (such as strategic, organizational, project, process). It also includes both negative and positive impacts on objectives. The risk is often expresses as a combination of the consequences of an event and the associated likelihood of occurrence. As we discuss risks management frameworks for IT-centric micro and small companies, the main focus are the organizational risks. There are various types of organizational risks such as program management risk, investment risk, budgetary risk, legal liability risk, safety risk, inventory risk, supply chain risk, and security risk. (NIST, 2011)

For the needs of the management of the IT-centric micro and small companies, all these risks could not be approached independently, and an integrated approach is necessary. This approach should be focused on the main drivers in the company, like the continual operations thru IT operation and known business processes so that the employees can understand what they should do. The reliance on IT as well puts the information security risks among the top as well. For the purposes of the research questions, we make the assumption that the management of these IT-centric micro and small companies deals with the legal and financial risks intuitively, and that they are not necessary to be included in the integrated risk management framework and approach of the company.

Having said that, for the purposes of the paper, we will look into the IT risk, information security risk and operational risk, which are respectively defined as:

- IT risk—that is the business risk associated with the use, ownership, operation, involvement, influence and adoption of IT within an enterprise. (ISACA, 2009)
- Information security risk—that is, the risk associated with the operation and use of information systems that support the missions and business functions of their organizations.(NIST, 2011)

- Operational risk - The most common definition, first published in The Next Frontier and also adopted in recent operational risk documents issued by the Basel Committee, is that "Operational risk is the direct or indirect loss resulting from inadequate or failed internal processes, people and systems, or from external events." (Haubenstock, 2001)

The next concept to be introduced is the risk management. ISO31000(ISO, 2009) defines the risk management very broadly as the coordinated activities to direct and control an organization with regards to risk. Other institutions have a more precise definition, as described for example in NIST special publication SP800-39 (NIST, 2011), where risk management is defined as a comprehensive process that requires organizations to:

- frame risk (i.e. establish the context for risk-based decisions);
- assess risk;
- respond to risk once determined; and
- monitor risk on an ongoing basis using effective organizational communications and a feedback loop for continuous improvement in the risk-related activities of organizations.

Risk management is carried out as a holistic, organization-wide activity that addresses risk from the strategic level to the tactical level, ensuring that risk based decision making is integrated into every aspect of the organization.

With the development of risk management as an organizational discipline, a more defined concept evolved, named Enterprise Risk Management (ERM). There are many definitions of ERM, but a representative one is from the COSO framework: "Enterprise risk management is a process, effected by an entity's board of directors, management and other personnel, applied in strategy setting and across the enterprise, designed to identify potential events that may affect the entity, and manage risk to be within its risk appetite, to provide reasonable assurance regarding the achievement of entity objectives" (COSO, 1992).

After setting the stage with the definition of the concept, lets look at the available risk management frameworks. Nowadays, there are several types of risk management methodologies, some of them issued by national and international organizations such as ISO, NIST, AS/NZS, BSI, others issued by professional organizations such as ISACA or COSO, and the rest presented by research projects. Each of these methods has been developed to meet a particular need so they have a vast scope of application, structure and steps. The common goal of these methods is to enable organizations to conduct risk assessment exercises and then effectively manage the risks by minimizing them to an acceptable level (Saleh & Alfantookh, 2011).

Table 1 provides a comparative overview of the elements of the various frameworks, methodologies and/or standards.

Vorster and Labuschagne in their work(Vorster; & Labuschagne, 2005) go even deeper in the analysis focusing solely on the methodologies for information security risk analysis and define a framework for comparing them. The objective of their framework is to assist the organization in the selection process o the most suitable

methodology and/or framework. The elements than they are taking into consideration include:

- Whether risk analysis is done on single assets or groups of assets
- Where in the methodology risk analysis is done
- The people involved in the risk analysis
- The main formulas used
- Whether the results of the methodology are relative or absolute

Some of these criteria are tightly related to the risk management considerations we have identified in the following section for the IT-centric micro and small companies.

Table 1. Overview of elements in risk management frameworks and methodologies

Type of framework	Main elements	Resource
Generic risk management frameworks	• 11 Principles for managing risks • 5 segment framework: mandate and commitment; design framework; implement risk management; monitor and review the framework; continual improvement • 5 step process: establish the context; risk assessment; risk treatment; monitoring and review; communication and consultation	ISO31000:2009 Risk Management Standard(ISO, 2009)
	• It has 4 sub-processes: Risk assessment process; Risk treatment process; Risk communication process; Risk review and monitoring process.	Corpuz and Barnes in their 2010 paper on integration information security policy into corporate risk management (Corpuz & Barnes, 2010)
Information Security Risk Management Frameworks	The tiers are: Organization, Mission / business processes and Information Systems, while the phases are Frame, Assess, Respond and Monitor	NIST SP800-39: Managing Information Security Risk (NIST, 2011).
	6 step process: context establishment; risk assessment; risk treatment; risk acceptance; monitoring and review; risk communication.	ISO27005:2008 Information Security Risk management (ISO, 2008).
	Views: STROPE - strategy, technology, organization, people, and environment Phases: DMAIC - define, measure, analyze, improve, and control cyclic phases.	Information security risk management (ISRM) framework for enterprises using IT (Saleh & Alfantookh, 2011)
IT Risk management frameworks	Domains: Risk governance, Risk evaluation and Risk response	RiskIT framework (ISACA, 2009)

Type of framework	Main elements	Resource
Operational Risk Management Framework	Components: identify, assess, respond to and control risk	COSO Enterprise risk management integrated framework (Aguilar, 2004)
	Elements: 1. leadership, 2. management, 3. risk, and 4. tools.	RMA Operational risk management framework (Taylor, 2006)

3 Risk Management considerations for IT-Centric Micro and Small Companies

As further elaborated in the authors' paper on challenges or implementation of risk management frameworks in IT-centric micro and small companies(Trajkovski J., 2012Trajkovski J., 2012) , there are several specifics and challenges that need to be taken into consideration when designing a suitable integrated framework for such companies. These include:

- Specifics of IT-centric micro and small companies:

– Exposure to various types of risks
– Limited resources for risk management
– Low resilience of the organizations to operations and information security risks

- Challenges related to meeting the following requirements:

– Need for integrated approach to treat various types of risks
– Need for comprehensive and usable methodology

These considerations, together with the findings from the analysis of the various existent risk management frameworks, standards and/or methodologies demonstrate the need for development of an integrated risk management framework for IT centric micro and small companies.

The developed framework should be applicable for the identified key risks groups: operational, IT and information security risks, as well as be implementable i.e. doable for an average team for risk management that includes 3-5 people, within 5-10 days annually, and the respective level of effort of 20-25 man/days on annual basis.

Regarding the need for comprehensive and usable methodology, the integrated risk management framework for IT-centric micro and small enterprises should be comprehensive and should present an understandable set of steps and activities, with clear inputs and defined outputs. The elements of the organizational context should be clearly defined and should allow for small or even non-existent organizational hierarchies and procedures. The framework should allow for unclear segregation of duties, and for very high criticality of managers and/or owners, as well as lack of documented knowledge.

The integrated approach should as well include an adjusted valuation model as part of the risk assessment phase that can be used on various types of risks. This model

should not be data intensive as experience shows that micro and small companies do not have access to historical data about risks, probabilities and impacts. It as well should not be based on complex calculations, require advanced skills, nor should it require too much time or people to conduct the risk assessment exercise.

4 Proposed Risk Management framework for IT-Centric Micro and Small Companies

As defined in the ISO31000:2009 Risk Management Standard (ISO, 2009), a risk management framework is a set of components that provide the foundations and organizational arrangements for designing, implementing, monitoring, reviewing and continually improving risk management in the organization.

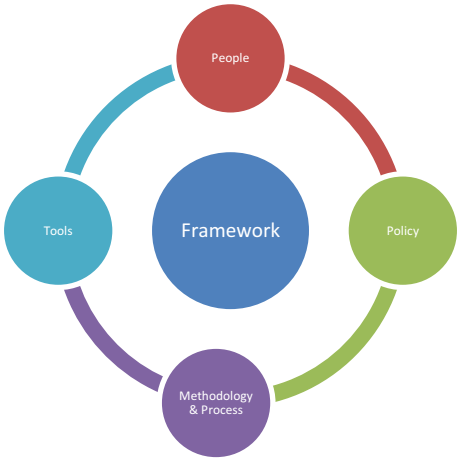


Fig. 1. Risk Management Framework

We presented in the previous section an overview of the various risk management frameworks connected to the identified key risks for an IT-centric micro and small company, showing the main elements of the frameworks. In order to provide a usable framework which covers all the key risks, and which takes into consideration the specifics of the IT-centric micro and small company, we took the common elements of the analyzed frameworks and created a customized framework. This new framework reflects the experiences gained from implementing risk management in over 20 micro and small IT-centric companies in the Balkan region which the authors have.

In future work, the authors will focus on further development of all the elements of the proposed framework and their elaboration in detail. The framework will then be tested on IT-centric micro and small companies. The results will be used to fine-tune the framework to the specific needs of these companies in the Balkan Region.

The customized framework for risk management in IT-centric micro and small companies presented on Figure 1 consists of: people, policy, methodology and process, and tools.

4.1 People

The People component of this framework deals with the Risk management team and the Risk management officer. It is described first, as it reflects the facts that risk management is people intensive process and the people are crucial for the successful implementation and maintenance of risk management in the organization

- Risk management team – to include the representatives from the main processes or units in the company, as well as the management team. Optimal number of representatives is 5 to 7.
- Risk management officer – a responsible person in the company, the owner, and managing director or other person from the management team in the forefront of the activities for risk management.

4.2 Policy

The risk management policy is a simple but straightforward document summarizing the intent and the approach for risk management. As main elements, the policy includes:

- Scope and purpose of the risk management
- Main objectives
- Risk management principles
- The commitment of management to risk management
- Allocated responsibilities for the process and results
- References to the methodology and process to be used
- Level of acceptable risk for the company.

The document is public and circulated to all employees. Its optimal length is 1 to 2 pages, and it should be in line with other management policies, if they exist in the company, such as Quality management policy or Information Security Management policy. The policy should be reviewed at least annually to reflect the changes in the environment of the company in which the risk are identified, assessed and managed, as well as the level of acceptable risk.

4.3 Methodology and Process

The risk management methodology includes comprehensive and implementable guidelines for conducting the risk management process. It enables the company to:

- Identify process and asset related threats and vulnerabilities

- Repeatedly conduct risk assessment with comparable/consistent results taking into account the already implemented mitigation steps
- Get a prioritized list of key risks using a common qualitative scale
- Decide on the level of acceptable risk for the company
- Identify further mitigation action necessary
- Define a realistic Risk Treatment Action Plan with necessary resources, priorities for the actions, responsible person

The methodology enables consistent implementation of the Risk management process and its activities. The Risk Management team should understand the Risk Management methodology and should be trained and competent to implement the Risk Management process.

Risk management process is a series of inter-related activities that enable the company to address risk. It includes the steps grouped in 3 phases as described in the in Figure 2. The main results of the process are:

- the process and/or asset register,
- the risk identification register,
- decision on range of value for probability and impact of risks, and calculation formula;
- the risk assessment register,
- decision on acceptable risk,
- the risk treatment plan,
- the risk treatment action plan,
- the risk measurement/monitoring log.

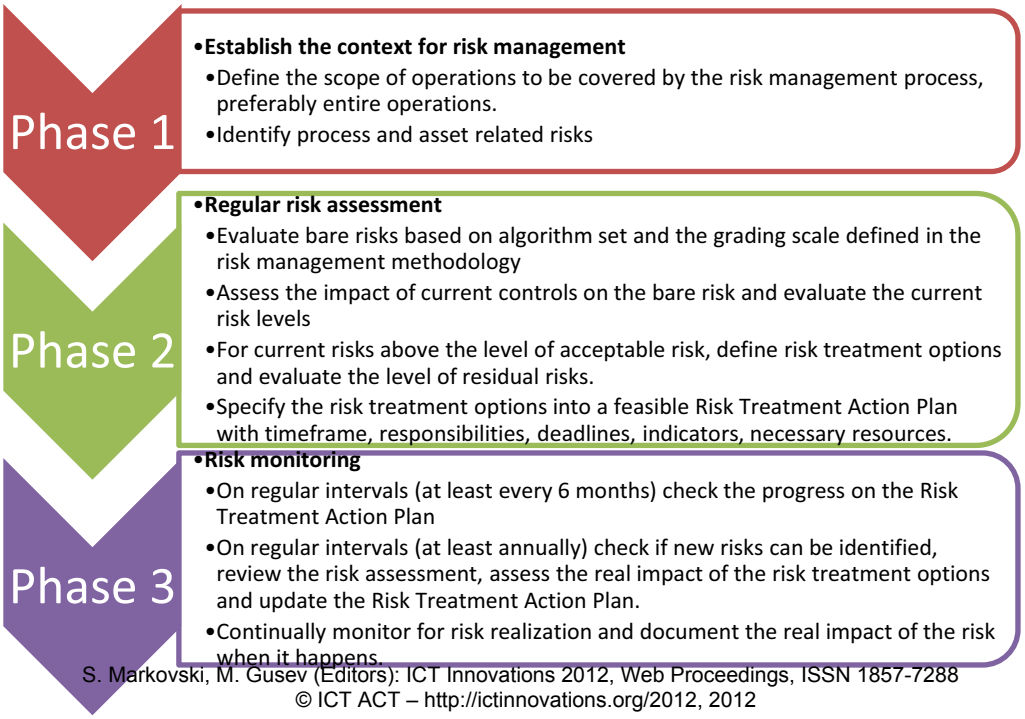


Fig. 2. Diagram of the phases in the Risk Management Process

4.4 Tools

Risk management toolkit is a usable software tool (spreadsheet or more advanced software) for gathering, calculating and presenting risk assessment results as well as other related information. It can include:

- the process and/or asset register,
- the risk identification register,
- the risk assessment register,
- the risk treatment plan,
- the risk treatment action plan,
- the risk measurement/monitoring log.

The benefits of using a toolkit is the automation of the calculations required for the risk assessment (Figure 3), as well as possibility for manipulation of the risk management results for better and more understandable presentation which will allow for adequate decision making by the management.

The toolkit should allow a maximal human influence on the results, as the risks identification and their assessment cannot be automated and still provide a usable result for the IT centric micro and small enterprise.

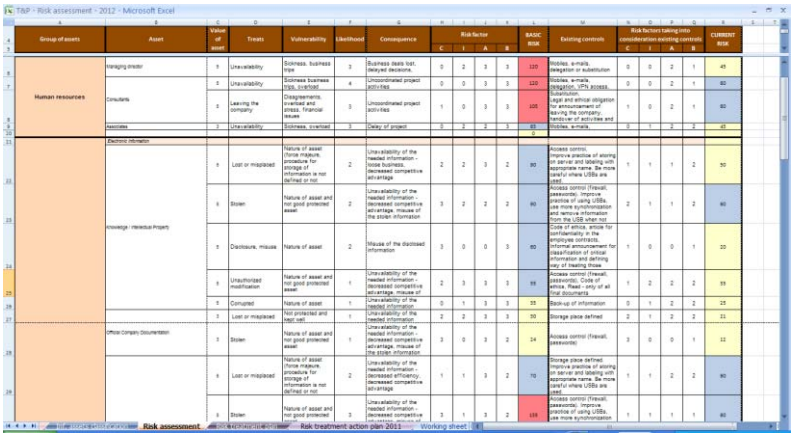


Fig. 3. Sample Risk Assessment spreadsheet

5 Conclusion

This paper reviewed the risk management concepts, specifically the risk management frameworks, and discussed a newly proposed risk management framework that is designed to reflect the specifics of the IT-centric micro and small companies. The

proposed framework is focused on the 4 main elements in risk management: people, policy, methodology and process, and tools. Due to the simplicity and the generalized approach of the framework, it can be used to deal with various types of risks to which micro and small companies are exposed. Its main benefit is that it addresses the challenge of very limited human resources for risk management in the companies.

The open questions which remain are concerned with the availability of adequate models for risk assessment i.e. valuation of risks so that the management of the micro and small companies can compare the various types of risks to which they are exposed, prioritize them and set appropriate risk mitigation controls.

In future work, the authors will focus on development in detail of all the elements of the proposed framework, elaboration of them in detail and specifically identifying a suitable and scientifically valid risk valuation model. The framework and model will then be validated and tested on IT-centric micro and small companies. The results will be used to fine-tune the framework and model to the specific needs of these companies in the Balkan Region.

6 References

1. Aguilar, M. K. (2004). COSO releases a new risk management framework. Accounting TOday, 18(19), 1.
2. Corpuz, M., & Barnes, P. H. (2010). Integrating information security policy management with corporatre risk management for strategic allignment. Paper presented at the 14th World Multi-conference on Systemics, Cybernetics and Informatics (WMSCI 2010).
3. COSO. (1992). Internal Control-Integrated Framework: Committee of Sponsoring Organizations of the Tread way Commission (COSO), AICPA/COSO.
4. Haubenstock, M. (2001). The Evolving Operational Risk Management Framework. The RMA Journal, 84(4), 5.
5. ISACA. (2009). The RISK IT framework: ISACA.
6. ISO. (2008). ISO/IEC 27005:2008 Information technology -- Security techniques -- Information security risk management: ISO.
7. ISO. (2009). ISO 31000:2009 - Risk Management - Principles and guidelines: ISO.
8. NIST. (2011). Managing Information Security Risk, SP800-39 NIST Special publication.
9. Saleh, M. S., & Alfantookh, A. (2011). A new comprehensive framework for enterprise information security risk management. Applied Computing and Informatics, 9(2), 107-118. doi: 10.1016/j.aci.2011.05.002
10. Taylor, C., ;. (2006). The RMA operational risk management framework. The RMA Journal, 88(5), 3.
11. Trajkovski J., A. L. (2012). Overview of risk management frameworks and challenges for their implemetation in ITcentric micro and small companies. Paper presented at the EuroSPI 2012, Vienna.
12. Vorster;, A., & Labuschagne, L. (2005). A framework for comparing different information security risk analysis methodologies. Paper presented at the SAICSIT 2005

On the General Paradigms for Implementing Adaptive e-Learning Systems

Emilija Kamceva[1] and Pece Mitrevski[2]

[1]FON University, Bul. Vojvodina bb, 1000 Skopje, Republic of Macedonia
emilija.kamceva@fon.edu.mk
[2]St. Clement Ohridski University, Faculty of Technical Sciences, Ivo Lola Ribar bb
7000 Bitola, Republic of Macedonia
pece.mitrevski@uklo.edu.mk

Abstract. Adaptive e-Learning systems are the newest paradigm in modern learning approaches. Nowadays, adaptive e-Learning systems are accountable for selecting learning materials or contents according to the learners' style, profile, interest, previous knowledge level, goal, pedagogical method, etc., all in order to provide highly personalized learning sessions. A number of researches have been conducted in the area of adaptive learning and different adaptive learning methods have been proposed. Due to similarity between learning objects graph and the formalism of Petri Nets, an approach based on Petri Nets for controlling the learning path among learning activities has brought many improvements to the paradigm of adaptive e-Learning systems. A model based on High-Level Petri Nets (HLPN) for modeling and generating behavioral pattern of students has been presented recently, bringing numerous advantages, but also a lot of open issues that need to be resolved. This paper exposes the fundamental characteristics of adaptive e-Learning systems, as well as the students' behavior, such as learning styles and levels of knowledge, thus providing a basis for a novel modeling framework for performance analysis of such systems, as our main goal in future research.

Keywords: e-Learning systems, Petri Nets, learning style, adaptive learning

1 Introduction

e-Learning for decades is one of the most attractive areas of scientific research that involves both pedagogical aspects and application of ICT in various domains (software engineering, semantic technology, communications, computer networks, etc.). The problem of development of personalized adaptive systems has been recognized from the outset as one of the permanent objectives, and current research efforts on the global level have resulted in the development of platforms that support and provide access to the general use.

Tang and Mccalla [1] proposed an evolving web-based learning system, which finds a relevant learning content according to the accumulated ratings given by students. Liu and Yang [2] introduced the Adaptive and Personalized e-Learning System

(APeLS), which provides dynamic learning content and adaptive learning processes for students to enhance the quality of learning. Papanikolaou *et al.* [3] suggested integrated theories of instructional design with learning styles to develop a framework that adapts different students to provide the presentation sequence of the learning content in a lesson. Chen [4] proposed genetic-based personalized e-Learning system to (i) generate appropriate learning paths according to the incorrect testing responses of an individual student in a pre-test, and (ii) provide benefits in terms of learning performance promotion.

2 The Concept of Adaptivity in e-Learning

During the past years, a lot of companies, faculties, universities and other institutions developed systems for common or personal use. There are many systems developed for e-Learning, such as: LRN; Blackboard; BSCW[1]; CLIX; InterWise; Moodle; OLAT[2], etc. Some of them have extensive support for adaptivity and provide both adaptive presentation support and adaptive navigation support [5]: AHA[3]; ALFANET[4] [6]; ELM-ART; EPIAIM; SQL-Tutor.

This section deals with adaptivity in e-Learning systems and explains *why* and *how* adaptivity is able to improve the quality of e-Learning environments.

The goal of adaptive e-Learning is aligned with exemplary instruction: delivering the right content, to the right person, at the proper time, in the most appropriate way - any time, any place, any path, any pace (NASBE)[5]. According to [7] adaptivity is of particular importance in the field of e-Learning for two main reasons:

- learners differ in their goals, learning styles, preferences, knowledge and background and the profile of a single learner changes (e.g. the knowledge increases as an effect of learning);
- the system can help the learner to navigate through a course by providing user-specific (not necessarily linear) paths.

Providing personalized access to the content (fitting the individual user's needs) and taking care of a single user (decisions on what is presented are based on the user's goals, knowledge, etc.) compensate for one significant problem of common e-Learning systems that provide the same view of the information for all learners. There are several different ways to categorize adaptivity features: Adaptive presentation support, Adaptive navigation support, Criteria for adaptation [7].

[1] Basic Support for Cooperative Work
[2] Online Learning And Training
[3] Adaptive Hypermedia Architecture
[4] Active Learning For Adaptive Internet
[5] National Association of State Boards of Education Study Group

2.1 Studies on Adaptive e-Learning

The adaptation of the teaching and learning process can be divided in four elements, based on a hypothetical e-Learning system, as described below: Adaptive content aggregation, Adaptive presentation, Adaptive navigation, and Adaptive collaboration support.

There were some studies related to adaptive e-Learning but they had different focuses and approaches. Semet *et al.* [8] used an artificial intelligence approach, called *"ant colony optimization"*. In this system ants are presumed as learners, and adaptive learning path is made by considering pheromone (a secreted or excreted chemical factor that triggers a social response in members of the same species) which is released by other learners. Zhu [9] used a *learning activity graph* and allocated Boolean expression to every edge of the graph. This expression includes required precondition and post condition for traveling through the edge. If this expression is evaluated correctly, corresponding edge is traversed by learner. Chen *et al.* [10] used *"Item response theory"* for providing individual learning path for each learner. Chang *et al.* [11] used a *Behavioral Browsing model* (B^2 model) based on High Level Petri Nets (HLPNs), which was introduced for modeling and generating behavioral pattern of students. Su *et al.* [12] proposed *object-oriented course modeling* based on HLPNs, whereas Liu *et al.* [13] proposed an approach based on Petri Nets for controlling learning path among learning activities.

2.2 Style and Models for e-Learning

Learning styles are various approaches or ways of learning. They involve educating methods, particular to an individual, which are presumed to allow that individual to learn best. Most people prefer an identifiable method of interacting with, taking in, and processing stimuli or information [14].

A learning style is a student's consistent way of responding to and using stimuli in the context of learning. There are several definitions of learning styles.

Keefe [15] defines learning styles as the "composite of characteristic cognitive, affective, and physiological factors that serve as relatively stable indicators of how a learner perceives, interacts with, and responds to the learning environment." Stewart and Felicetti [16] define learning styles as those "educational conditions under which a student is most likely to learn". Thus, learning styles are not really concerned with what learners learn, but rather *how* they prefer to learn. Honey and Mumford [17] defined the learning style as the attitudes and behaviors that determine student's preferred ways of learning. Learning style influences the effectiveness of training, whether that training is provided on-line or in more traditional ways.

This section describes the models developed for styles in e-Learning.

(Kolb's model). The ELT model outlines two related approaches toward grasping experience: Concrete Experience and Abstract Conceptualization, as well as two related approaches toward transforming experience: Reflective Observation and Active Experimentation. According to Kolb's model, the ideal learning process engages all four of these modes in response to situational demands. In order for learning to be

effective, all four of these approaches must be incorporated. As individuals attempt to use all four approaches, however, they tend to develop strengths in one experience-grasping approach and one experience-transforming approach. The resulting learning styles are combinations of the individual's preferred approaches. These learning styles are as follows: Converger; Diverger; Assimilator; Accommodator [18]. Kolb's model gave rise to the Learning Style Inventory, an assessment method used to determine an individual's learning style. An individual may exhibit a preference for one of the four styles – accommodating, converging, diverging and assimilating – depending on their approach to learning via the experiential learning theory model [19].

(Honey and Mumford's model.) Two adaptations were made to Kolb's experiential model. Firstly, the stages in the cycle were renamed to accord with managerial experiences of decision making/problem solving. The Honey & Mumford stages are: Having an experience, Reviewing the experience, Concluding from the experience and Planning the next steps. Secondly, the styles were directly aligned to the stages in the cycle and named Activist, Reflector, Theorist and Pragmatist. These are assumed to be acquired preferences that are adaptable, either at will or through changed circumstances, rather than being fixed personality characteristics [17].

(Neil Fleming's VAK/VARK model). One of the most common and widely-used [20] categorizations of the various types of learning styles is Fleming's VARK model (sometimes VAK), which expanded upon earlier neuro-linguistic programming (VARK) models [21] visual learners; auditory learners; kinesthetic learners or tactile learners [22]. Fleming claimed that visual learners have a preference for seeing (think in pictures; visual aids such as overhead slides, diagrams, handouts, etc.). Auditory learners best learn through listening (lectures, discussions, tapes, etc.). Tactile/kinesthetic learners prefer to learn via experience – moving, touching and doing (active exploration of the world; science projects; experiments, etc.). Students can also use the model to identify their preferred learning style and maximize their educational experience by focusing on what benefits them the most [14].

2.3 Components of e-Learning

The *content model* houses domain-related bits of knowledge and skill, as well as their associated structure or interdependencies. This may be thought of as a knowledge map of what is to be instructed and assessed, and it is intended to capture and prescribe important aspects of course content, including instructions for authors on how to design content for the model. A content model provides the basis for assessment, diagnosis, instruction, and remediation.

A *learning object* is "a collection of content items, practice items, and assessment items that are combined based on a single learning objective". The requirements for any content model fall into two categories: requirements of the delivery system and requirements of the learning content that is to be delivered [23].

Knowledge structures. The purpose of establishing a knowledge structure as part of the content model in any e-Learning system is that it allows for dependency relations to be established. Each element (or node) in the knowledge structure may be classified in terms of different types of knowledge, skill, or ability. Some example knowledge types include:

- *Basic knowledge (BK):* This includes definitions, examples, supplemental links (jpgs, avis, wavs), formulas, and so on, and addresses the *What* part of content.
- *Procedural knowledge (PK):* This defines step-by-step information, relations among steps, subprocedures, and so on, and addresses the *How* part of content.
- *Conceptual knowledge (CK):* This refers to relational information among concepts and the explicit connections with BK and PK elements, draws all into a "big picture" and addresses the *Why* part of content [23].

The *learner model* [23] represents the individual's knowledge and progress in relation to the knowledge map, and it may include other characteristics of the learner as a learner. As such, it captures important aspects of a learner for purposes of individualizing instruction. This includes assessment measures that determine where a learner stands on those aspects. Different student models have been explored: Han B. et al [24], AHAM [25] and NetCoach [26].

The *instructional model* manages the presentation of material and establishes (if not ensures) learner mastery by monitoring the student model in relation to the content model, addressing discrepancies in a principled manner, and prescribing an optimal learning path for that particular learner. Information in this model provides the basis for deciding how to present content to a given learner and when and how to intervene.

3 Petri Nets as a Modeling Tool

Many researchers have conducted research in the field of adaptive learning and have used different methods. At first, Liu *et al.* [13] proposed an approach based on Petri Nets for controlling the learning path among learning activities. This model has brought many improvements in adaptive systems but also many disadvantages, such as learning styles of students. Recently, the "B^2 model" [11] has been constructed, based on High-Level Petri Nets (HLPN), which was proposed for modeling and generating behavioral pattern of students. This new paradigm brings numerous advantages, but also a lot of open issues that need to be resolved.

3.1 Comparison of Models

To evaluate the practicability and accuracy of the B^2 model, behavioral patterns generated by the B^2 modeling tool have been compared with the actual behavioral patterns collected from elementary school students [27]. Actual behavioral patterns have been acquired by Chiu, Chuang, Hsiao, and Yang [28]. By two-stage cluster analysis, they classify three kinds of learning styles:

1. *Dilatorily type:* The student belonging to this type takes more time to browse a learning unit than other students. She/he often reviews the same learning unit and skips learning units.
2. *Transitory type:* The student spends the least amount of time in browsing and has the least browsing depth. Browsing order is irregular.
3. *Persistent type:* The browsing depth is the highest and browsing order is regular.

The B^2 model can capture students' learning behavior and then generate behavioral patterns for facilitating the verification of the researches regarding e-Learning.

For the generated behavioral patterns, whether or not a student quits an e-Learning course is determined by:

$$Prob_{quit} = C\frac{e^{-\lambda\theta}(\lambda\theta)^x}{x!}$$

where $\theta = \theta_{total} / \theta_{unit}$ refers to the time that a student has already stayed in an e-Learning course; θ_{unit} is the unit of time; λ refers to the rate of leave; and C is a constant for regulating the distribution curve. Since $Prob_{quit}$ is used to determine that a student quits, x=1. After the result derived using this equation determines that a student remains in an e-Learning course, the roulette wheel selection is used to select the next browsing action from: depth-first, breadth-first, skip, and review.

A conclusion has been drawn that the generated behavioral patterns based on the B^2 model are fairly similar to the actual behavioral patterns acquired from Chiu *et al.* [27] and confirms that the number of students who remain in an e-Learning system exponentially decreases in time. Furthermore, the experimental e-Learning system must also possess the additional function of collecting learners' behavior. It needs to take redundant time and effort in the development of the additional function for collecting learners' behavior. Obviously, the B^2 model, indeed, reduces cost and time of collecting learners' behavior. In terms of practicability, the generated behavioral patterns based on the B^2 model can serve as test data to validate the accuracy of an intelligent tutoring system. Therefore, the B^2 model can facilitate the verification process of an intelligent tutoring system.

In contrast the previous two models, Omrani *et al.* [29] have also developed a model based on High Level Petri Nets (i.e. Colored Petri Nets (CPNs), where a value (color) is allocated to each token, and Timed Petri Nets, where a real value corresponds to each token, representing the token age). This model is quite similar to the B^2 model and uses the same formulas (Poisson and normal distribution) to calculate the time of searching, the time required for assessment and the time the student spends in the queue. Three factors to adaptation are used: 1) Learning Style; 2) Knowledge Level; 3) Score. It should be noted that this adaptive model uses Petri Nets for the determination of adaptive learning style. Four learning styles are considered: Visual, Audio, Read/Write and Kinesthetic. A learning object in the learning object graph, which can be a chapter, is mapped to a place in the Petri Net. Colored tokens are learners, and they travel through learning object via transitions.

Although the method constructed on the basis of High Level Petri Nets uses the same mathematical formulas as well as the B^2 method, it differs in that the calculated

response time is computed in two viewpoints of black and white box. In viewpoint of black box, authors calculate only response time for each learner, whereas, in viewpoint of white box, in addition to response time, they show traveled path, time spent for each middle learning unit, earned score in each test, learning style and knowledge level for each learner.

Characteristic of this model is that the learning style of a new learner is checked before entering into introduction node and based on its type he/she is guided to two individual paths. These two paths represent the same course, but their content, due to its style, is presented with different media. However, authors use CPN tools (a fast simulator that efficiently handles untimed and timed nets) for editing, simulating, and analyzing the proposed model.

4 Discussion and Conclusion

In this paper we presented an overview of e-Learning systems and thoroughly described the concept of adaptivity. Various researchers, in different ways, have constructed models to improve adaptive e-Learning systems. We made a comparison of models that are created by means of Petri Nets and described their advantages and disadvantages.

Original Petri nets did not contain a concept of time, and an enabled transition would fire instantaneously. The introduction of deterministic time delays (that is, firing takes place if a transition is enabled for a specified amount of time, or tokens have memory, i.e. "aging tokens") has led to Timed Petri nets that were later extended to Stochastic Petri Nets (SPNs), where such delays are random variables based on given distributions. The firing of a transition corresponds to any discrete event of the modeled system, and represents a fundamental feature of Petri nets: the ability to graphically depict the *dynamic behavior* of a system. The analysis of an SPN model is usually aimed at the computation of more aggregate performance indices than the probabilities of individual markings. Several kinds of aggregate results are easily obtained from the steady-state distribution over reachable markings. Some of the most commonly and easily computed aggregate steady-state performance parameters include: the probability of an event, the probability mass function (*pmf*) of the number of tokens in a place, the average number of tokens in a place, the frequency of firing of a transition, the average delay of a token, etc. Moreover, efficient numerical algorithms for transient analysis of deterministic and stochastic Petri nets (DSPNs) and other discrete-event stochastic systems with exponential and deterministic events have been widely introduced [30].

Herewith, motivated by the aforementioned research papers, we propose our promising idea of using *analytical modeling approach* and Stochastic Petri Nets as a tool for adaptive model development and performance analysis. In this manner, lots of questions about the characteristics of adaptive e-Learning systems could be answered. Inspired by these ideas, our future work will be focused on developing (D)SPN model for performance analysis of adaptive e-Learning systems, by first investigating the dynamic behavior of learners.

S. Markovski, M. Gusev (Editors): ICT Innovations 2012, Web Proceedings, ISSN 1857-7288

References

1. Tang, T. Y., & Mccalla, G.: Smart recommendation for an evolving e-Learning system. In Proceedings of workshop on technologies for electronic documents for supporting learning, international conference on artificial intelligence in education (AIED 2003) (pp. 699–710). Sydney, Australia (2003)
2. Liu, H. I., & Yang, M. N.: QoL guaranteed adaptation and personalization in Elearning systems. IEEE Transactions on Education, 48(4), 676–687 (2005)
3. Papanikolaou, K. A., Grigoriadou, M., Magoulas, G. D., & Kornilakis, H.: Towards new forms of knowledge communication: The adaptive dimension of a web-based learning environment. Computers and Education, 39(4), 333–360 (2002)
4. Chen, C. M.: Intelligent web-based learning system with personalized learning path guidance. Computers and Education, 51(2), 787–814 (2008)
5. David H. and Mirjam K.: State of the Art of Adaptivity in e-Learning Platforms. Institute for Information Processing and Microprocessor Technology Johannes Kepler University, Linz
6. *ALFANET*. http://rtd.softwareag.es/alfanet. last download: 16.7.2007.
7. Brusilovsky, P. and Weber, G.: Collaborative Example Selection in an Intelligent Example-based Programming Environment, in *Proc. Of International Conference on Learning Sciences,* pp. 357 362 (1996)
8. Semet Y., Lutton E. and Collet P.: Ant Colony Optimisa-tion for e-Learning: Observing the Emergence of Peda-gogic Suggestions, *Proceedings of the* 2003 *IEEE Swarm Intelligence Symposium*, Indianapolis, pp. 46-52 (2003)
9. Zhu X.: Semi-Supervised Learning with Graphs, doctoral thesis, Language Technologies Institute, School of Computer Science, Carnegie Mellon University (2005)
10. Chen C. M., Lee H. M. and Chen Y. H.: Personalized e-Learning System Using Item Response Theory, *Computers & Education*, Vol. 44, No. 3, pp. 237-255. doi:10.1016/j.compedu.2004.01.006 (2005)
11. Chang Yi-Chun, Huang Ying-Chia, Chu Chih-Ping: B^2 model: A browsing behavior model based on High-Level Petri Nets to generate behavioral patterns for e-Learning, Expert Systems with Applications 36 12423–12440 (2009)
12. Su J. M., Tseng S. S., Chen C. Y., Weng J. F. and Tsai W. N.: Constructing SCORM Compliant Course Based on High-Level Petri Nets, *Computer Standards & Interfaces*, Vol. 28, No. 3, pp. 336-355. doi:10.1016/j.csi.2005.04.001 (2006)
13. Liu X. Q., Wu M. and Chen J. X.: Knowledge Aggregation and Navigation High-Level Petri Nets-Based in e-Learning, *International Conference on Machine Learn-ing and Cybernetics*, Vol. 1, pp. 420-425 (2002)
14. From Wikipedia, the free encyclopedia, http://en.wikipedia.org/wiki/Learning_styles
15. Keefe, J. W.: Learning style: An overview. NASSP's *Student learning styles: Diagnosing and proscribing programs,* pp. 1-17, Reston, VA. National Association of Secondary School Principals (1979)
16. Stewart, K. L., & Felicetti, L. A.: Learning styles of marketing majors. *Educational Research Quarterly,* 15(2), 15-23 (1992)
17. Honey, P & Mumford, A: The Learning Styles Questionnaire, 80-item version. Maidenhead, UK, Peter Honey Publications (2006)
18. Kolb, D.: *Experiential learning: Experience as the source of learning and development*. Englewood Cliffs, NJ: Prentice-Hall. ISBN 0-13-295261-0 (1984)
19. Smith, M. K.: *David A. Kolb on experiential learning*. Retrieved October 17, 2008, from: http://www.infed.org/biblio/b-explrn.htm (2001)

20. Leite, W. L., Svinicki, M., and Shi, Y.: Attempted Validation of the Scores of the VARK: Learning Styles Inventory With Multitrait–Multimethod Confirmatory Factor Analysis Models, pg. 2. SAGE Publications (2009)
21. Hawk T. F., Shah A. J.: Using Learning Style Instruments to Enhance Student Learning *Decision Sciences Journal of Innovative Education* doi:10.1111/j.1540-4609.2007.00125.x (2007)
22. LdPride. (n.d.). *What are learning styles?* Retrieved October 17, 2008
23. Shute, V. J. & Towle, B.: Adaptive e-Learning. Educational Psychologist, 38(2), 105-114 (2003)
24. Han H. C., Giles L., Manavoglu E., Zha H., Zhang Z. and Fox E.A., Automatic document metadata extraction using support vector machines, in Proc. of the third ACM/IEEE-CS Joint Conference on Digital Libraries, pp. 37 – 48 (2003)
25. Bra D. P., Houben G.J. and Wu H.: AHAM: A Dexter based Reference Model to support Adaptive Hypermedia Authoring, *in Proc. of the ACM Conference on Hypertext and Hypermedia,* Darmstadt, Germany, pp. 147-156 (1999)
26. Weber G., Kuhl H. C. and Weibelzahl S.: Developing Adaptive Internet Based Courses with the Authoring System: NetCoach, *in Proceedings of the third workshop on adaptive hypermedia* (2001)
27. Chang Y.C., Kao W.Y., Chu C.P. and Chiu C.H., A learning style classification mechanism for e-Learning, Computers & Education 53, 273–285 (2009)
28. Chiu, C. H., Chuang, C. H., Hsiao, H. F., & Yang, H. Y.: Exploring the patterns of computer mediated synchronous collaboration by elementary school students, Computers in Human Behavior (in press)
29. Omrani F., Harounabadi A. and Rafe V.: An Adaptive Method Based on High-Level Petri Nets for e-Learning, Journal of Software Engineering and Applications, Vol. 4 No. 10, pp. 559-570. doi: 10.4236/jsea.2011.410065 (2011)
30. Choi H., Kulkarni V.G., Trivedi K.S.: Transient analysis of deterministic and stochastic Petri nets, in: M. Ajmone Marsan (Ed.), Application and Theory of Petri Nets 1993, Lecture Notes in Computer Science 691, Springer, Berlin, pp. 166–185 (1993)

S. Markovski, M. Gusev (Editors): ICT Innovations 2012, Web Proceedings, ISSN 1857-7288

Research, Implementation and Application of the SQBC Block Cipher in the Area of Encrypting Images

Zlatka Trajcheska, Vesna Dimitrova

Faculty of Computer Science and Engineering

zlatka.trajcheska@gmail.com, vesna.dimitrova@finki.ukim.mk

Abstract. The application of block ciphers in the modern society is enormous. One branch of this development is getting a new dimension – the application of quasigroups and quasigroup transformations in the block ciphers. This paper contains the results of the research involving the family of block ciphers named SQBC (Small Quasigroup Block Cipher), which is based on quasigroups and quasigroup transformations. Actually, we will discuss the results of encrypting images in different formats using the SQBC block cipher based on quasigroups of order 4.

Keywords: cryptography, block cipher, images, quasigroup, encryption, SQBC

1 Introduction

Cryptography is not only a scientific field, it is a fundamental part of our everyday living. We could not possibly imagine a world without the usage of its benefits. The modern society relies on the cryptographic algorithms for the purpose of simple electronic communications, electronic transactions or military causes. Considering this, the development and improvement of cryptography is essential for today's way of living.

Generally, cryptographic algorithms can be divided into two groups: symmetric and public key (or asymmetric) algorithms. On the other side, symmetric algorithms can be stream or block ciphers. Block ciphers are the ones that we are interested in, as SQBC is a family of block ciphers. In particular, SQBC is based on quasigroups and quasigroup transformations of small order, and we will use quasigroups of order 4. For the purposes of this research, before we continue with the results of the encryption, we will first discuss some basic terms about block ciphers, quasigroups and their classification, and finally the SQBC algorithms for encryption and decryption.

2 Block Ciphers

Encryption of a plaintext message using a block cipher is done by separating the plaintext message into blocks with fixed length, and encrypting them individually, so

S. Markovski, M. Gusev (Editors): ICT Innovations 2012, Web Proceedings, ISSN 1857-7288

that every block from the plaintext message is encrypted into a block of the encrypted message that has the same length. Both sides of the communication have the same secret key. The block length varies, but usually we are considering block length of 64, 128, 256 bits and so on. To obtain the final encrypted message from the individually encrypted blocks, we can use different ways to combine them. These are known as modes of operation. Two of them (ECB and CBC) are used in the research, so they will be briefly discussed in addition.

The ECB (Electronic Code Book) mode of operation is the simplest and the least secure of the modes. It simply concatenates the encrypted blocks.

The encryption with CBC (Cipher Block Chaining) mode is done by combining the encrypted messages the following way: at the beginning we have an initial vector (IV) which is of same length as the block. This initial vector is XORed with the first block, and then encrypted, so that is the first encrypted block. Afterwards, we get every new encrypted block by XORing the previous encrypted block and the current plaintext block and encrypting the result.

3 Quasigruops and Their Classification

A groupoid (G,*), where * is a binary operation, is called a quasigroup if for each a and b in G there are unique x and y in G so that:

$$a * x = b = y * a \tag{1}$$

or formally

$$(\forall\, a, b\ \in G)\ (\exists x, y\ \in G)(a * x = b = y * a) \tag{2}$$

For every quasigroup five other quasigroups can be derived, which are called parastrophes. In particular, we are interested in the so called left and right parastrophe which are defined by the following equivalences [1]:

$$x*y=z \qquad y=x\backslash z \qquad x=z/y \tag{3}$$

- There are 576 quasigroups of order 4, and we will use them in the research. Considering the results of a previously conducted research [1], there are several classifications on the quasigroups of order 4. In this paper we consider the classification by fractality (fractal and non-fractal) and classification by Boolean representation of quasigroups (linear, non-linear and purely non-linear). The results from [1] reveal several properties of the quasigroups which will be used in the research. Actually, several representative quasigroups were taken in our research and also their left and right parastrophes. We use the following quasigroups given with its lexicographical number quasigroup 1, 6, 158 and 181 – as a representative of the fractal and linear quasigroups, the non-fractal and linear quasigroups, the non-fractal and non-linear quasigroups and the non-fractal and purely non-linear quasigroups, respectively.

- It is expected that the quasigroups that are linear by Boolean representation or fractal give bad results used for encryption and we want to see if these quasigroups cause some undesirable structures or fractals[1] in the encrypted image.

4 The Algorithm of the SQBC Block Cipher

SQBC is a family of block ciphers that use quasigroups and quasigroup transformations to encrypt the plaintext message in encrypted message, using a working key generated directly from a secret key. Obviously, the cipher includes decrypting the encrypted message, which also is based on quasigroup transformations.

The algorithm of this cipher is described in "SQBC - Block cipher defined by small quasigroups" [2] and in this section the discussed content is referenced to it.

Without going into detail about the algorithms used and their construction, we can more generally say that this cipher is different from most of the common and well known block ciphers, as it uses two algorithms for encryption and two for decryption. Actually, one encryption algorithm is used to encrypt the first block and another one to encrypt all the other blocks. Similarly is for the decryption. The encryption/decryption algorithm uses *e* transformation or *d* transformation, respectively. The *e* and *d* transformations are defined as:

$$e_{*,l}(\alpha) = b_1 b_2 \dots b_n, \text{so that } b_{i+1} = b_i * a_{i+1} \quad (4)$$

$$d_{*,l}(\alpha) = c_1 c \dots c_n, \text{so that } c_{i+1} = a_i * c_{i+1} \quad (5)$$

There is also an algorithm for generating working key out of the secret key which should be at least 80 bits long.

In order to examine the avalanche effect[2] of this block cipher, it was implemented in Java [3]. The interface allows the user to enter a number of rounds and the block length. Also, it is required to enter the secret key and generate the working key before we proceed to encrypting and decrypting. The avalanche effect can be calculated both for the encryption and the working key generation algorithm. This implementation can be used in various purposes. In this paper, we will apply it in the area of encrypting images.

5 Application in the Area of Encrypting Images

The modern way of living and the fact that the information technologies are ubiquitous in all scientific fields, the culture and life in general demand paying a lot more attention to the data security, especially when concerning sensible data. Lately, those sensible data are often some images. And this is where the idea for encrypting images

[1] the term fractal here is used as a general term for any recognizable pattern or structure that can appear on the picture

[2] the term avalanche effect represents the percentage of different bits between the encryptions of two very similar plaintext messages (that are differ only in one bit)

comes from. Actually, the SQBC block cipher may be used in several devices like personal identification cards, where there is an obvious need for encrypting images. Before we enclose the results of the experimental research conducted with various image formats, we should mention that the header of the images, which varies from one format to another, is not encrypted, because we want to display the results of the encryption like an image as well. That way we can visually see any fractals that may appear.

5.1 Encryption of *.bmp Images with the SQBC Block Cipher

Using the mentioned implementation an experimental research was conducted in order to see the effect of encrypting images with this cipher. The results are given in details in the paper "Encrypting images with SQBC"[4]. In this paper we will shortly discuss this part too, as we want to compare the results with the ones obtained from our research.

In the previous research given in [4] were used 24-bit Bitmap image format. This format is made up from 4 blocks – Header, InfoHeader, RGBQuadArray and color indexed array. More details about the format are given in Table 1. This image format is encoded with Windows-1251 encoding, so some slight changes were made in the previous implementation. Actually, to make the testing faster and easier, the manual input of the plaintext message was avoided and was replaced from reading the plaintext message from an external file. To avoid problems with the encoding the HxD hexadecimal editor was used. As said before, the first 54 bytes (the header) was not encrypted.

Table 1**.** Bitmap file structure

Name	Size	Description
Header	14 bytes	Windows Structure: Bitmap File Header
InfoHeader	40 bytes	Windows Structure: Bitmap Info Header
RGBQuad array	4 bytes	
color-index array	Varies	

At first, the CBC mode was used for encryption. The results were very similar to each other and all results showed that a fractal does not appear. Figure 3 shows the original *.bmp image, as for 4 and 5 show the results of the encryption using quasigroups number 1 and 181, respectively. The block length used is 128 bits.

Fig. 1. The original *.bmp image (1)

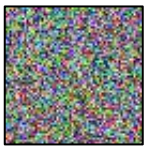


Fig. 2. Encryption of (1) with quasigroup number 1 and CBC mode

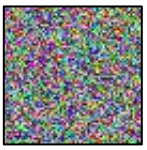


Fig. 3. Encryption of (1) with quasigroup number 181 and CBC mode

Also, the same experiment was conducted using ECB mode to see more clearly the direct impact of the quasigroups used and its properties. Another change was made, which was to encrypt all blocks with the same algorithm. The testing was done with 8, 16 and 24 bits, because of the structure of *.bmp images. Actually, by encrypting 8 bits long blocks every color in the pixel is encrypted as one block. Encrypting blocks of 16 bits is actually encrypting two colors of the pixel as one block, and finally encrypting blocks which are 24 bits long result in encrypting exactly one pixel.

Using the ECB mode as described some fractal structures started to appear. They are not that obvious at first, but we must consider that the algorithm of block cipher itself has a great impact on the fact that the encrypted image is not periodic. When using ECB mode, as expected, whenever the input was periodic, the encryption was periodic, too. The images below show the results of the encryptions with ECB mode, using blocks which are 8, 16 and 24 bits long, and the quasigroups 1, 6, 158 and 181 for the properties mentioned above.

The Figures 6, 7 and 8 show the results of the encryption of the original image (1) using the quasugroup 1, ECB mode and block length of 8, 16 and 24 bits.

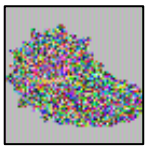


Fig. 4. Encryption of (1) with quasigroup 1, 8 bit block

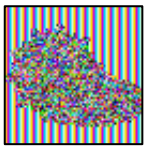


Fig. 5. Encryption of (1) with quasigroup 1, 16 bit block

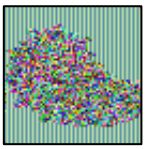


Fig. 6. Encryption of (1) with quasigroup 1, 24 bit block

The Figures 9, 10 and 11 show the results of same experiments using the quasigroup 6, the Figures 12, 13 and 14 using the quasigroup 158 and the Figures 15, 16 and 17 using the quasigroup 181.

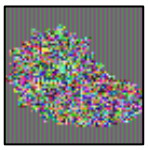


Fig. 7**.** Encryption of (1) with quasigroup 6, 8 bit block

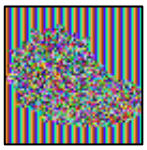


Fig. 8**.** Encryption of (1) with quasigroup 6, 16 bit block

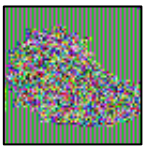


Fig. 9**.** Encryption of (1) with quasigroup 6, 24 bit block

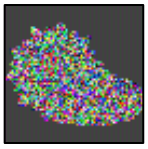


Fig. 10**.** Encryption of (1) with quasi. 158, 8 bit block

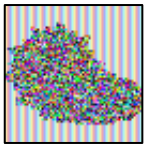


Fig. 11**.** Encryption of (1) with quasi. 158, 16 bit block

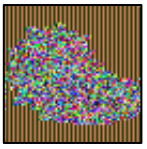


Fig. 12**.** Encryption of (1) with quasi. 158, 24 bit block

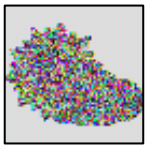


Fig. 13**.** Encryption of (1) with quasi. 181, 8 bit block

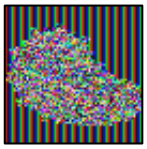


Fig. 14**.** Encryption of (1) with quasi. 181, 16 bit block

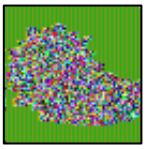


Fig. 15**.** Encryption of (1) with quasi. 181, 24 bit block

The encryption of images with SQBC is more elaborated in [4] and in the next sections, we will discuss more about this research – encrypting images of other formats.

5.2 Encryption of *.jpg images with the SQBC block cipher

The *.jpg image format is massively accepted and used nowadays. It provides a very powerful compression which contributes to the small size of the *.jpg image file. But, this reflects on the quality of the picture.

The *.jpg format is very different from other image formats. While *.png, *.bmp and even *.gif have the so called lossless compression and allow the image to be fully restored, the *.jpg format is intended to have loss of quality in order to obtain small size of the file [5].

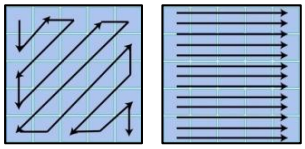
The structure of a *.jpg file according to its official documentation is complex and may contain a lot of markers. But, most often the *.jpg images are saved as JFIF (JPEG File Interchange Format). Table 2 shows the structure of a JFIF segment.

Table 2**.** Structure of a JFIF segment

Name	Size	Description
APP0 marker	2 bytes	It is always the same value: 0xFFE0
Length	2 bytes	The segment length including the APP0

		marker including the APP0 marker
Identifier	5 bytes	Always has the same value 0x4A46494600, which is JFIF0 in ASCII
Version	2 bytes	The first byte is for the major version, and the second one for the minor version
density units	1 bytes	Holds the information about the units for pixel density fields (0 – only aspect ratio, 1 – pixels per inch, 2 – pixels per centimeter)
X density	2 bytes	horizontal pixel density
Y density	2 bytes	vertical pixel density
thumbnail width (tw)	1 bytes	thumbnail width in pixels
thumbnail height (th)	1 bytes	thumbnail height in pixels
thumbnail data	3 x tw x th	uncompressed 24-bit RGB rasterised thumbnail

Without discussing about all the details of this format, we will consider only the aspects which are important about the encryption of the images. Actually, while encoding the image, a chromatic subsampling occurs. In simple words, this means that the encoding uses the flaws of the human eye to compress the image and save it in smaller resolution, so that it wouldn't be noticeable on first glance. After the subsampling, the image channels are divided on blocks of 8x8 bits. They undergo several operations, which finally leads to dependencies between the blocks. Figure 18 a) shows the so called zigzag encoding.



a) b)

Fig. 16. a) JFIF zigzag encoding b) Encrypting the image with SQBC

Now that the picture is divided into dependent blocks it will be encrypted sequentially, bit by bit, as shown on Figure 18 b). So, with the encryption of the image, the dependencies are lost and the output cannot be shown as a *.jpg image. This does not mean that the images of *.jpg format cannot be encrypted or decrypted with SQBC, but the result will have to be stored textually.

Because the output of the encryption cannot be shown as an image, Figures 19 and 20 show the original and the encrypted image in textual and hexadecimal representation.

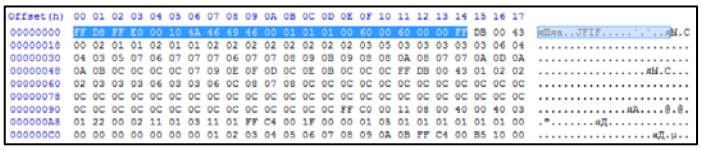


Fig. 17**.** The original *.jpg image in textual/hexadecimal representation

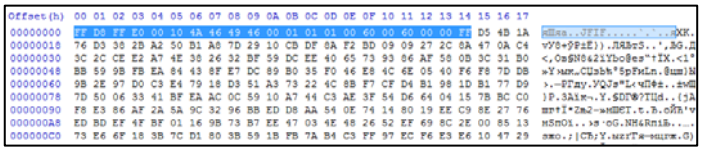


Fig. 18**.** The encrypted *.jpg image in textual/hexadecimal representation

5.3 Encryption of *.png Images with the SQBC Block Cipher

Unlike the *.jpg image format, the *.png image format is bitmapped and enables lossless compression. There are two aspects of its structure that are important for the encryption and that are the *.png file header and several markers called chunks that the file might contain .

The *.png file header is represented by the fixed 8 bytes 89 50 4E 47 0D 0A 1A 0A hexadecimal. The details are given in Table 3.

Table 3. Structure of a *.png file header

Hexadecimal bytes	Purpose
89	For the systems that don't support 8-bit data, this is a marker to distinguish the *.png file from any other text file
50 4E 47	This is PNG in ASCII, so that it would be recognizable in text editor too
0D 0A	Marking the end of conversion in DOS-Unix style.
1A	Stops the preview under DOS when the command type is used
0A	Marking the end of conversion in Unix style.

As said before, a *.png file can contain several chunks which should not be changed during encryption if we want the output as an image. The most important are so called critical chunks. The chunks are made of four fields: length, chunk type, chunk data and CRC. The critical chunks are: IHDR – the first chunk, PLTE – marks the palette, IDAT – marks the image content and IEND – marks the file end. There are also ancillary chunks which are considered that may not cause problems in the file decoding, but the experience in our research shows otherwise. So, we decided not to change the ancillary chunks. Some of them are bKGD, gAMA, hIST, iCCR, pHYs, sRGB. As explained in the previous sections, we will use the HxD hexadecimal editor to prepare the images for encryption. First, we used CBC mode and block lengths of 8, 16 and

128 bits. The 8 and 16 bit long block is used to observe because the RGB values are represented in 8 bites for each pixel, and sometimes there are additional 8 bits in the pixel to represent the alpha values of the palette. The 128 bit block length is chosen in order to observe the encryptions for longer blocks. We used 20 rounds and different keys while encryption. As before, we used the same original image given in Figure 3 in *.png format and quasigroups number 1, 6, 158 and 181 for the discussed reasons.

The Figures 21, 22 and 23 show the results of the encryption of the original image using the quasugroup 1, CBC mode and block length of 8, 16 and 128 bits.

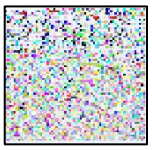


Fig. 19. Encryption with quasigroup 1, 8 bit block

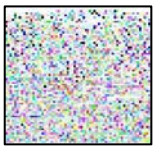


Fig. 20. Encryption with quasigroup 1, 16 bit block

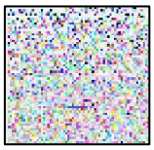


Fig. 21. Encryption with quasigroup 1, 128 bit block

The Figures 24, 25 and 26 show the results of same experiments using the quasigroup 6, the Figures 27, 28 and 29 using the quasigroup 158 and the Figures 30, 31 and 32 using the quasigroup 181.

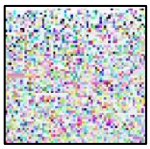


Fig. 22. Encryption with quasigroup 6, 8 bit block

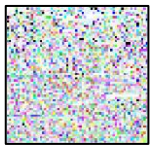


Fig. 23. Encryption with quasigroup 6, 16 bit block

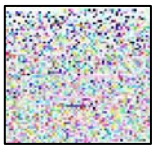


Fig. 24. Encryption with quasigroup 6, 128 bit block

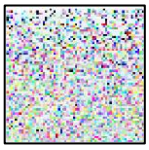


Fig. 25. Encryption with quasigroup 158, 8 bit block

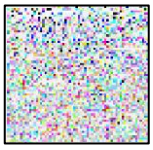


Fig. 26. Encryption with quasigroup 158, 16 bit block

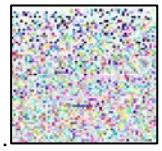


Fig. 27. Encryption with quasigroup 158, 128 bit block

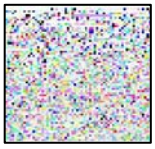


Fig. 28. Encryption with quasigroup 181, 8 bit block

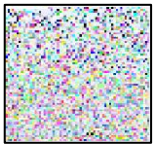


Fig. 29. Encryption with quasigroup 181, 16 bit block

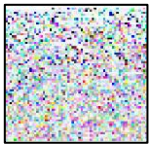


Fig. 30. Encryption with quasigroup 181, 128 bit block

It is obvious that no fractal structures appeared using the CBC mode. That is why we tried encrypting with ECB mode. It is clear that if some fractal structures appeared it will be while using the quasigroup number 1, as it is one of the quasigroups with worst properties. Even using the ECB mode we don't see any fractal structures to appear. The *.png file has more complex compression than the *.bmp file so this is the most probable reason that no obvious fractals can be recognized.

6 Conclusion

In this paper we gave the results of research of different types of images using the SQBC block cipher and we can conclude that the differences between the image file formats lead to different results of the encryption. This way, when encrypting the *.bmp images we can clearly distinguish the results of encryption using CBC and ECB mode, because of the very structure of the file and the way that the data is organized within it – every color is represented by 8 bits, making a pixel represented by 24 bits. This means that by encrypting a color the result is also a color. Here, the use of some quasigroups results in fractals in the encrypted image.

Unlike the *.bmp format, the *.png format provides encryptions that cannot be distinguished when using the CBC and ECB mode. Also, no matter which quasigroup we use, a fractal structure doesn't seem to appear. This is probably a consequence of the specific compression and the special palette that the *.png format uses.

As discussed before, the *.jpg format doesn't allow the encryptions to be shown as images, because of the dependencies between the blocks and the special compression. Maybe it will be possible if the design of the block cipher is changed, but that was not our goal.

About the future work, there are still a lot of things to consider about the SQBC block ciphers. These results should be theoretically examined. The research can be expanded to encrypting sound as well, which is currently done. In the future work we can include quasigroups of higher order.

7 References

1. Dimitrova V.: Quasigroup processed strings, their Boolean representations and applications in cryptography and coding theory, PhD Thesis, Skopje, 2005
2. Markovski S., Dimitrova V. and Mileva A.: "SQBC - Block cipher defined by small quasigroups", Loops'11, 2011
3. Trajcheska Z., Petkovska M., Kostadinoski M. and Velkoski G.: Implementation of SQBC in Java, FCSE, Skopje, 2012
4. Dimitrova V., Trajcheska Z., Petkovska M.: Encrypting images with SQBC, The 9th Conference for Informatics and Information Technology (CIIT), Bitola, April 2012
5. http://www.scantips.com/basics9j.html

Social Signal Processing and Human Action Recognition in Communication Services

Tomaz Vodlan[1] and Andrej Kosir[2]

[1] Agila d.o.o.,
Tehnoloski park 19, 1000 Ljubljana, Slovenia
tomaz.vodlan@agila.si
[2] Faculty of Electrical Engineering, University of Ljubljana,
Trzaska cesta 25, 1000 Ljubljana, Slovenia
andrej.kosir@ldos.fe.uni-lj.si

Abstract. We present a novel communication and evaluation scenario for user interaction with a communication device (television). To improve user experience with communication services, we merge three domains: social signal processing, human action recognition and human-computer interaction. Our scenario includes applied and theoretical solutions about user interaction with a communication device. We declared the atom actions, which are represented by unique hand gestures and replace functions on the remote controller. On the other hand, we include some social signals like agreement and disagreement, which can represent a user's reaction to currently played content. The advantage of our solution is the hierarchical communication scenario that allows the user to quit the procedure at any step. On the other hand, the computation time of low-level features within a recognition step represents the biggest drawback. The use of our method allows the user to control the communication device using hand gestures. Likewise, the use of social signals in interaction can improves the user experience with the communication device.

Keywords: social signal processing, human action recognition, human-computer interaction, communication scenario

1 Introduction

A user's experience with communication services in his/her natural environment is still unpleasant and requires a high level of attention. The human-computer interaction in communication device handling can be improved if we utilise the user's past and present behaviour to ease user interaction by predicting his/her decisions (recommend service or content). The recommender system approach using only past behaviour was later enhanced by incorporating the user and service context. However, these approaches are socially ignorant, and we are not aware of any attempt that has

S. Markovski, M. Gusev (Editors): ICT Innovations 2012, Web Proceedings, ISSN 1857-7288

been made to utilise the user's social signals during his/her interaction with the communication service. Our goal is to prove that we can increase the usability of human-computer interaction if social signals are taken into account.

Social signal processing [1] is the research domain that aims to provide computers with the ability to sense and understand human social signals [2]. Social signals are expressed with a group of nonverbal behavioural cues initiated by the human body as a reaction to a current situation in everyday life. These behavioural cues are grouped into five groups [3]: physical appearance, gestures and postures, face and eye behaviour, vocal behaviour and space and environment. Context plays a crucial role in understanding human behavioural cues and includes all cues in the physical and social environment of the observer. The social signals are most distinctly expressed in the interaction between two people. If we replace one person with a communication device, we get human-computer interaction (Fig. 1). In that case, the computer (communication device) recognises the user's social signals and uses them in the user-communication device interaction. This procedure can improve the user experience and increase the efficiency level of a communication service.

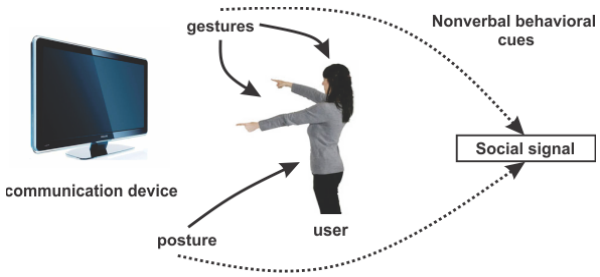


Fig. 1. Nonverbal behavioural cues (hand gestures, head movements, change in posture) in the case of the user-communication device interaction.

The extraction of social signals that will be used in the user-communication device interaction will be based on the action recognition method. Human action recognition in video is the process of labelling image sequences with action labels [4]. The current methods for action recognition of social signals do not take into account that the recognised event has more than one meaning depending on the context.

The action recognition domain ignores the social aspect of events. On the other hand, the introduction of social context in the human-computer interaction can increase the usability of the current applications. The proposed idea of this paper is to utilise automatically extracted and processed social signals in the user-communication device interaction in order to induce (provide the relevant list of available options) and support the user's decisions while passing through the user's interaction procedure. It is based on utilisation of human social intelligence and on the fact that it is natural for humans to transmit social signals in several verbal and nonverbal ways. In terms of signal processing, we are dealing with the extraction of social signals. The recognised social signals are then used to assist users while interacting with the communication device.

At the beginning of the paper, we present the problem statement of using social signals in the human-computer interaction (Sec. 2). Our proposal includes theoretical and applied solutions. Since there is no appropriate dataset that we can test our idea with, we will record an annotated dataset for our purposes. In section 3, we present the communication and evaluation scenario using mock interfaces. The evaluation methodology represents an important part of the whole solution. Evaluation is based on two approaches: a comparison between the test and control set of users and on the accuracy of action recognition and the social signal extraction method (Sec. 4).

2 Problem Statement

The domain of human-computer interaction is relatively unexplored in terms of automatic social signal extraction in video using action recognition methods. In the literature, we can find the following problems: the problem of using real world data for social signal extraction [2], and the general problem of social signal extraction in human-computer interaction [3]. The identification of applications that could have benefit from social signal processing in [2] and [5] is another challenging issue. On the other hand, the design of smart environments represents a big challenge for the current researchers [6].

2.1 Data and resources

The relevant test data are crucial for a successful evaluation of our proposed method – for automatisation of communication between the user and communication device. The datasets available for research purposes can be divided into three classes: datasets for human action recognition and detection in video, datasets proposed for social signal processing and datasets for hand gesture recognition.

There are many datasets for human action recognition and detection. They are categorised into two classes [7]: datasets for action recognition are in [8-10], datasets for action detection are in [11-13]. In action recognition datasets, each video clip contains only one action. The aim is to classify the type of action. On the other hand, in action detection datasets, each video sequence contains multiple actions and the aim is to recognise these actions in the scene (location and time).

Among the datasets that are appropriate for social signal processing, we focus on those where the social signals are presented with hand gestures and head poses. In most cases, this information is included in the bodies of discussion sessions as in [14] and [15]. In [16], the dataset of head poses that are recorded during meetings and office tasks is presented. On the other hand, the HPEG dataset [17] is a dataset of hand gestures and head poses.

Datasets with hand gestures that are not recorded in the context-aware environment belong to the group of datasets for hand gesture recognition. The keck gesture dataset [18] presents military signals, while the Cambridge hand gesture dataset [19] includes basic hand motions (gestures) that are not specifically used in everyday situations.

Since there is no appropriate dataset to test our proposed method, we plan to create an annotated video dataset with included atom actions and social signals. The communication and evaluation scenario with mock applications will be presented later in this paper.

2.2 Proposed solutions

This paper presents theoretical and applied solutions that can improve user-communication device interaction. A hierarchical communication scenario of human-computer interaction where social signals are used represents the first theoretical solution. The scenario provides decision support for the user at each level of communication (choice of device, choice of service, choice of content). The second theoretical solution presents the methodology for evaluation of the impact of social signals on a user's decisions. In this paper, only the basic evaluation concept of our solution is mentioned.

Applied solutions include automatisation of user interactions with a communication device, social signal extraction based on action recognition in video and the already mentioned annotated dataset. The first includes the introduction of atom actions that will be used for automatic handling of a communication device. The second solution includes the action recognition method in a context-aware environment with the aim extracting social signals. The last includes the recording of an annotated dataset.

3 Annotated Dataset

Since there is no appropriate dataset to determine the impact of the application of social signals, we will record our own annotated dataset. A hierarchical communication scenario will be constructed that aims to predict the use of social signals in human-computer interaction in the real world. The evaluation procedure for a communication scenario and a mock user interface and human operator interface will also be presented.

3.1 Communication scenario

We propose a user scenario in human to communication device interfacing that takes place in a living room. There are several communication devices such as a remote control, several mobile phones, etc. The structure of this scenario will guide every user along the planned steps in a predefined order. At every step, there is a decision fork where the user can either continue or quit the chosen path. The communication scenario by steps is described in Table 1. However, parallel with each step, feedback to the system is included. It presents data acquisition control provided by the human operator. For test data recording purposes, automatic action recognition and social signal extraction will be provided at each step by a human operator (real-time). Later, recognition and extraction will be based on video analysis algorithms and machine learning algorithms. Based on that scenario, an annotated dataset will be created.

Table 1. Steps in a communication scenario.

Step	Description	Recognition
I.	Start to use communication device.	Did someone approach the device?
II.	Decide on communication action.	Which communication device was selected?
III.	Social signal induction.	User reaction to induction.
IV.	Perform a selected communication action.	Actions of user interface control.
V.	User decides to end or interrupt the current communication action.	End of the action? Did user leave the observation area?
VI.	Decide to go back to step II or go to next.	New communication action?

Test data acquisition scenario

The scenario for acquisition of test data about communication is shown in Fig. 2. In addition to the user and communication device, the scenario also includes a human operator and persons who encourage the user to express his/her emotions more distinctly. The user is observed by the human operator. He/she recognises the actions and signals produced by the user and provides system feedback in real time. His/her decisions are marked as ground truth data. Our goal is to establish the natural environment for the user who will use the device, therefore some additional people are included in the scenario. The scenario preparation is based on an experimental design that aims towards social signal acquisition and evaluation of the impact of the application of social signals in a specific communication scenario – controlling the device.

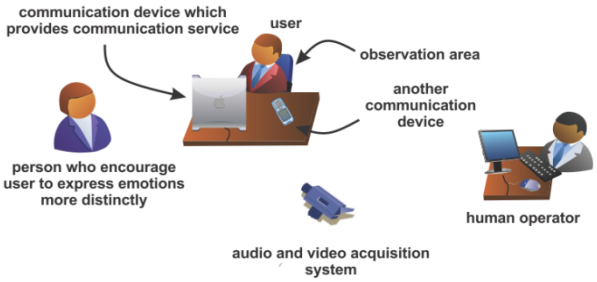


Fig. 2. Test data acquisition scenario.

3.2 Evaluation procedure for communication scenario

The user data acquisition procedure will involve a dedicated set of communication services. Human operators provide ground truth action recognition, social signal extraction and system feedback to the test user in real time. The evaluation procedure consists of four steps. The first step includes the implementation of a specific set of communication services on a remote control and mobile phone together with a user interface and user guidance. A human operator interface will be also integrated. The second step includes establishment of controlled environment that will emulate the

natural home environment of the user. The environment will be equipped with audio and video recording devices. The third step includes the selection of a test and control group of users according to the experimental design. All users will be informed about the procedure. They will, one by one, perform the communication scenario through their own choices at decision forks. The system will be driven by the human operator, who provides ground truth action recognition and social signal extraction. The video and audio will be recorded during the sessions. The fourth step includes evaluation of the data. It will focus on comparison between the test and the control group, on the accuracy of social signal extraction and on the accuracy of action recognition.

3.3 Appropriate activities for action recognition and social signal extraction

Action recognition and social signal extraction will be based on video sequences. Hereinafter, we will describe the actions that are appropriate for that type of analysis. We divide actions into three groups: atom actions, social signals and other actions. We will briefly describe each of these groups in the context of the predefined scenario.

Atom actions present hand gestures through which the user communicates with the communication device. These actions replace the basic functionalities of a communication device user interface (remote control). The basis for the selected gestures is the "gestures with mouse" of a Mozilla Firefox application [20].The functions on a remote control that will be defined with hand gestures are: power on, power off, stop, pause, play, next channel, previous channel, increase volume, decrease volume, mute and remote unit dial (1-9). Each of the hand gestures will be defined with a unique movement. Based on that, some kind of codebook with hand gestures will be proposed. Some gestures from that codebook are presented in Fig. 3. These hand gestures – defined as atom actions – will be mapped into social signals that allow a user to manage the communication device in a context-aware environment.

Social signals are actions that in a context-aware environment represent a user's reaction to current content. The focus in this scenario will be on hand gestures, head movements and some other actions that are clearly defined with movements. The basic scenario includes two different contexts where social signals are presented. The first context includes communication when the user approaches the communication device and the device offers him/her various video contents. The second context includes a scenario where the user watches the same video content for a longer time. Social signals that can be extracted in both cases are: agreement (nodding), disagreement (head shaking), thinking (put hand to a chin, move hand), fear (cover eyes), etc. Performance of these signals is presented in Fig. 4.

Other actions are defined as actions that cannot be expressed as reactions to current video content but are defined as a reaction to outside influences. Activities are determined based on user observation during his/her interaction with the communication device in a natural environment. The focus will be on whole body movements as well as on upper body movement recognition. The first group of actions includes the following situations: user approaches the observation area and displays that he/she wants to watch TV (starts communication action) and user stands up and leaves the observa-

tion area (pause communication action, stops communication action). The second group includes the following situations: a user takes a new communication device in his/her hand (if mobile phone – mute the sound on TV), a user lies down during a communication action (he/she is tired – decrease the volume on TV), etc.

Fig. 3. Examples of atom actions presented with hand gestures.

Fig. 4. Examples of social signals presented with hand gestures and head movements.

3.4 Annotation of dataset

As we have already mentioned, the subset of the recorded audio-visual dataset with video information about the actions recorded in the context-aware environment will be the basis for our work. The information about each action will be presented as a record in an XML document. In the head of XML document, there will be basic information about the whole set of actions including: number of actions, frame rate and space and time units. Each action will then be described in the following form: action name, action number (unique ID number), time interval when action appears (starts, ends) and spatial position in space.

3.5 Applications used by user and human operator

As we see in the scene of the scenario (Fig. 2), there are two devices where applications will be used. The first will be used by a human operator and the second by the user.

The human operator interface (Fig. 5a) consists of various buttons through which the human operator makes notes (pushing the buttons on interface) about user behaviour. His/her recognition will be used as feedback information for the user. The buttons on the interface are grouped in four groups: social signals, atom actions, other activities and video content. In addition to that, the selection bar, a window for displaying the user's previous activities and a window where the current time is displayed are also shown on the interface.

The user interface (Fig. 5b) represents the applied version of the communication device. The interface consists of two parts: the window in which the video content is

playing and the window in which the feedback of the human operator is displayed. Based on that, the user always knows how the human operator recognised his/her reaction to the currently played video content. In this case, the human operator replaces the automatic recognition of activities and signals. This feedback information from the human operator is important from various aspects. If the user does not know how its social signals are interpreted, his/her emotional response is much less distinctive than if he/she knows. The consequence of that is an unpleasant user experience that leads to useless test results.

Fig. 5. a) Human operator interface and b) User interface.

4 Data Evaluation

The annotated video dataset will be evaluated in two different ways. The first one involves a comparison between the test and the control group. The aim of the comparison is to determine the intensity of impact of social signals on a user's decisions in a communication scenario. The second way is based on evaluation of the action recognition and social signal extraction method on the test set.

4.1 Comparison between test set and control set

The test set will be represented by a group of users whose induced social signals during interaction with a communication device will be taken into account. On the other hand, the control group will be represented by a numerically comparable group of users whose induced socials signals will not be taken into account. The comparison between both sets of users will be based on pre and post questionnaires. The questionnaires will be based on one of the methods for user experience evaluation. One of the most appropriate is the system usability scale (SUS), which represents a simple questionnaire. It uses the Likert scale with ten questions and provides a high degree of reliability. AttrakDiff, usability heuristics, etc are also appropriate methods. The aim of comparison is to verify the impact of social signals on a user's decisions. If the impact of social signals is big enough, the results of comparison between both sets

must show the difference in contentment with the selected content, in contentment with the communication device, and in user's interaction time with device.

Since the application of social signals is not the only factor that has an impact on user's decisions, the experimental design will be used to control other factors in the way that the realistic evaluation of social signals impact can be evaluated. Each decision of a user $u \in U$ is modeled as

$$\varphi = f(u, v_1, \dots, v_m, w_1, \dots, w_q), \quad (1)$$

where $v_1, \dots, v_m$ are the factors we can control and $w_1, \dots, w_q$ are the factors we cannot control.

4.2 Accuracy of the action recognition and social signal extraction method

The second method of evaluation is evaluation of the accuracy of the action recognition and social signal extraction method on the test set. From video sequences of actions and signals, we will extract low-level features using the space-time interest point detector (STIP) [21] that has proven successful thus far [22]. Then hierarchical Bayesian feature selection (HBFS) [23] will be used to select most representative local features for each class of atom actions. A dimension of features space will be reduced through dimensionality reduction methods (LPP, LDA, etc.). The most appropriate method for our dataset will be selected. Among the many techniques of pattern recognition, various SVM classifiers will be tested for our purposes. A test video dataset will be divided into two groups. The performance of the proposed algorithm will be examined based on a confusion matrix and these measures; precision, recall and F-measure. The time complexity of the algorithm will also be measured.

5 Conclusion

The proposed communication model will be used for data acquisition, including social signals, about the user. During the interaction, the user will use atom actions for communication device handling. On the other hand, the system will extract the user's social signals that represent his/her reaction to the current video content.

The important predisposition for successful evaluation of data is the initial emotional state of user, which must be taken into account. Based on our experimental design, the impact of other factors will be minimised. Evaluation will be based on pre and post questionnaires and on the accuracy of the action recognition method.

The aim of the proposed model is automatisation of the interaction between the user and communication device which including user's social signals. Improved user experience and usability might be the biggest advantages of the proposed model. On the other hand, the drawbacks might be the additional learning of atom actions and a slow algorithm for action recognition and social signal extraction.

Acknowledgments. "Operation part financed by the European Union, European Social Fund."

References

1. Pentland, A.: Social Signal Processing. IEEE SP Magazine 24(4), 108–111 (2007)
2. Vinciarelli, A., Pantic, M., Bourlard, H.: Social Signal Processing: Survey of an Emerging Domain. Image and Vision Computing 27(12), 1743–1759 (2009)
3. Vinciarelli, A., Salamin, H., Pantic, M.: Social Signal Processing: Understanding social interactions through nonverbal behaviour analysis. In: Proceedings of the International Workshop on CVPR for Human Behaviour vol. 231287, pp. 42–49. IEEE, (2009)
4. Poppe, R.: A survey on vision-based human action recognition. Image and Vision Computing 28(6), 976–990 (2010)
5. Murino, V., Cristani, M., Vinciarelli, A.: Socially Inteligent Surveillance and Monitoring: Analysing Social Dimensions of Physical Space. In: Proceedings of the International Workshop on Socially Intelligent Surveillance and Monitoring, pp. 51–58 (2010)
6. Cook, D.J.:Learning Setting – Activity Models for Smart Spaces. IEEE Intelligent Systems 27(1), 32–38 (2012)
7. Yuan, J., Liu, Z.: TechWare: Video-Based Human Action Detection Resources [Best of the Web]. IEEE Signal Processing Magazine 27(5), 136–139 (2010)
8. Gorelick, L., Blank, M., Shechtman, E., Irani, M.: Weizmann Human Action Database. http://www.wisdom.weizmann.ac.il Cited 16 May 2012
9. KTH. Action Database. http://www.nada.kth.se Cited 16 May 2012
10. Liu, J., Luo, J., Shah, M.: YouTube Action Dataset. http://www.umiacs.umd.edu Cited 17 May 2012
11. Gilbert, A.: MultiKTH. http://www.andrewjohngilbert.co.uk Cited 17 May 2012
12. Laptev, I., Marszalek, M., Schmid, C., Rozenfeld, B.: Hollywood Human ActionsDataset. http://www.di.ens.fr/~laptev/download.html Cited 18 May 2012
13. Marszalek, M., Laptev, I., Schmid, C.: Hollywood 2 Human Actions and Scenes Dataset. http://www.di.ens.fr/~laptev/actions Cited 18 May 2012
14. Hung, H., Chittaranjan, G.: The Wolf Corpus: Exploring Group Behaviour in a Competitive Role-Playing Game. In: ACM Multimedia, pp. 879–882 (2010)
15. Vinciarelli, A., Dielmann, A., Favre, S., Salamin, H.: Canal 9: A Database of Political Debates for Analysis of Social Interactions. In: Proceedings of the International Conference on Affective Computing and Intelligent Interaction, pp. 1–4 (2009)
16. IDIAP. Head Pose Database. http://idiap-head-pose-db.sspnet.eu/ Cited 16 May 2012
17. Asteriadis, S.: Head Pose and Eye Gaze (HPEG) Dataset. http://www.image.ece.ntua.gr/~stiast/HPEG/ Cited 17 May 2012
18. Lin, Z., Jiang, Z., Davis, L.S.: Keck Gesture Dataset. http://www.umiacs.umd.edu Cited 16 May 2012
19. Imperial College London. Cambridge Hand Gesture Dataset. http://www.iis.ee.ic.ac.uk/~tkkim/ges_db.htm Cited 17 May 2012
20. Mozilla Firefox. Gestures. https://addons.mozilla.org/sl/firefox/addon Cited 28 Aug 2012
21. Laptev, I.: On Space-Time Interest Points. IJCV 64(2–3), 107–123 (2005)
22. Thi, T.H., Zhang, J., Cheng, L., Wang, L., Satoh, S.: Human Action Recognition and Localisation in Video Using Structured Learning of Local Space-Time Features. In: Proceedings of the 7th IEEE International Conference on AVSS, pp. 204–211. IEEE, (2010)
23. Carbonetto, P., Dorko, G., Schmid, C., Hendrik, K., de Freitas, N.: Learning to Recognise Objects with Little Supervision. International Journal of Computer Vision 77(1–3), 219–237 (2008)

Investigating Students' Acceptance of a Learning Management System in University Education: A Structural Equation Modeling Approach

Arsim Fidani, Florim Idrizi

Faculty of Natural Science and Math, Department of Informatics, State University of Tetova, 1200 Tetovo, Macedonia
{arsim.fidani, florim.idrizi}@unite.edu.mk

Abstract. Tools emerged from information technology, such as Learning Management Systems (LMS), have been proven to significantly boost the students' performance and productivity. Hence, many higher education institutions have adopted such systems. Given the fact of the growth rates in adopting LMS in university education and the imperativeness of the factors' evaluation before the adoption of such a system at State University of Tetova in Macedonia, adds to the urgency to conduct a research by which factors that predict the students' intention to adopt LMS to be examined. In this paper, based on extant literature review and by utilizing the Unified Theory of Acceptance and Use of Technology (UTAUT), as the best model for predicting individual's technology acceptance, a research model is developed. By employing survey, the relations between the factors are described. Empirical data is gathered through purposive sampling technique. The results of this study, beside the theoretical contribution, are of a great importance to the university itself in promoting the use of technology in teaching and learning, and also confirm the need for adopting a LMS.

Keywords: Learning Management System, UTAUT, Intention, Adoption

1 Introduction

Over the years information technology (IT) is increasingly becoming an integral or imperative part of everyone's working and personal life. Information and knowledge is needed to be reached everywhere and at any time [1].

Learning Management System (LMS), as a tool emerged from IT, and it is considered as one of the most significant developments in the use of IT in universities in the last decades[2]. Such information system (IS) provides many benefits to individuals and organizations [3]. From the students' perspective, these systems provide them with the ability to access the course materials, delivered by the instructors, and use communication and interactive features in their learning activities, which in turn have been proven to significantly boost their performance and productivity [4]. Hence, many higher education institutions have adopted such

S. Markovski, M. Gusev (Editors): ICT Innovations 2012, Web Proceedings, ISSN 1857-7288

systems. The adoption of LMS is influenced by a number of factors, grouped as critical success factors into four categories [5] namely, instructor, student, IT, and university support, hence this paper takes into consideration the students' personal decision to adopt such a system.

Although, there are several studies that have focused specifically on the users' – students' intention to adopt such systems [6-11], however we should bear in mind that there are different institutional cultures and characteristics [12], thus the richness of each findings can help the worldwide community to enrich their understandings about the factors influencing LMS's acceptance. Moreover, given the fact of the growth rates in adopting a LMS in university education and the imperativeness of the factors' evaluation before the adoption [5] of such a system at State University of Tetova in Macedonia, adds to the urgency to conduct a research by which factors that predict the students' intention to adopt LMS to be examined.

Based on extant literature review and by utilizing the Unified Theory of Acceptance and Use of Technology (UTAUT), as the best model for predicting individual's technology acceptance by explaining nearly 70% of the variance in usage intention [13], a research model is developed.

The results of this study, beside the theoretical contribution- validation of UTAUT, are of a great importance to the university itself (the State University of Tetova), and also to other universities in promoting the use of technology in teaching and learning, confirm the need for adopting-deploying a LMS, if one is not deployed yet, and most importantly they identify and enrich the understandings of the factors that influence its acceptance.

2 Literature Review and Theoretical Framework

2.1 Learning Management System (LMS)

LMS is viewed differently among different perspectives [14] and also, as a relatively new concept, it is often confused with other similar concepts, such as e-learning, digital learning, virtual learning and distance learning etc.[1]. Since, all of these concepts represent modern achievements in the education process, for the purposes of this paper the term LMS will be used accordingly.

As an important platform to support effective learning environment, a LMS is defined as a system which employs a range of ICTs to offer an online platform over the internet, where a whole course can be planned, facilitated and managed by both the teacher and the learner [15]. Moreover, they are defined as web-based tools used to manage, implement and assess online learning and teaching. examples of which are categorized into proprietary systems such as Blackboard, WebCT, Desire2Learn, Angel etc and open – source systems such as Sakai, Moodle, OLAT etc [16]. All of the functionalities provided by these systems tend to motivate students, enhance their efficiency and cost-savings [14] and most importantly, tend to accelerate their learning processes [17].

2.2 Unified Theory of Acceptance and Use of Technology (UTAUT)

Many studies have been conducted and several theoretical models have been proposed to explain users' intention to use IT. The UTAUT model [18], developed in an effort to improve the predictive power of a user acceptance model, integrates elements across the eight prominent theoretical models, which are build up on each other. The newly developed UTAUT model includes four variables, performance expectancy, effort expectancy, social influence, and facilitating conditions, and up to four moderators of key behaviors, gender, age, experience, and voluntariness. While the eight models taken individually varied in explanatory power from 17% to 53% of the variance in user intentions to use IT, UTAUT explained 69% of the variance [19].

2.3 Reseach Model and Hypothesis development

Although TAM is considered to be one the most common theories in the field of adoption, due to the contradictory results in e-learning studies that adopted TAM [20, 21] cited in [22], and also due to UTAUT's novelty in the field of user acceptance [23], its high predictive power and the new significant relations-paths between its constructs found in recent e-learning studies [3, 22, 24], a research model is proposed, as shown in Fig. 1, which clearly utilizes UTAUT as a base theory.

2.3.1 Performance Expectancy, Effort Expectancy and Attitude

Within each individual model used to develop UTAUT, performance expectancy is the strongest predictor of intention [18]. Moreover, extant studies [24] and [22] found that performance expectancy has a positive effect on the users' attitudes.
H1: Performance Expectancy has a positive effect on the users' Behavioral Intention.
H2: Performance Expectancy has a positive effect on the users' Attitude.

On, the other hand, effort expectancy is defined as the degree of ease associated with the use of the system [18]. Similarly, as perceived ease of use of TAM which is found to have significant effect on perceived usefulness, attitude and behavioral intention, effort expectancy has been found to be influential on performance expectancy, attitude [3, 22, 24] and behavioral intention.
H3: Effort Expectancy has a positive effect on the users' Performance Expectancy.
H4: Effort Expectancy has a positive effect on the users' Attitude.
H5: Effort Expectancy has a positive effect on the users' Behavioral Intention.

Attitude has been found to be strong and important determinant of the users' behavioral intention in many studies. In the study in which attitude is incorporated into UTAUT model, the results indicated that attitude toward using technology can predict the users' behavioral intention [24].
H6: Attitude has a positive effect on the users' Behavioral Intention.

2.3.2 Social Influence and Facilitating Conditions

Social influence, as a strong and direct determinant of behavioral intention in UTAUT, it is incorporated in other models as well as an external factor having a positive effect on attitude [22].
H7: Social Influence has a positive effect on the users' Attitude.
H8: Social Influence has a positive effect on the users' Behavioral Intention.

Facilitating conditions are a direct determinant of the users' behavioral intentions [18]. Moreover, new significant relationships are identified between facilitating conditions and effort expectancy [3].
H9: Facilitating Condition has a positive effect on the users' Behavioral Intention.
H10: Facilitating Condition has a positive effect on the users' Effort Expectancy.

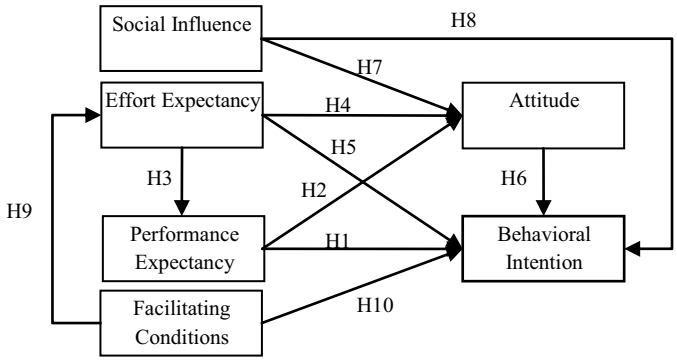


Fig. 1. Research Model

3 Methodology

3.1 Research Approach and Strategy

Since the study involves hypothesis testing through data collection and statistical analysis then it adopts a positivist knowledge claim [25], hence a quantitative approach was used in this study. Regarding the research strategy, survey was employed since the study focuses on contemporary event and it does not require control over behavioral events [26].

3.2 Sample, Data Collection, Questionnaire and Data Analysis Method

Purposive sampling, as a form of nonprobability sampling [27], was employed in this study as a sampling method. Thus, the empirical data is gathered through the first year students studying at different departments at State University of Tetova, to whom the course outlines and course materials were presented through authors' - instructors'

personal website. The designed questionnaire for the purposes of this study, was firstly distributed among 27 students, namely a pilot study was conducted to ensure its appropriateness and understandability. The final questionnaire, after the pilot study, comprised of 24 observed and 6 latent variables. The five point Likert scales, ranging from strongly disagree to strongly agree, was used for all the questions measuring the observed variables. Totally, 213 students completed the questionnaire.

IBM SPSS Statistics 20 is used for computing the reliability coefficients and the explanatory factor analysis, whereas for the confirmatory factor analysis, known as Structural Equation Modeling (SEM), LISREL 8.80 is used.

3.3 Validity and Reliability

To assess the validity in this study, distinct aspects of validity are considered such as content validity, which is achieved from the extensive literature review [28], and factorial validity as a subtype of construct validity, which can assess both convergent and discriminant validity are examined by using factor analytic techniques.

The results obtained from the factor analysis technique, namely the KMO and Bartlett's test of sphericity, show that that the partial correlations among variables are all greater than 0.5, ranging from 0.684 to 0.827, and the factor model is appropriate whereby all the relationships are significant ($p < 0.05$). To assess convergent and discriminant validity, Average Variance Extracted (AVE) of at least 0.5 [29] and the shared variance is calculated by the using the following equation AVE= Σ (factor loading)2/Σ(factor loading)2 + Σ measurement error [3]. The results show that the amount of AVE ranged from 0.615 to 0.827, and the shared variance (squared correlation coefficient between constructs) was below the amount of average variance extracted.

To assess the reliability in this study, internal consistency reliability is applied by using the Cronbach's alpha value higher than 0.70 for the study to be internally consistent and acceptable. The results of this study indicate that the research instrument used is internally consistent and acceptable, having the total reliability equal to 0.835. The validity and reliability results of this study are shown in Table 1 and Table 2.

Table 1. Instrument Validity and Reliability

	KMO >0.5	AVE >0.5	Cronbach's Alpha >0.7	Total Reliability
Performance Expectancy (PE)	0.792	0.689	0.841	0.835
Effort Expectancy (EE)	0.779	0.615	0.782	
Attitude (A)	0.789	0.769	0.848	
Facilitating Condition (FC)	0.827	0.827	0.875	
Social Influence (SI)	0.800	0.821	0.871	
Behavioral Intention (BI)	0.684	0.747	0.790	

Table 2. Discriminant Validity

	PE	EE	A	FC	SI	BI
PE	0.83					
EE	0.58	0.78				
A	0.29	0.29	0.88			
FC	0.38	0.36	0.32	0.91		
SI	0.17	0.18	0.22	0.19	0.91	
BI	0.77	0.69	0.30	0.42	0.22	0.86

Inter variable correlation and the square root of Average Variance Extracted on the diagonal.

4 Results and Discussion

As mentioned above, structural equation modeling (SEM) or confirmatory factor analysis (CFA), which is one of the most used multivariate data analysis techniques in information systems (IS) research [30], is used for evaluation of the research model and the proposed hypothesis.

4.1 The Measurement Model

The measurement model fit in this study was estimated by using the common model-fit measures. As presented in Table 3, all the model-fit indices exceed their respective common acceptance levels [31] indicating a good model fit with the data collected.

Table 3. Measurement Model

Fit indices	Recommended Value	Result
χ^2/d.f.	< 3	1.77
GFI (Goodness of Fit Index)	>0.8	0.85
RMSEA (root mean square error of approximation)	< 0.08	0.06
RMR (root mean square residual)	< 0.08	0.03
NFI (normed fit index)	> 0.9	0.96
NNFI (non-normed fit index)	> 0.9	0.98
CFI (comparative fit index)	> 0.9	0.98

4.2 Structural Model Test

Given the satisfactory model fit, standardized path coefficients and t-values of the structural model were studied to evaluate the research hypotheses. The research results of the structural model and discussion are presented as follows:

- The results reveal that performance expectancy is significantly related to behavioral intention (γ=0.69, t=2.29), hence the first hypothesis **(H1)** is supported. It implies that students find web-based learning – LMS useful in their education, improving their performance and also increasing the possibilities of communication with other students and instructors. On the other hand, the relationship between performance expectancy and attitude is not significant (γ=-0.41, t=-0.74), hence, the second hypothesis **(H2)** is not supported. Although H1 and H2 results with some prior related studies contradict and with some are consistent [3, 22, 24, 32], most importantly, they confirm prior findings of Venkatesh et al. [18] who developed the Unified theory of acceptance and use of technology (UTAUT).
- The third hypothesis **(H3)**, namely the relation between effort expectancy and performance expectancy appears to be very significant (γ=0.95, t=8.72), which confirm the findings of Chiu and Wang [33], Sumak et al. [22] and Alrawashdeh et al. [3]. The fourth **(H4)** and the fifth **(H5)** hypothesis are not supported, since the relation between effort expectancy and attitude (γ=0.95, t=1.69) and the relation between effort expectancy and behavioral intention (γ=0.47, t=1.37) are not significant. These results indicate that students are not concerned with the ease of use of the system, rather, as mentioned earlier, the degree to which they believe that using the system will them attain gains in performance. This also in line with the TAM (Technology Acceptance Model), as Davis [34] claims that ease of use, referred to as effort expectancy in UTAUT, may be an antecedent to usefulness referred to as performance expectancy in UTAUT, rather than a parallel, direct determinant of usage.
- Hypothesis six **(H6)**, according to the gained results is not supported (γ=-0.11, t=-1.57), confirms the finding of Venkatesh et. al. [18], that attitude does not have direct effect or it is not a direct determinant of behavioral intention.
- The results from the structural equal modeling reveal that both hypothesis seven **(H7)** (γ=0.28, t=3.82) and hypothesis eight **(H8)** (γ=0.09, t=2.03) are significantly related, hence both hypothesis are supported and the results in line with prior related studies [3, 22, 24], whereas contradicting with findings of Chiu and Wang [33] and Islam [32], who found significant relation between social influence and behavioral intention.
- Hypothesis nine **(H9)**, which is about the relationship between facilitating condition and effort expectancy, is supported as the results show that facilitating condition has a significantly positive effect on effort expectancy (γ=0.77, t=9.21), which confirms the newly detected path found in the study conducted by Alrawashdeh et al.[3]. Hypothesis ten **(H10)**, is not supported as the results show that the relationship of facilitating condition with behavioral intention is not significant (γ=-0.06, t=-0.74), which confirms prior study results and most importantly it confirms the findings of Venkatesh [18], who theorized facilitating condition as a construct not being a direct determinant of behavioral intention, rather, facilitating condition is a determinant of use intention.

5 Conclusion and Research Limitations

5.1 Conclusion

This study, by utilizing UTAUT, as a base theory, by integrating an additional construct ("attitude") and also by adding new construct relationships, identified as significant in extant related studies, this study investigated the student's behavioral intention to accept LMS, with an aim to advance the knowledge about the factors influencing its acceptance.

The results of this study show that performance expectancy and social influence are direct determinants of behavioral intention, whereas effort expectancy is identified as an antecedent of performance expectancy. Facilitating conditions and attitude are not found to have significant relationship with behavioral intention.

From this study's results, several implications can be drawn. First, an important contribution is the use of UTAUT, namely its validation, as a novel model in the field of user acceptance [23] in an educational context, respectively at State University of Tetova, given the fact that there are different institutional cultures and characteristics [12]. Second, the findings implicate that students' perception of the performance expectancy is crucial in fostering their behavioral intention to accept and use a technology, similar to prior findings [10] that the strongest driver of technology use is perceived usefulness, whereas their attitude may not be equally important. Moreover, beside confirming the university's need for adopting-deploying a LMS at State University of Tetova, and also in other universities if such a system is not deployed yet, a key role in promoting the use of technology in teaching and learning, play course instructors – professors, who have to take advantage of the web technologies and other related services to facilitate the students' needs.

5.2 Limitations and Future Work

Investigation of the research problem, construction of the research model, data collection and analysis, and documentation of the results in a limited time frame cause limitations in the scope and size of this study. Respectively, the sample size which is limited to students, to whom the course outlines and course materials were presented through instructors'-authors' personal website. Further on, the fact the adoption of LMS is not only influenced by students, but also a number of other factors such as instructors, IT, university support [5] etc., which need to be addressed in future studies. In addition, due to time limitation, the UTAUT moderating effects such as gender, age, experience and voluntariness of use are not considered in this study, which make an avenue of future research.

References

1. Kritikou, Y., Demestichas, P., Adamopoulou, E., Demestichas, K., Theologou, M., Paradia, M.: User Profile Modeling in the context of web-based learning management systems. Journal of Network and Computer Applications 31 (2008) 603-627
2. Coates, H., James, R., Baldwin, G.: A critical examination of the effects of learning management systems on university teaching and learning. Tertiary Education & Management 11 (2005) 19-36
3. Alrawashdeh, T.A., Muhairat, M.I., Alqatawnah, S.M.: Factors affecting acceptance of web-based training system: Using extended UTAUT and structural equation modeling. Arxiv preprint arXiv:1205.1904 (2012)
4. Wang, M., Vogel, D., Ran, W.: Creating a performance-oriented e-learning environment: A design science approach. Information & Management (2011)
5. Selim, H.M.: Critical success factors for e-learning acceptance: Confirmatory factor models. Computers & Education 49 (2007) 396-413
6. Abdel-Wahab, A.G.: Modeling Students' Intention to Adopt E-learning: A Case from Egypt. The Electronic Journal of Information Systems in Developing Countries 34 (2008)
7. Lee, M.C.: Explaining and predicting users' continuance intention toward e-learning: An extension of the expectation–confirmation model. Computers & Education 54 (2010) 506-516
8. Mabed, M., Koehler, T.: An Empirical Investigation of Students Acceptance of OLAT as an Open Web-Based Learning System in an Egyptian Vocational Education School. International Journal of Web-Based Learning and Teaching Technologies (IJWLTT) 7 (2012) 36-53
9. Lai, J.Y., Ulhas, K.R.: Understanding acceptance of dedicated e-textbook applications for learning: involving Taiwanese university students. The Electronic Library 30 (2012) 1-1
10. Edmunds, R., Thorpe, M., Conole, G.: Student attitudes towards and use of ICT in course study, work and social activity: A technology acceptance model approach. British Journal of Educational Technology 43 (2012) 71-84
11. Chang, C.S., Chen, T.S., Hsu, H.L.: The Implications of Learning Cloud for Education: From the Perspectives of Learners. IEEE (2012) 157-161
12. Babo, R., Azevedo, A.: Higher Education Institutions and Learning Management Systems: Adoption and Standardization. Information Science Reference (2011)
13. Oye, N.D., A. Iahad, N., Ab. Rahim, N.: The history of UTAUT model and its impact on ICT acceptance and usage by academicians. Education and Information Technologies (2012) 1-20
14. Al-Busaidi, K.A., Al-Shihi, H.: Instructors' Acceptance of Learning Management Systems: A Theoretical Framework. Communications of the IBIMA (2010)
15. Wang, Y., Chen, N.S.: Criteria for evaluating synchronous learning management systems: arguments from the distance language classroom. Computer Assisted Language Learning 22 (2009) 1-18
16. Almrashdah, I.A., Sahari, N., Zin, N.A.H.M., Alsmadi, M.: Distance learners acceptance of learning management system. IEEE (2010) 304-309
17. Cavus, N., Momani, A.M.: Computer aided evaluation of learning management systems. Procedia-Social and Behavioral Sciences 1 (2009) 426-430

18. Venkatesh, V., Morris, M., Davis, G., Davis, F., DeLone, W., McLean, E., Jarvis, C., MacKenzie, S., Podsakoff, P., Chin, W.: User acceptance of information technology: Toward a unified view. INFORM MANAGEMENT 27 (2003) 425-478
19. Dwivedi, Y., Wade, M.R., Schneberger, S.L.: Information Systems Theory: Explaining and Predicting Our Digital Society, Vol. 1. Springer Verlag (2011)
20. Lee, M.K.O., Cheung, C.M.K., Chen, Z.: Acceptance of Internet-based learning medium: the role of extrinsic and intrinsic motivation. Information & Management 42 (2005) 1095-1104
21. Van Raaij, E.M., Schepers, J.J.L.: The acceptance and use of a virtual learning environment in China. Computers & Education 50 (2008) 838-852
22. Sumak, B., Polancic, G., Hericko, M.: An empirical study of virtual learning environment adoption using UTAUT. IEEE (2010) 17-22
23. Straub, E.T.: Understanding technology adoption: Theory and future directions for informal learning. Review of Educational Research 79 (2009) 625-649
24. Jong, D., Wang, T.S.: Student acceptance of Web-based learning system. (2009) 533-553
25. Orlikowski, W., Baroudi, J.: Studying information technology in organizations: Research approaches and assumptions. Information systems research 2 (1991) 1-28
26. Yin, R.: Case study research: Design and methods. Sage Pubns, London (2009)
27. Creswell, J.: Research design: Qualitative, quantitative, and mixed methods approaches. Sage Publications, Thousands Oaks (2009)
28. Muijs, D.: Doing quantitative research in education with SPSS. Sage Publications Ltd (2004)
29. Fornell, C., Larcker, D.: Structural equation models with unobservable variables and measurement error: Algebra and statistics. Journal of Marketing Research 18 (1981) 382-388
30. Gefen, D., Straub, D., Boudreau, M.: Structural equation modeling and regression: Guidelines for research practice. Communications of the Association for Information Systems 4 (2000) 7
31. Hair, J., Anderson, R., Tatham, R., Black, W. (eds.): Multivariate data analysis (1998)
32. Islam, A.K.M.N.: Understanding continued usage intention in e-learning context. 24th Bled eConference AIS e-library, Slovenia (2011)
33. Chiu, C.-M., Wang, E.T.G.: Understanding Web-based learning continuance intention: The role of subjective task value. Information & Management 45 (2008) 194-201
34. Davis, F.: Perceived usefulness, perceived ease of use, and user acceptance of information technology. Mis Quarterly 13 (1989) 319-340

KupiKniga.mk: Transforming a Website into a Profitable E-Commerce System Using Assisted Conversions Funnel

Ljupcho Antovski, Goce Armenski

University Ss. Cyril and Methodius,
Faculty of Computer Sciences and Engineering,
Rugjer Boshkovikj 16, 1000 Skopje, Macedonia

{ljupcho.antovski, goce.armenski}@finki.ukim.mk

Abstract. There is a lack of published research data on the goal conversions of the real e-commerce systems and e-commerce users' behavior in Macedonia and the Balkan countries. We present the findings of the 6 months research on the strategies applied on the regional e-commerce online bookstore Kupikniga.mk. Several conversion channels and combination funnels were applied, like paper flyers, TV and news presence, multi-level referral lists, social networks, search engines optimization, AdWords, social advertising, online video advertising, email advertising, newsletters, personalized search and recommendations. Based on the analytical data from 2,000 registered byers for 10,000 offered items, we draw conclusions on the conversion rates of different channels and model the behavior of the average e-commerce user in the region.

Keywords: goals, analytics, conversions, online bookstore, e-commerce, user behavior, supply chain, social networking, search engine optimization, referrals

1 Introduction

There are numerous research publications concerning the e-commerce status in Macedonia and the region 1234. But there is lack of indexed published research providing information regarding real commercial e-commerce websites performance indicators 5. Especially for the following research questions: which conversion channels lead to higher percentage of accomplishments of e-commerce goals, and what is the behavior of the average e-commerce user in Macedonia and the Balkan region?

We applied the research on the e-commerce web site Kupikniga.mk 6. It is an online bookstore that sells books and small goods to customers mainly in Macedonia, the Balkan countries and the Macedonian diaspora around the world. Different approaches and several conversion channels and combination funnels were applied, like paper flyers, TV and news presence, multi-level referral lists, social networks, search engines optimization, AdWords, social advertising, online video advertising,

S. Markovski, M. Gusev (Editors): ICT Innovations 2012, Web Proceedings, ISSN 1857-7288

email advertising, newsletters, personalized search and recommendations. The period of monitoring was 6 months.

This paper is organized in the following sections: section 2 presents the overview of the e-commerce site Kupikniga.mk, section 3 defines the goals that were setup for the website and the completion outcome that was measured during the research period, section 4 presents the different channels and the combinations conversions funnels that were applied and the impact results that were obtained, section 5 present the findings and the discussion.

2 Overview of Kupikniga.mk

Kupikniga.mk is an online bookstore that sells books and small goods to customers in Macedonia, the Balkan region and the regional diaspora around the world. It is an investment of the authors.

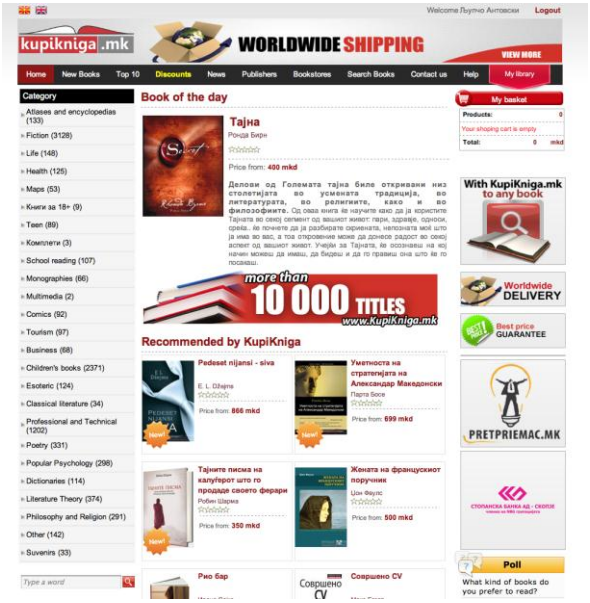


Fig. 1 End users front end of the system

The website is developed from scratch and does not use any third party of-the-shelf non-configurable components. It is constantly upgraded and new features are added on daily bases. Kupikniga.mk includes the front end e-shopping website for the users (Figure 1) and back end administrative system that is a combination of a basic customer relationship management system, e-commerce back end system, finance management system, and supply chain and tracking system (Figure 2).

The books are organized in numerous categories and subcategories. At this moment of writing there are more than 10,000 books and 2,000 registered buyers. All the analysis in this paper is based on this representative number.

The web site implements a rating and recommendations system and several social connectors that enable sharing and expressions for the offered books.

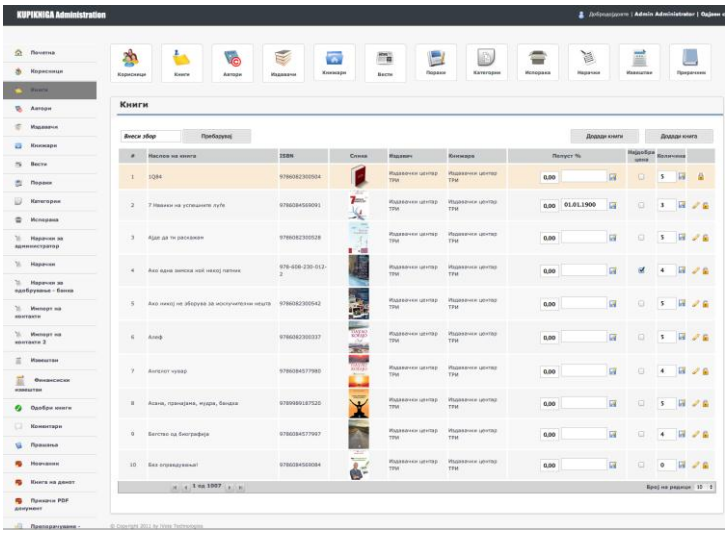


Fig. 2 Back end administrative part of the system

Kupikniga.mk implements an innovative supply chain concept. There are no books and goods kept on stock at Kupikniga.mk. Instead it has a comprehensive tracking and delivery system. The leger of the books is automatically maintained in connection to the status of the publishers' legers, and the ordered books are ordered and shipped directly from the publisher's collection center to the end user. There are special cases when one user orders books from several publishers. In order to minimize the expenses for delivery, the system calculates the optimal route and date for collection of the books from several publishers that are later joined in one package at the Kupikniga.mk fulfillment center and shipped to the end customer.

The second special case is when a user orders books that are not from the country of the end shipment address. Kupikniga.mk has regional partner collection centers in several Balkan countries. To minimize the impact of the transport and import expenses on the end user price, the systems calculated the optimal time when a group order will be shipped from a collection center. There is a minimal order quantity that is covered by the end user price margin. If the number of ordered books is below this quantity, the Kupikniga.mk admin system places automatic orders of additional books. The books are selected by the system using several criteria: the best sellers in the last 10 days, the books that are most visited and socially shared, books with best ratings. It is expected that these books will be as well purchased shortly after the order. With these techniques, the average delivery time for in country ordered books

with normal shipment is 3 days and for priority shipment if ordered during working hours is the next working day. The same periods apply to orders from Macedonian publishers to customers abroad. The above times apply but for start of delivery taking into account the additional time is needed for delivery with normal and priority post. For out of country ordered books, the delivery time to the customer is around 9 days in average.

3 Goals

There are several goals defined and monitored with different conversion channels. The first goal is the number of visits. Since there are new and returning visitors that behave differently, apart from the number of visit, as well the number of visited pages and visit duration is monitored.

The second goal is the registered users. For the website this is a very important goal, since apart from the number of visits these registered users pay to the site, they can as well be targeted with personalized emails, calls and offers.

The ultimate goal of Kupikniga.mk, like for other e-commerce sites 7 is the successful purchase when a client orders and pays. There is a long-lasting argument which goal is more important. At the end, higher number of visitor usually leads to more registered customers and this to more purchases if handled well. But a clear distinction needs to be made between informational websites and e-commerce sites. The main revenue for informational web sites is from commercials that usually refer to more visitors to the site, but for e-commerce sites, the core revenue is from concluded purchases.

4 Multi-Channel Conversions

Our approach in the researched period was to apply different strategies that included use of a specific conversion channel, but also to explore strategies that combined several assisted conversions channels 8. The main goal was to determine the right strategy that gives best results in the context of the geographical location of the customers 910. The following channels were tried:

Hardcopy promotional fliers – the leaflets were distributed to special groups that have interest in book. We covered events like the Skopje Book fair and specific promotional event for books. We tracked the impact with a special promotional code for registration.

Television and press presence - the strategy was to be present on the major television channels and in the press in a period of one week so we could easily measure the impact. Different shows were targeted like morning shows, classical interviews, and press conferences.

Cross-referencing – banners were exchanged with several other sites. The referral traffic was monitored.

Loyalty and referral program - We implement a loyalty program were the regular customers are rewarded with discounts from 3 to 10 %. More a customer spends, the

bigger is the discount for all future purchases. We encourage the customers to invite more friends to join and register on the site. After accomplishing a certain number of registered referrals, the inviting party gets a bigger discount. To encourage the referred user to register, we offer a startup discount for this group of users as well. The registered users are targeted with personalized e-newsletters with latest offers.

Optimization of the search tools and recommendations system – Kupikniga.mk implements the algorithm of collaborative filtering for the search tool on the website and item-to-item collaborative filtering for the recommendation system where when a product is presented, several additional recommendations are given 16.

Social networks – we had two approaches for the social networks. The first one was to create and maintain profiles on the social networks: Facebook, Google +. LinkedIn and Tweeter. The second approach was to attract more visitors with paid advertisements on Facebook. For the second approach several campaigns were tried. We targeted special age groups, with special interests and specific locations.

Search engines – the first approach was to optimize the website and every page for search engines optimization (SEO). Several steps widely published were conducted with linking to other sites and the metadata of every webpage 1718. We also ran paid AdWords 19 campaigns on Google in specific periods to measure the impact of paid searches.

5 Findings and Discussion

We have measured the impact of the conversions channels using the data from the internal administration system of Kupikniga.mk and the analytical tool provided by Google Analytics 20.

The data for the audience of the website shows that there is a high percentage of new visitors 61.72% compared to 38.28% returning visitors. This indicates that in order to complete a purchase, an average e-commerce site in the region needs to have a good navigation in order to lead on the users to make a decision to purchase in the first visit.

Majority of the visitors are from Macedonia with more than 90%, followed by a fair representation from the USA, Serbia, Croatia, Germany, Bulgaria, Slovenia, Italy and other countries. Majority of the site visitors from Macedonia live in the capital Skopje 67,93%, but there is noticeable domination of purchases from other cities in Macedonia compared to Skopje. The reason will be discussed further on in this section.

Most of the visits with 49.01% are with duration of less than 10 seconds. Usually visits last from 1 to 10 minutes in 25,79% of the visits and remaining for other durations. We conclude that a usual e-commerce user in the region need a good impression in the first 10 seconds to stay on the site, and usually the purchase is made in the time span from 1 to 10 minutes.

The dominant operating system in use is Windows with 94.57% followed by Mac OS 1.51%, Linux 1.42%, and other. Dominant browser is Chrome 54.52%, than Firefox 31.58%, Internet Explorer 8.36%, Safari 2.08% and other. At this moment the

average e-commerce users in the region dominantly use Windows and browse on Chrome. All the future services need to be fine-tuned especially for this profile of OS and browser.

When it comes to the access devices, only 2.35% of the users access the site on a mobile device, from which dominant are the Apple products iPhone 28.76% and iPad 22.61%. At this moment the market for e-commerce in Macedonia and the region is not ready for mobile access, but it is an emerging market, since the cost for mobile Internet is getting more affordable in the region.

The traffic on the website is mostly referral from Facebook 51.86%, direct traffic 19.86%, Google organic search 11.62% and insignificant remaining sources. We reason that an ordinary e-commerce user in the region heavily uses social networks and this is a very good channel to refer traffic to the website. When it comes to the AdWords campaigns, they are insignificant compared to the advertisements on Facebook, because there is an estimation of only 170 thousand overall users that use Google search 19, further narrowed with specific search keywords, compared to more than 1.5 million Facebook profiles in Macedonia only 12.

Regarding the conversion rates of the channels presented in section 4, the promotional leaflets did not lead to any goal completion and we could conclude that this channel is not appropriate for marketing activities connected to e-commerce.

The television and press presence had an impact on the goals completion, especially for the number of visits and the registration. In the period of intense presence in the media, there was an increase in the goals completion in the span of 4-9% compared to the periods without media activity.

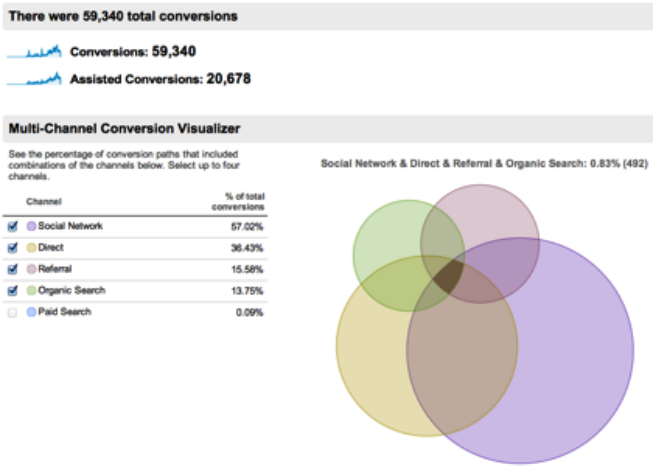


Fig. 3 Multi-channel conversion funnel

The cross referencing and the referral from other local sites had a significant impact on the goals completions. Since these were sites that had normal rate of traffic, this channel will need to be trialed further with referrals from top ranked sites by visits from Macedonia and the region.

The loyalty and referral program increased the completion of the goal for registration nearly 21%, but only when used in a funnel with other promotional activities like social networking, paid search and e-newsletters.

Kupikniga.mk has more than 10,000 items in the offer. There was a significant number of drop-offs because the customers could not find products that do exist in the offer. After implementing the algorithm of collaborative filtering for the search tool on the website and item-to-item collaborative filtering for the recommendation system, the number of drop-offs for searches of products that do exist in the offer has lowered by nearly 36%.

The research showed that the activities thru the channels of social networking and search engines have the highest rate of conversions (Figure 3). From total of 59,340 goal completions, 20,678 were assisted conversion than included several channels before the final step of goal completions. There were 57.02% of goal completions with referral from social networks, with 99.38% coming from Facebook, direct conversion 36.43%, referral from other sources 15.58%, organic search 13.58% and insignificant paid search 0.09%.

The assisted conversion with 34.84% of the goal completions came from the interaction of several channels. The best performing combination were: the combination of social and direct traffic assisted in 10.25% of the assisted conversions, social and referral in 3.96%, direct and referral 6.05%, organic search and direct 4.42%, social with direct and referral 2.97%. The other combinations are with impact of below 1%. Having said this, we can conclude that the social networks are the best performers in assisting goal completions, but in many cases there is a need for assisted conversions funnel with several channels to complete a specific goal.

Two aspects that are important are the time lag and the goal path length. The observed data showed that the e-commerce users in Macedonia are impulsive and buy on the first day of visit with 70.94%. The second significant group is buyers returning and buying after 29 days with 2.81% and all other periods from 1-28 days distributed evenly around 1% each.

The ordinarily e-commerce users in the region are very inpatient to complete a purchase. The completion of a goal is either in one interaction 65.15%, two interactions 13.68%, three interactions 5.32%, four interactions 2.76%, or 12+ interaction 7.20%. All other paths participate with bellow 1%.

Interesting phenomenon was observed with the type of payments. On Kupikniga.mk the users can pay online with a credit card on via bank transfer. Even though this is an e-commerce site, nearly 60% of the users prefer to pay for the purchase via bank transfer, not to use a credit card. There is a huge mistrust that is mainly subjective, since there have not been any major reported abuses of credit cards on e-commerce sites in Macedonia or in the region.

6 Conclusion

In this paper we have presented our observations on the different strategies applied to improve the conversion rate of the specified goals on the website Kupikniga.mk. The observed site had 2,000 registered clients and 10,000 items.

The paper based printed traditional channels underperform. The media still contribute but only with supportive electronic campaign as well.

From the electronic channels, for the specific users that were observed, the best channel to increase the number of conversions is Facebook. Google, especially the paid search underperforms, due to the reason that there is a significant difference in the number of social users and users that search on the web.

Apart from the direct conversion, there are assisted conversion paths that are combination of several conversion channels. The best performer is the combination of social and direct traffic.

The e-commerce users in the region are impulsive, inpatient and complete the purchase either instantly or in the period from one to ten minutes.

Even buying on the Internet, more than 60% of the e-commerce users prefer to pay for the purchase via bank transfer, not to use a credit card. Since electronic payment systems are the main enabler of e-commerce, this phenomenon of low trust for credit card payments needs to be researched further.

References

1. Sekulovska, M.: Internet Business Models for E-Insurance and Conditions in Republic of Macedonia, XI International Conference, Service Sector in Terms of Changing Environment, 27-29 October 2011, Ohrid, Macedonia, doi: http://dx.doi.org/10.1016/j.sbspro.2012.05.016 (2011)
2. Abdullai, B.: The EPS as an e-commerce enabler: The Macedonian perspective, MPRA Paper No. 13996 (2009), doi: http://mpra.ub.uni-muenchen.de/13996/
3. Ribarski, P. and Antovski, Lj., Introducing Strong Authentication for E-Government Services in Macedonia, Proceedings of the ICT-ACT conference, Ohrid, Macedonia (2009)
4. Antovski, Lj. And Gusev, M., M-Government Framework, Proceedings of the First European Conference on Mobile Government, (editors Ibrahim Kushchu and M. Halid Kuscu), 10-12 July, University Sussex, Brighton UK" pp 36-44 (2005)
5. Google Scholar Customized Search, http://scholar.google.com/scholar?hl=en&q=e-commerce++macedonia&btnG=&as_sdt=1%2C5&as_sdtp=
6. Kupikniga.mk, http:// http://www.kupikniga.mk
7. Moe, W., Fade, P.: Dynamic Conversion Behavior at E-Commerce Sites, Journal of Management Science, Volume 50 Issue 3, March 2004, Pages 326 – 335 (2004)
8. Putsis, P., Narasimhan S.: "Buying or Just Browsing? The Duration of Purchase Deliberation," Journal of Marketing Research, 31 (August), 393-402. (1994)
9. Business Intelligence at Debenhams: Case Study, http://www.bluemartini.com/bi
10. Lee, J, et all: Visualization and Analysis of Clickstream Data of Online Stores for Understanding Web Merchandising, Data Mining and Knowledge Discovery, 5(1/2), (2001)
11. Skopje Book Fair, http://www.eragrupa.mk/

12. Facebook, http://www.facebook.com
13. Google+, http://plus.google.com
14. LinkedIn, http://www.linkedin.com
15. Tweeter, http://www.tweeter.com
16. Linden, G. and Smith, B. and York, J.: Amazon.com recommendations: Item-to-item collaborative filtering, Internet Computing, IEEE, 7/1, pp. 76-80 (2003)
17. What SEO isn't Blog, http://blog.v7n.com. June 24, (2006)
18. Burdon, M.: (March 13, 2007). "The Battle Between Search Engine Optimization and Conversion: Who Wins?". Grok.com. (2007)
19. Google AdWords, http://adwords.google.com
20. Google Analytics, http://analytics.google.com

S. Markovski, M. Gusev (Editors): ICT Innovations 2012, Web Proceedings, ISSN 1857-7288

Prediction of Video materials offered to a user in a Video-on-demand system

Zoran Gacovski, Gjorgji Ilievski, Sime Arsenovski

FON University, Bul. Vojvodina, bb, Skopje, Macedonia
zoran.gacovski@fon.edu.mk, gjorgji.ilievski@yahoo.com,
sime.arsenovski@fon.edu.mk

Abstract. Prediction of the customer behavior is a subject that is considered to be "the holy grail" in the business. Data mining techniques are not a new subject, but the amount of data that can be processed by the modern computers and the global market that the world has become has opened a lot of opportunities. This paper considers a method for proposal of video materials to the customers in a video on demand (VOD) system, but its broader usage covers any closed system in which the user is identified before the purchase and history of previous user actions is available. By usingthe data from previous purchases in the systemand applying the well-known Apriori algorithm, a set of association rules is generated. An algorithm that uses the history of the client for which the recommendation should be made, compares it with the association rules found previously and produces the prediction for the best fit videos that will be recommended to the customer. The method is simulated using WEKA for the association rules and using T-SQL procedures and functions for the prediction algorithm. Real data from an existing and publicly available VOD (T-home's MAX TV) system is used for the simulation. The data is put in a relational MS SQL database.

Keywords: Data mining, prediction, Apriori algorithm, association rules, video-on-demand, WEKA.

1 Introduction

Making high-quality predictions about the customer's behavior is very important for any business, for planning, marketing, pricing, product management, human resources, training and almost any subject related to a management of business processes. Data mining techniques have been around for a long time, but the development of the IT infrastructure today allows complex data mining techniques to be implemented for a variety of problems. Even problems that require real time response can be modeled and implemented using data mining techniques. At the same time, the implementation costs are reasonable.

VOD systems are closed environments in which offered videos and the users that access them are known. Making a good prediction about the videos that users will rent is crucial for a good and profitable system. It will help both the sales and the procurement of new videos that will be put in the system. There are a few techniques that can be implemented for some prediction to be made, but most of them require

classification of video materials and/or users in groups [1], [3]. Then, the quality of the prediction is dependent on the chosen classifiers as well as the right classification of the videos and the users in them. Another approach is to use the "market basket analysis" that will produce a set of association rules for the videos. The "baskets" will be created from the history of previous rentals of the users [5], [7]. These baskets will be associated with the basket of the target user. His basket will be created from all the videos he has rented in an interval of time ***t***. This interval will be chosen, having in mind the average time the offered videos are available in the video store.

This approach offers one big advantage, dividing the preparation of the data in a useful format and the process of creation of the association rules in an "offline" phase, and making the association "online" when the user is logged on in the system. This makes the process applicable in practice without burdening the existing process of renting videos. The Apriori algorithm is chosen as the algorithm for the preparation of the association rules in the proposed method in this paper. The confidence and the support of the association rules are chosen as the factors that will make the distinction of the valid and not valid association rules. The data used is extracted from a real existing and publicly available VOD system (T-home's MAX TV), for a period of 3 months. The data is part of a relational database created in MS SQL 2007. The Apriori algorithm for the creation of the association rules is implemented in WEKA.

The main objective of our paper is to offer a simple, straight forward model that makes valid and logical predictions in a VOD environment. The model uses a combination of reliable and proven techniques and it is applicable in practice due to its speed, implementation price, simplicity and clearness. The organization of this paper is as follows. First, we introduce available data and how it is organized as well as the basics for the Apriori algorithm. Next, we define our baskets as sets of items and the parameters that will be used to run the Apriori algorithm. Furthermore we discuss the algorithm used for association of the users' videos with the association rules and providing predicted videos. At the end, results from the simulation are provided. Finally, we pose some of our conclusions.

2 Association Rule Mining

For applying association rule mining, three sets of data from the VOD system are needed: all accounts that have been active in the VOD system, the data for all rentals in the period considered and all videos available in the system. The data is part of a relational database.

We define:

$$I=\{i_1, i_2,\ldots,i_n\}$$

as a set of items, in our case a set of active videos.

$$D=\{T_1, T_2,\ldots,T_m\}$$

is a set of transactions. Every transaction is a set of items (videos) that one user has rented in the considered period of time.

Every transaction T_j, $j=1,..,m$ is a subset of all items I, $T_j \subseteq I$.

A set of *k* items is called *k-itemset*.

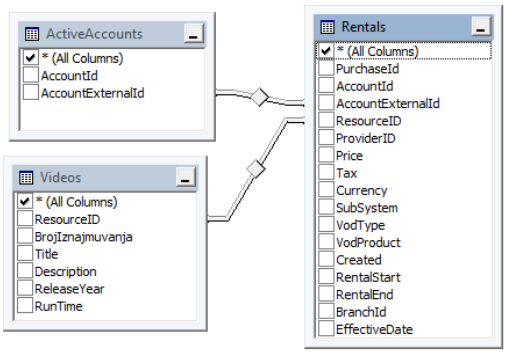


Fig.1. Data relation

Let X and Y are sets of specific items. An implication in form $X=>Y$, where $X \subset I$, $Y \subset I$ *and* $X \cap Y = \emptyset$, is called association rule. X is called antecedent, and Y is called consequent.

We define:

$|X|$ - number of items in X
$|Y|$ - number of items in Y
$|D|$ - number of items in D

To find the significant association rules we use the confidence and the support for every association rule. We say that the item-set X in the transactional set D has support *s,* if *s%* of the transactions in D include X.

$$s(X) = \frac{|X|}{|D|}$$

We say that the rule X=>Y in the transactional set *D,* has support *s,* if *s%* of the transactions in D include all the items in X and in Y $(X \cup Y)$.

$$s(X \Rightarrow Y) = \frac{|X| + |Y|}{|D|}$$

We say that the rule X=>Y in the transactional set *D,* has confidence *c,* if *c%* of the transactions in D that include the set of articles X also include the set of articles Y.

$$c(X \Rightarrow Y) = \frac{s(X \Rightarrow Y)}{s(X)}$$

We find our association rules from the set of transactions D by finding the rules $X=>Y$ whose support is larger than the previously selected minimum, called *minsupp* and whose confidence is larger than a previously selected minimum called *minconf*. A set of items is called large itemset if its support is larger than *minsupp*.

As an example, we need to find association rules in form:

$video_{1x}$, ...,$video_{nx}$ => $video_{1y}$,...,$video_{ny}$

We will read this association rule: The users that have rented videos: $video_{1x}$, ...,$video_{nx}$ have also rented videos: $video_{1y}$,...,$video_{ny}$. If the support of this rule is 0.3, it means that 30% of all transactions include the videos: $video_{1x}$,...,$video_{nx}$ and videos $video_{1y}$,...,$video_{ny}$. The confidence of this rule is 0.8, if 80% the transaction that include $video_{1x}$, ..., $video_{nx}$ also include $video_{1y}$,...,$video_{ny}$.

To find all available association rules, we've used the Apriori algorithm [1].

```
L1 = Large 1-itemsets
for (k = 2 ; Lk ≠Ø; k + +)
{
Ck = Apriori_Gen(Lk-1)
forall transactions t ∈ D
        {
Ct= subset(Ck , t)
forall candidates c ∈ Ct
                {
c.count + +
                }
              Lk = {c ∈ Ck|c. count ≥ minsupp ∗ |D|}
        }
}
return ∪k Lk
```

Fig. 2. Pseudo code of Apriori algorithm.

```
insert intoCk
selectp.item1, p.item2, …, p.itemk- 1, q.itemk- 1
from Lk- 1 p, Lk- 1q
wherep.item1 = q.item1,…, p.itemk-2 = q.itemk- 2, p.itemk-1 <
q.itemk- 1

forallitemsets c ∈Ckdo
    forall(k – 1)-subsets s of c do
          if (s ∉Lk- 1) then
                delete c from Ck;
```

Fig. 3. Pseudo code of Apriori_Gen algorithm.

By selecting different values for *minsupp* and *minconf*, different number of association rules is generated. If these values are smaller the number of association rules rises but the validity of these rules loses strength and vice versa. Choosing the appropriate values for *minsupp* and *minconf* is crucial for making a valid prediction.

The values can be adjusted by making statistical investigation when the model is applied for a longer period of time.

As discussed previously, we define all videos that one user has rented in a specific period of time as one transaction e.g. as one "basket". When a particular user logs into the VOD system, his/hers basket is compared to the baskets of the rest of the users by finding the association rules that are the best fit for this user. These association rules will give the videos that will be presented to the user. If the user data cannot be associated to any of the association rules, the overall most popular videos are presented to the user.

The data for these "transactions" is then formatted to be applied to the Apriori algorithm.

3 Predictions made by associations rules

To find the videos that will be best suited for a particular user, his/her history of rentals has to be compared to the association rules that were "mined" with the Apriori algorithm. An algorithm is proposed that should do the job. Because there might be users that have no history of rentals or their rented videos cannot be fitted in any of the association rules, a list of three most popular videos is prepared. If there is such a case, this user will be given a list of these 3 most popular videos.

The algorithm will present the user 3 videos in any case. All of these can be calculated from the association rules, or if the association rules cannot give 3 videos, the number will be populated from the most popular videos.

The popularity $p_i$of a video is calculated as a quotient from the total number of times the video was rented b_i and the total number of days the video is available in the system d_i.

$$p_i = \frac{b_i}{d_i}$$

We note N as the set of all videos that a user has rented in a period of time t. If N=∅, then the user will be presented with the 3 most popular videos: p_a, p_b and p_c. The set of the 3 videos is noted as P.

$p_a = max(\{b_i : i=1,..,n\})$
$p_b = max(\{b_i : i=1,..,n\} \setminus \{p_a\})$
$p_c = max(\{b_i : i=1,..,n\} \setminus \{p_a, p_b\})$

If $N \neq \emptyset$, then the user has his "own" basket of videos. This basket is compared with the association rules of type X=>Y. The algorithm starts from the rule with the highest confidence and continues until it reaches the rule with the lowest confidence, or while it finds 3 resulting videos. Starting from the first association rule the algorithm checks if $X \subseteq N$. If so, the videos from Y are the potential resulting videos. If some of the videos in Y are already members of the set N, then these videos are disregarded. We calculate:

$Y'=Y\backslash N$

If the number of videos in Y', noted as $|Y'|$ is larger or equal to 3, ($|Y'|\geq 3$), then the algorithm finishes and the first 3 videos of Y' are presented to the user as the predicted videos. If $|Y'|<3$, then the videos in Y' are presented as resulting videos and the algorithm continues to the next association rule, until it finds a total of 3 videos, or there are no more association rules. As it was mentioned previously, if there are no 3 resulting videos from the association rules, the rest of the videos are found from the list of the 3 most popular videos.

The pseudo code of the algorithm that compares the user's "basket" with the association rules is given below. The proposed algorithm can be divided in 2 parts – one that can be performed "offline" and one that can be performed "online". In real systems, when the user logs on to the VOD system, he/she should get the resulting predicted videos in a reasonable time. For this to happen, the association rules generation (call of the Apriori procedure in the pseudo code) as well as the calculation of the table of the most popular videos (FindMax function) can be done in the background.

These operations can be done when the VOD system has available resources. The "basket" of every user can also be prepared in this "offline" mode. The "online" part is performed when the user logs in the system. Then, only the search through the previously calculated association rules and the prepared most popular videos can be performed.

```
PROCEDURE comparison_algorithm

INPUT VALUES
minsupp, minconf;  // constants, numbers of type REAL
varv(1,n);  //vector of n videos available in the VOD
varb(1,n);  //vector of rentals in the VOD system
vard(1,n); // number of days a video is available

OUTPUT VALUES pa, pb, pc; //predicted videos

begin
 for i=1 to n
  p(i)=b(i)/d(i);
  i++;
next i;

 pa =FindMax(p(1,n));//find the video with maximum
rentals
 pb =FindMax (p(1,n)\{pa});//find the second max rented
video
 pc=FindMax (p(1,n)\{pa, pb});//find the third max
rented video

 Call Apriori(in:TRANSACTIONS, minsupp, minconf; out:
```

```
X_i=>Y_i; supp_i; conf_i; k);
  //generation of assoc.rules with Apriori algorithm with
support larger than
  //minsupp and confidence larger than minconf,
  //ordered descending by the confidence, that returns
  //the rules with their support and confidence and the
// total number of rules k.

  var N(0,r);  // r videos rented by user in interval t.
  var f=0;  //number of found videos. = 0 on the start
  var vid(1,5);  // finds max. 5 videos in a search

  if N is NOT NULL
  for i=1 to k // k - number of associat. rules generated
  if f<3 then
  ifX(i)⊆N then
  Y(i)=Y(i)\N; //eliminate videos that were rented before
        if|Y(i)|≥3then f=f+1;
  vid(f)= |Y(i)|(1);f=f+1;vid(f)= |Y(i)|(2); f=f+1;
  vid(f)= |Y(i)|(3);
  end if;
  if |Y(i)|=2 then f=f+1; vid(f)=|Y(i)|(1);f=f+1;
vid(f)=|Y(i)|(2);end if;
  if |Y(i)|=1then f=f+1; vid(f)=|Y(i)|(1);end if;
  end if;
  end if;
  i++;
  next i;
  if f>2 then vid=f(1); pb= vid(2); pc= vid(3); end if;
  if f=2 then pa= vid(1); pb= vid(2); end if;
  if f=1 then pa= vid(1);
  end if;
   end if;
return pa, pb, pc;       //return 3 videos
end.
```

Fig. 4 Pseudo code of the algorithm that compares the user data with the association rules and returns predicted videos.

The proposed model has a cycle of six steps given in figure 5. The pseudo code above is an implementation of the fifth step. The model can be calibrated by performing a statistical analysis on the videos rented from the proposals, as well as the total videos rented with different values for *minsupp* and *minconf*. When satisfactory results are gained, they can be used for valid prediction making. In practice, the calibration process can be done on a regular basis.

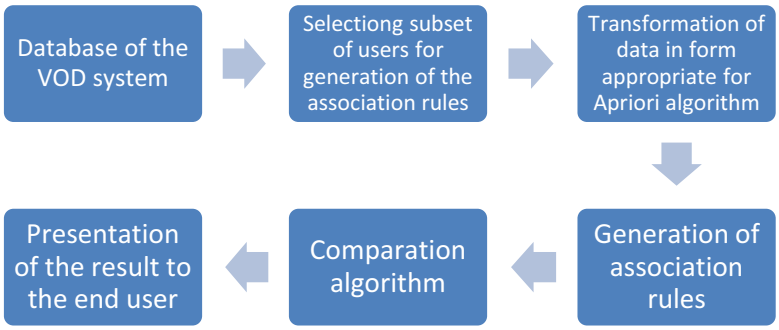


Fig. 5 Steps of the prediction process.

4 Simulation Results

The data used in the simulation is extracted from an existing and publicly available VOD system. The system is based on Microsoft Mediaroom 1.1 platform. The data is for the period of 3 months. It is a part of a relational MS SQL 2007 database.

We have developed an application in Microsoft SQL 2007 that implements the algorithm that compares the users' data with the association rules and returns predicted videos. The application consists of 2 stored procedures and 1 function. It gives results for the real data for the users and their activity in the VOD system.

The association rules are generated using WEKA. Different sets of association rules are produced by supplying different values for the *minsupp* and *minconf* values. The association rules are imported in new tables of the MS SQL database.

The simulation method is consisted of 6 steps:

Step1: Extracting the data for the users, videos and rentals in a relational database. A period of 3 months is used. 7576 users were active in this period. 161 videos were available for rental. Total of 26077 rentals were performed in this period of time.

Step 2: A subset of users is selected as a valid group whose rentals will be used for the generation of the association rules. All users that have rented 10 or more videos in the selected period of 3 months were selected in this group. There are 502 such accounts.

Step3: The data from the rentals is prepared in a format suitable for performing the Apriori Algorithm.

Step 4: Association rules are generated using the Apriori algorithm.

Step 5: Running the algorithm that compares the users' sets of rented videos with the association rules.

Step 6: Prediction of 3 videos for every one of the 7576 users is done and written in a results table.

	Iteration 1	Iteration 2	Iteration 3	Iteration 4	Iteration 5
Minconf	0,4	0,4	0,3	0,7	0,5
Minsupp	0,1	0,08	0,08	0,1	0,2
No. of Assoc.rules (AR)	156	333	487	18	5
No. of users for which results cannot be found from the Association rules	2289	1488	1487	6922	5809
% of users for which results cannot be found from the Association rules	30,17 %	19,64 %	19,63 %	91,37%	76,68%
No. of vids found with Assoc. rules	13802	16564	17284	753	1767
% of videos found with the Associat. rules	60,73 %	72,88 %	76,05 %	3,31%	7,77%
No. of users for which 1 result is found from the Assoc. rules	829	619	397	559	1767
% of users for which 1 result is found from the Assoc. rules	10,94 %	8,17%	5,24%	7,38%	23,32%
No. of users for which 2 results are found from the Assoc. rules	410	462	189	91	0
% of users for which 2 results are found from the Associat. rules	5,41%	6,1%	2,49%	1,2%	0%
No. of users for which 3 results are found from the Assoc. rules	4051	5007	5503	4	0
% of users for which 3 results are found from ARs	53,47	66,09 %	72,64 %	0,05%	0%

Fig. 6 Table of results of the significant iterations.

Over 100 iterations with different values for *minconf* and *minsupp* were conducted. Five iterations were chosen for this paper as the most relevant, to show the prediction percentage based on the association rules. The results from the 5 iterations are given in the table below. By analyzing the iterations' results, it is obvious that the values of *minsupp* that is around 0,1 and *minconf* - around 0.4 will give the most logical results. If these values are higher, the number of predicted videos is very low. On the other hand, by lowering these values the number of association rules and predicted videos is rising, but the strength of the prediction is dropping.

In any real case scenario, these values can be calibrated and the most appropriate values can be found by analyzing the impact that the offered videos will have on the amount of videos rented.

5 Conclusion

In this paper we have proposed a method for recommending videos to users that log on in a VOD system. The model is used for the prediction of videos that will be most suited for the logged-in user. The data used was extracted from a real existing and publicly available VOD system (T-home's MAX TV), for a period of 3 months. The basic data mining technology is the association rule generation using the Apriori algorithm. The Apriori algorithm for the creation of the association rules was implemented in WEKA. The system should give the best prediction for the videos that a user can rent according to the previous rentals of that particular user, compared to the rentals of all users in the system.

The model can be calibrated on the fly, by choosing different values for the minimal support and confidence of the association rules. The process can be divided in two parts, one that can be prepared in advance and a second that will do the calculation when the user logs in. This makes it plausible in a real case scenario due

to the fact that the time latency that will be caused with the calculation step is reasonably low. The simulation is realized in WEKA and in MS SQL 2007. The results show that the system will give predictions for a significant number of users.

6 References

1. RakeshAgrawal, RamakrishnanSrikant, „Fast Algorithms for Mining Association Rules", *IBM Almaden Research Center,* 650 Harry Road, San Jose, CA 95120, 1999
2. Mehmet AydinUlas, „Market Basket Analysis for Data Mining", (PhD thesis) Bogazici University, 1999.
3. Sotiris Kotsiantis, Dimitris Kanellopoulos, „Association Rules Mining: A Recent Overview", *GESTS Intern. Trans. on Computer Science and Engineering*, Vol.32 (1), 2006, pp. 71-82, 2006.
4. R. Agrawal, H. Mannila, R. Srikant, H. Toivonen& A.I. Verkamo, „Fast discovery of association rules", *Advan. in Knowledge Discovery and Data Mining*, pp. 307 - 328, 1996.
5. YaseminBoztuğ, Lutz Hildebrandt, „A Market Basket Analysis Conducted with a Multivariate Logit Model", *Schmalenbach Business Review* (sbr), Vol. 60(4), pp. 400-422, 2005.
6. Sally Jo Cunningham, Eibe Frank, „Market Basket Analysis of Library Circulation Data", Proc. of *6th International Conference on Neural Information Processing*, vol. II, *Perth, Australia*, pp. 825-830, 1999.
7. LuísCavique, „A Scalable Algorithm for the Market Basket Analysis", *Journal of User Modeling and User-Adapted Interaction*, Vol. 19 Issue 1-2, February 2009.
8. E.García, C.Romero, S.Ventura, C.de Castro, „An architecture for making recommendations to courseware authors using association rule mining and collaborative filtering", *Jour. of User Modeling and User-Adapted Interaction*, Vol. 19 Issue 1-2, 2009.
9. Troy Raeder, Nitesh V. Chawla, „Market Basket Analysis with Networks", *Social Networks Analysis and Modeling Journal*, vol. 1, No. 2, pp. 97-113, 2010.
10. Ian H. Witten &Eibe Frank, „Data Mining Practical Machine Learning Tool and Techniques", second edition, Morgan Kaufmann Publishers, 2005.
11. Xiaoyuan Su and Taghi M. Khoshgoftaar, „A Survey of Collaborative Filtering Techniques", *Advances in Artificial Intelligence*, Vol. 2009 (2009), Article 421425, 2009.

Lecture Notes in Computer Science: Scaling Scrum for Large Projects

Naum Puroski [1],

[1] Seavus, 11 Oktomvri 33A, Skopje, Macedonia
{naum.puroski}@gmail.com

Abstract. We all know what the *best is*, but not what *enough is*. We design and implement things aiming at perfection and thus spending more money and effort than real need and real justifications are. This problem is also present in the world of software engineering. In traditional software development processes a significant number of bureaucratic procedures exist that are spending a lot of money and time, leading to insignificant or unnecessary quality. Agilists are trying to solve this problem by creating lightweight development processes with minimal set of actions needed for successful project completion. In the past ten years, agile methodologies are successfully applied for small and medium projects, but there is still mistrust when it comes to large and critical projects. This paper summarizes the most common problems of Scrum agile process for large projects and gives guidance how Scrum process should be scaled in order to be more applicable for them.

Keywords: Scrum, Agile, Large Projects, Software Development Processes.

1 Introduction

Traditional or heavy software development processes are used by software teams more than 60 years. In their life they clashed with various problems; problems were analyzed, addressed and processes were upgraded in order same problems to not recur anymore. In this way over time, mature processes were created which are very accurate and resistant to bugs, but on the other hand they are bureaucratized, with high cost, slow start and long duration for development, fixed scope, heavy documentation and poor communication between people.

Nowadays software companies run into frequent changes in business demands and pressure to come to market as quickly as possible. In such dynamic environments traditional methodologies are not suitable because as previously mentioned in these methodologies requirements are fixed, start slow, and changes very easily lead to failure of the project [1]. Agile methodologies as an alternative to traditional methodologies, eliminate many of these weaknesses because they are adaptive rather than predictive; they rely more on people rather than process; they are open for changes and current needs of the client; they rely on code and communication rather than documentation; they implement the simplest possible solution instead solution which may never be used.

S. Markovski, M. Gusev (Editors): ICT Innovations 2012, Web Proceedings, ISSN 1857-7288

Agile methodologies in the past ten years are proved as successful approach for small and medium projects [10], but despite on these successful stories it is still difficult for software companies to decide to use them for large or critical projects. The biggest reasons why companies stay away from agile methodologies for large projects are difficulties in managing of large teams with this approach [2, 7, 4] and the lack of initial architecture, design and documentation [4, 5]. However, most experts agree that agile and traditional approaches are philosophically compatible [6, 3], even some of the methodologies (like Scrum [8]) give directions how to scale process in order to be more applicable for large projects. On one conference Ken Schwaber, founder of the Scrum methodology, was asked "Are there projects that are not meant to be Scrum projects?" on this he replied "Don't use Scrum if you want to ignore your impediments and if you don't want to know that your project may fail in an early stage" [9]. This kind of analyzes and activities by the experts and the agilists gave courage to people from industry to try agile methodologies for large projects.

Minimal use of agile methodologies for large projects and insufficient literature on this topic motivated me to analyze and examine whether the mistrust of companies for using agile methodologies, i.e. Scrum as their representative, for large projects is justified. I performed this analysis as part of my master thesis [11] and conclusions from my master thesis are briefly presented in this paper.

2 Top 10 Problems in Scrum for Large Projects

1. Although there is a general mistrust of companies for using agile methodologies for large project, there are companies who applied Scrum for large projects and shared their experience in papers or web articles. Analyzing this shared experience, taking in consideration the existing critics about agile methodologies and my previous professional experience with Scrum, I derived the most crucial and critic problems with which companies were faced when they tried Scrum for large projects and I categorized them based of their root problem:*Lack of initial solution design*. Scrum does not invest in creating of the initial architecture and design because in the early start of the project all requirements for the system are not well known and on the other hand as drafted they are expected to change. But anyway, initial design is essential for large projects. Lack of initial design leads to production of non-scalable, rigid, fragile and immobile code which is hard for maintenance i.e. each modification requires major changes in late stages of implementation. Also the lack of initial design complicates the process of creating independent teams that would work for implementation of isolated modules.

2. *Scaling of development teams*. The lack of a complete scope and design of the system disables proper sizing of the project teams at the beginning of the project. This is why agile methodologies have ad-hoc strategy for creation of new teams. Such a strategy leads to creation of teams that strongly depend on each other and which are not effective because most of the time they spend communicating with each other and correlating with other teams. On the other hand agile teams rely on people who are able to complete all types of tasks, from collecting requirements to testing. If we consider the required type and number (over 100) of such a people needed for large

projects, finding them is a real challenge. Impossibility of finding a sufficient number of this type of people leads to the need to introduce people with specific skills (technical writers, testers, GUI designers, etc.) in agile teams and their proper integration into them.

3. *Not effective face to face communication.* Face to face communication is effective in cases where the information should be shared with one or two people, but in cases when information should be shared with 10+ people this type of communication is extremely ineffective. In such situations there is a need for more documentation, not less.

4. *Unavailability or inconsistency of the client*. Client is directly involved in the Scrum development process and has a role of product owner. At the beginning of project and during sprint planning meetings he presents his needs in general, and after that during iteration, on daily bases, he defines details about them and gives feedback about completed tasks. This approach is good for small/medium projects, but it is almost impossible for large projects where project owner has to serve more than ten people concurrently. Because of this problem the development teams may be often stalled due to lack of information. If as solution we decide to add more people with role of product owner then we have problem with inconsistent requirements from different sources because there is no central point for verification and validation of customer needs.

5. *Managing and organizing product backlog*. Product backlogs for large projects can reach above 10000 customer stories. This size of product backlog is difficult for analyzing, monitoring and maintaining. On the other hand if we decide to create a one product backlog per module then we may lose the global picture of the system being built.

6. *Losing the big picture*. Creation of user stories in Scrum is a good idea because it allows frequent deliveries and focusing on small problems, but on the other hand it is a huge problem because by decomposing of the general functionalities into user stories we are losing the big picture of the system. Once when the general functionalities are decomposed and development teams start with implementation of the created user stories it is hard to reconstruct the big picture and to see exactly where the progress of the project is. Also user stories force agile teams to be focused on implementation of small functionalities rather on general functionalities. For large projects, top-down rather than bottom-up software development approach is more applicable in order the big picture to be kept and progress to be tracked better.

7. *Losing information about the system*. No investment in documentation leads to loss of information and decision points about developed functionalities. After a while without documentation no one, including developer and client, cannot remember why certain functionality is implemented as it is implemented. This problem grows exponentially with the size and length of the project, and is even greater when people in development teams are changed frequently and the code is not regularly commented or tested with unit tests.

8. *Planning and resolving dependencies***.** Unlike small projects where each requirement is assigned to one team in large projects new requirements depend on many teams. Client requirements need to be decomposed into multiple user stories

and they should be assigned to different teams. Synchronization and resolving dependencies between teams is needed and that presents problem because in Scrum there is no central point that define and assign tasks. Instead of this in Scrum tasks are defined by self-organized teams during sprint planning meeting according to some local priority.

9. *Working with distributed teams*. An inability to find sufficient number of people for large project on one geographical location induces companies to develop projects with distributed teams. Distributed teams have limited ability for communication because of their remoteness and different time zones. Thus violates one of the primary characteristic of agile methodologies and makes agile approach not very suitable for this type of teams. In order to apply agile methodology on project with distributed teams, teams have to be independent as much as possible and with that need for communication between team to be as low as possible.

10. *Change of people's habits and behavior*. Agile is completely new approach for managing of software projects and with that in contrast of traditional methodologies requires different behaving and acting of people. Changing of the mindset, habits and behavior of the people is not easy task, especially for large projects where number of people is large and management should encourage all of them to do something different than they usually do. It takes time and perseverance to train people to use new practices, to show and prove that this new approach gives results. Involvement of experienced Scrum Master into a project is crucial for success of this task.

3 Ten Rules for Successful Scaling Scrum for Large Projects

Based on findings presented in previous chapter and based on my previous professional experience with different software development processes, including Scrum, I am suggesting the following list of rules/practices which should be applied in order Scrum to be more applicable for large projects:

1. *Involvement of experienced Scrum Master*. Experienced Scrum Master's primary role is to encourage people to grow out of existing habits inherited from traditional methodologies and to try to transform hierarchical organized teams into self-organized teams. His experience will also contribute to the education of other team members and their motivation not to be distracted from the real road. From Scrum Master is also expected to assist in properly resolving the problems without adding bureaucratic activities into the process. For large projects, without an experienced Scrum Master the chaos could easily get out of control.

2. *Involvement of business analysts in agile teams*. At least one business analysts should be involved in each team who will act as a proxy to the client. In this way the role of Product Owner is scaled with people who are specialists in specifying software requirements, i.e. people who know best how to extract information from customers and who will best organize and pass on the desired functionalities to the team. Business analysts except as members of the development teams are acting as a special team of business analysts. The purpose of this team is to translate global requirements into user stories and to clarify dependencies and inconsistencies between the teams.

3. *Adding teams with special purpose*. In order the big picture to be kept and to cover tasks that are not in direct responsibility of any single development team, practice is adding of teams with special roles into a project. This kind of teams will cover segments of the software development lifecycle like verification and validation of customer needs; creation of user stories, their prioritization and resolving of dependences; creation of architecture and design; integration of solution and maintenance; system testing; etc. These teams may be composed of members of the existing development teams or by entirely new members.

4. *Involvement of specialists in the agile teams*. Except experienced people with wide range of software engineering knowledge, agile teams should also involve people with some specific field of knowledge. This practice will reduce need for such a large number of people with wide range of knowledge (specialists will perform work for which they are specialized; other will cover all other tasks) and also will improve effectiveness of the team (It is expected that the specialist will perform the specific task faster and with greater quality then others). For example, each agile team may involve GUI designer, QA engineer, technical writer, database developer, etc.

5. *Putting bigger accent on software architecture and design*. At the beginning of a large project we have to invest more in software architecture and design in order developed solution to be more robust, more scalable and more maintainable. Creation of the initial architecture and design will allow easier tracking of big picture, easier tracking of the impact from performed modifications, creating independent teams, including of less experienced people in teams, easier resolving of dependencies and inconsistencies in the system etc.

6. *Minimization of dependencies between teams*. Agile teams must be as independent as possible between each other in order need for communication and synchronization between them to be as low as possible. Ideal situation would be when the need for communication between agile teams would be at most once a week. This can be achieved by initially creating the architecture and design solution, and then assigning implementation of independent components on each team separately.

7. *Modification of the planning process*. In contrast to the small projects where one team is responsible for whole implementation and all tasks are defined and assigned during one sprint planning meeting, for large projects any required modification may have impact on more modules and with that on more teams. Taking into account that each team plans its tasks on separate sprint planning meeting, practices for synchronization of activities between teams is necessary in order modification to be completed at the end of a sprint. As a solution, each modification should be analyzed by team of business analysts and architects (equal on scrum of scrums team), during Scrum-of-Scrums meetings. During these meetings they have to perform impact analysis on required modification, to create user stories for all of the impacted modules and to define priorities for each of impacted teams. Expectations about these stories will be passed to the team by their representative on Scrum-of-Scrums meeting.

8. *Top-down approach for managing of the Product Backlog*. In Scrum general requirements are defined at the beginning of the project and then these requirements are decomposed into smaller user stories (top-down approach). For large projects,

systems used for managing of the product backlog has to allow managing of all level items (general requirements, user stories, tasks,…) and to allow establishment of links between them. Linking between backlog items will allow keeping of the big picture for the project, better tracking of the project scope, project status and dependences between user stories. The system for backlog managing should also allow filtering and grouping of items by components. This functionality allows keeping of all items in one product backlog and setting of different views by defining custom filters.

9. *Creation of minimum required documentation*. In large projects face to face communication is ineffective and therefore there is a need for creation of documentation that will contribute for easier communication between and within teams. In most cases it is practical to create software design document, technical documentation for specific algorithms and protocols used in the system, regularly commenting the code, test plan for system testing or SRS specification (not both because there is a 1:1 relationship between them). Type of documentation which will be created should be defined at beginning of the project and it must be created or updated in same iteration concurrently with the implementation of the required functionality. At the each iteration end we have to have potentially shippable product with completed documentation and code.

10. *Minimal introduction of new practices*. If new practices are introduced for all of the detected problems then agile process will soon become as bureaucratized as traditional ones. Instead of this, each problem should be analyzed during a sprint retrospective meeting, in order the root of the problem to be found and to be disused how, in an agile way, it can be avoided in future. New practices should be added only if the same problem occurs several iterations in a row. Experienced Scrum Master has the crucial role for control of this point.

4 Scaling of Scrum

The suggested rules and practices in previous chapter may be applied in Scrum very easily because Scrum is defined as a framework, not as a classical process where all activities and steps are defined in details. The only prerequisites in Scrum for adding of new practice/activity in the process are: people who suggest change have to understand the problem really; proposed modification must not violate any of the agile principles and at the end Scrum Master have to confirm that proposed modification will give positive results.

For the addressed problems about Scrum for large projects, explained in chapter 2 and following the rules for Scrum scaling, presented in chapter 3, I am suggesting scaling of the Scrum by adding of the these activities (see figure 1):

1. *Collection of requirements and creation of product backlog*. Initially a team of business analysts is created who should conduct the business analysis communicating with the client and the other stakeholders who are going to use the new system. Based on this analysis they create general requirements, after they decompose them into smaller user stories, detect dependencies and inconsistencies between the stories and finally all the stories are entered into the system for managing product backlog. This

stage can last for weeks; but it is practical to be limited on one iteration or maximum one month.

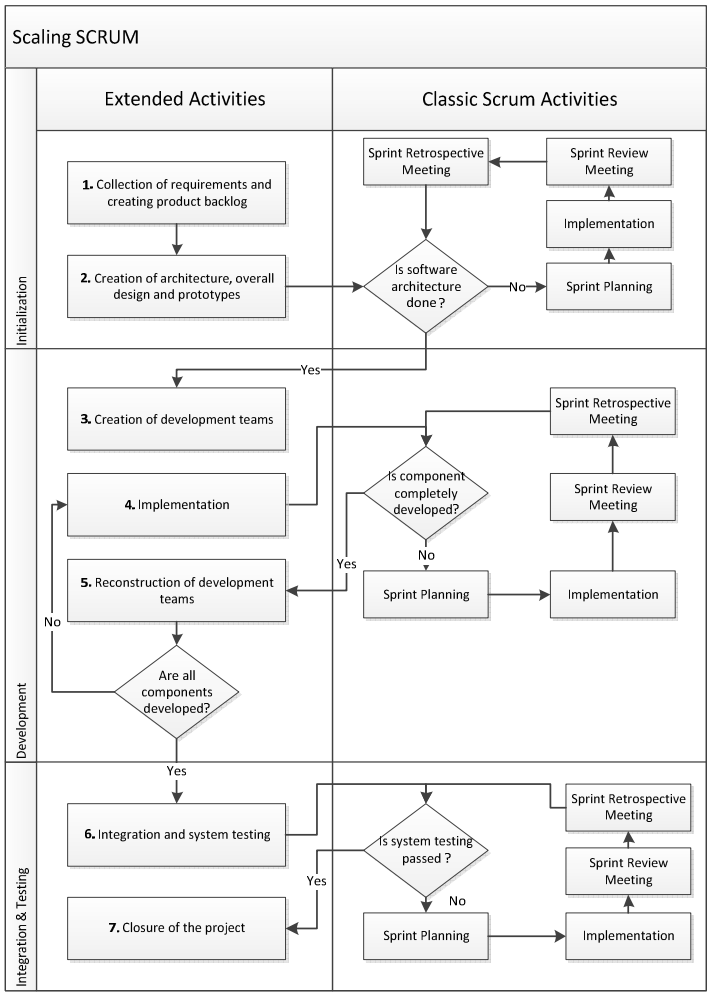


Fig. 1. Each phase of the scaled Scrum process uses classical Scrum process for execution of the team tasks. In the first phase classical Scrum process is used for creation of software architecture and design, in the second phase for implementation and in the third phase for integration and system testing.

2. *Creating the architecture, design and prototype of the solution*. Once the product backlog is created, a team for creation of initial architecture and solution design is established. In this team experienced technical people and business analysts who previously worked on collecting of requirements are included. The goals of business analysts in this stage are transfer of information to technical team and decomposition

of more general user stories into a smaller. Because this phase may last several months technical team, concurrently with software architecture, creates prototypes for proposed solution. Creation of prototypes allows showing something valuable to the client as also the end of iterations. In this way it is easier for the client to check whether the offered solution meets his needs. It is practical the team who works on this task to be not distributed in order communication between team members to be without limitations.

3. *Creating teams*. Teams should be created based on created software architecture teams. Ideally is for each separate component one separate team to be created. The teams created in this way very little depend on each other because they work on different components i.e. modules that can be easily isolated by mock/stub objects. The new teams, beside certain number of people with wide range of skills and certain number of specialists (tester, technical writer ...), have at least one of the team members from the architects and business analysts teams. They have a role to convey the current state, to present to the team into Scrum-of-Scrums meetings and do impact analysis on requested changes. In order to reduce the time needed for Scrum-of-Scrums meetings and increase the efficiency, Scrum-of-Scrums teams can also be defined by the architecture of the solution. One Scrum-of-Scrums team should have only members of teams that directly depend on each other (see figure 2).

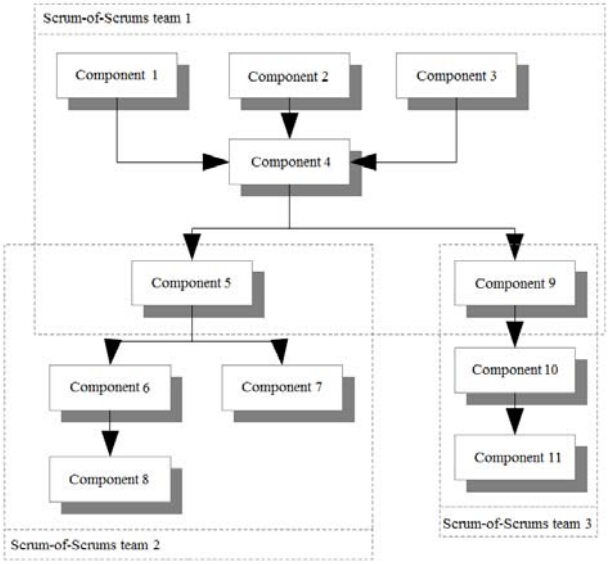


Fig. 2. One Scrum-of-Scrums team is composed of team members from teams which directly or indirectly depend. Team members from boundary teams (for example team 5 and 9) are part of two Scrum-of-Scrum teams and they have responsibility to pass information between two teams.

4. *Implementation*. Created teams implement assigned components using classic Scrum process.

5. *Reconstruction of teams*. Teams are reconstructed as soon as they finish with the implementation of some specific component. The team that has implemented some component may remain to function with smaller number of members if there are expectations for larger changes in the developed component. In opposite case the team is released and some of the members are transferred to the teams for system testing and maintenance. In both cases people who are experts on a certain part of the system are kept, these people will further be responsible for integration of all components in the solution and will be responsible for solving possible detected problems and system changes.

6. *Integration and system testing*. System can be integrated by using any integration approach: top-down, bottom-up, big-bang integration strategy, etc. The way how and when system will be integrated depends on selected integration strategy. In any case integration should be performed by the team that implemented the component which is about to be integrated or by the team members from the maintenance team. After integration is finished, system testing should be performed. This kind of testing should be performed by a special group of testers that can work within or without the maintenance team. This phase may last more than one iteration, so for this phase we also use classic Scrum approach.

7. *Closure of the project*. After integration and stabilization of the project whole team is released, except the maintenance team. This team is responsible for maintaining the system during his lifecycle. Optionally some of the teams may continue to work if they are responsible for some critical part of the system for which more frequent modifications are expected. After closing the product backlog product remains divided by components, but unlike before it will be run by one team.

Conclusion

Scrum methodology in the past 10 years had good results for small and medium projects, but it is still not too often used for large projects. Companies which tried Scrum for large project clashed with various problems as lack of initial design, loss of global picture, the impossibility of finding a sufficient number of “agile” people, spending too much time on communication and information loss. In this article I suggested several rules and practices that should be introduced into a Scrum in order these problems to be avoided or at least reduced. Some of suggested practices are investing more in software architecture, documentation and tools, introducing of the people with specific field of knowledge in agile teams, creation of agile team by some defined criteria, etc. These practices will make Scrum more robust and stabile and with this more applicable for large projects.

S. Markovski, M. Gusev (Editors): ICT Innovations 2012, Web Proceedings, ISSN 1857-7288

References

1. The Standish Group, Chaos Report, http://www.projectsmart.co.uk/docs/chaos-report.pdf, (1995)
2. Cockburn, A., Highsmit, J.: "Agile Software Development: The People Factor", Computer, vol. 34, no. 11, pp. 131--133 (2001)
3. Paulk, M.C.: "Extreme Programming from a CMM Perspective", IEEE Software, vol. 18, no. 6, pp. 19--26 (2001)
4. Boehm, B.: "Get Ready for Agile Methods, with Care", Computer, vol. 35, no. 1, pp. 64--69 (2002)
5. Rasmusson, J.,: "Introducing XP into Greenfield Projects: Lessons Learned", IEEE Software, vol. 20, no. 3, pp. 21--28 (2003)
6. Reifer, D.J.,: "XP and the CMM", IEEE Software, vol. 20, no. 3, pp. 14--15 (2003)
7. Fowler, B.,: "The New Methodology", http://www.martinfowler.com/articles/newMethodology.html (2005)
8. Cohn, M.,: "Advice on Conducting the Scrum of Scrums Meeting", http://www.scrumalliance.org/articles/46-advice-on-conducting-the-scrum-of-scrums-meeting (2007)
9. Smits, H.: "Implementing Scrum in a Distributed Software Development Organization". In Agile Conference (AGILE), pp. 371--375, IEEE Press (2007)
10. Results from the July 2010 State of the IT Union Survey, http://www.ambysoft.com/surveys/stateOfITUnion201007.html (2010)
11. Puroski, N.: "Scrum For Large Projects" ("Scrum за Големи Проекти"), master thesis in process (2012)

A Fuzzy Logic based Controller for Integrated Control of Protected Cultivation

Oliver L. Iliev[1] Pavle Sazdov[1], Ahmad Zakeri2

1. FON University, FICT, Laboratory for Protected Cultivation & Energy Efficiency, Bul. Vojvodina bb, 1000 Skopje, Republic of MACEDONIA
e-mail: Oliver.Iliev@fon.mk, Pavle.Sazdov@fon.mk
2. University of Wolverhampton, School of Engineering and the Built Environment, Wulfruna Street, Wolverhampton WV1 1SB, United Kingdom
e-mail: A.Zakeri2@wlv.ac.uk

Abstract. In terms of systems theory, the greenhouse represents a complex nonlinear system with emphasized subsystem interactions. System decoupling is used in order to obtain simplified control structures for independent control loops. This gives limited results because of the strong interaction between system variables. Such system does not allow system behaviour optimization primarily in terms of energy efficiency and/or water consumption. This paper presents a design of fuzzy logic based controller, which optimizes the Greenhouse heating, energy and water consumption. The design includes the main linguistic variables for sensors and actuators. Membership functions of Fuzzy Inference System (FIS) are generated and simulation and analysis of the behaviour of the designed control system is performed.

Keywords: Fuzzy control, Protected Cultivation, Agriculture, Complex Systems, Energy Efficiency, Water Management, Inference Engines, System Integration

1 INTRODUCTION

Protected Cultivation, as an alternative way for food production, has become more important in recent years due to several factors that change the global picture in the world of agribusiness.
The most relevant factors are:

- Global increase on food prices by 33% in 2010
- Reduced amount of quality water for irrigation
- Increased use of arable land for production of raw materials used for biodiesel
- Increased toxicity of arable land with heavy metals, excessive and/or misuse of fertilizers as well as long-term contamination due to the use of pesticides
- Global climate changes

Global trends show that these conditions will continue to rise in the future. Protected (Greenhouse production) allows for drastic reduction of amount of water for irrigation. So-called Hydroponics or soilless systems address the growing problem of soil pollution, allow increased density of plants per unit area, reduces impact of climate changes, as well as application of bio control as an effective alternative to traditional methods of plant protection.
The main advantage of these systems is the efficient use of water for irrigation. Using rock-wool as a growing substrate, offers the possibility to use water and fertilizers very sparingly. This is especially emphasized in the so called “closed irrigation sys-

S. Markovski, M. Gusev (Editors): ICT Innovations 2012, Web Proceedings, ISSN 1857-7288

tems", where water with fertilizers recirculate through the system and water is lost only through leaf transpiration. The water in rock-wool substrate is fully and easily available to the plant, as opposed to many other substrates used. There is also significant economic impact regarding the possibilities for off season and early season production.
The role of the greenhouse in protected cultivation is to provide optimal microclimate conditions for plant growth. From the systems theory aspect, protected food cultivation in greenhouses represents a complex nonlinear system, including number of subsystems with emphasized variables interdependency. Basically, the main controlled variables are:

Temperature control: The optimal growth condition assumes constant temperature inside the Greenhouse. Disturbance variables which affect the inside greenhouse temperature are; Outside temperature, Relative Humidity (RH), Light Irradiation, Plants growth stage, speed and direction of wind. It should be emphasized that the temperature can be controlled under various operating modes such as daily/night mode and different modes for each stage of plant development. Also, in more advanced systems, besides the inside temperature, the temperature of the plant and the temperature of the substrates are also measured.

Relative Humidity (RH): Air RH should be kept in the appropriate limits depending primarily on the type and stage of the plant growth. High RH levels cause development of bacterial diseases, while low RH causes difficulties in the pollination process and exceeded water loss. Disturbance variables are: Air temperature, Light Irradiation, Plants growth stage which increases RH by transpiration process, intake of fresh air and foliar irrigation. Beside the RH level of the inside air, in more advanced systems, the moisture content of the substrate is also measured.

Lighting: The amount of lighting is of decisive importance for the physiological processes affecting plant growth. Insufficient amount of natural (solar) lighting can be supplemented by so-called HID lamps (High Discharge Lights). Recent studies show substantial progress in usage of LED (Light Emitting Diodes) lighting, which is significant to HID both in terms of energy efficiency and in presetting the specific lighting spectral density corresponding to the two peaks of chlorophyll a and b (420 to 450 nm in blue and 630 to 660nm in red spectrum wavelength).

Irrigation and nutrient solution control: When it comes to irrigation of soilless systems, it is necessary to establish the correct ratio of macro and micro - nutrients, appropriate level of Ph and EC (Electrical Conductivity) or TDS (Total Dissolved Solids). This is opposite to the plant growth in soil, where all nutrient unbalances can be fixed by the soil itself. The control of an open irrigation systems has a relatively simple structure. The only problem refers mainly to the balance of all macro and micro nutrient ions, together with balanced Ph and EC. In the case of closed irrigation systems, all unused water is collected and reused, as opposed to the open irrigation systems, where usage of water for irrigation is 15 to 20% increased. In the closed irrigation systems water is reused, but in every irrigation cycle, nutrient solution should be rebalanced due to different absorption rate of nutrients. In both cases dis-

turbance variables are: Air temperature, RH, lighting and Plant growth stage. According to [6], only in recent years concrete efforts have been made to place problems particular to water loss in greenhouses. The vital relation between water supply and demand has not been adequately studied for greenhouse practice. Principal investigators have commented the limits of empirically obtained data [15] [18], using statistically derived relationships that have little relationship with physical principles.

Carbon Dioxide (CO2): The amount of carbon dioxide in the atmosphere is essential for plant growth. The natural concentration of CO2 is about 350 ppm, and given the limited space, this amount of CO2 can be absorbed within a few hours. Also, it has been proved that additional concentration of CO2 can significantly increase yields, and positively affect the shelf life of fruits. Here, the disturbance variables are Air temperature, RH, Light, and Plans growth stage.

2 GREENHOUSE CONTROL

The problem of Optimal Control of Greenhause can be defined as achieving the control trajectory

$$\vec{u}(t), t_0 \le t \le t_f \tag{1}$$

which minimize the predefined goal function

$$J(\vec{u}(t), t_0, t_f) = \Phi(\vec{x}(t_f), t_f) + \int_{\tau=t_0}^{\tau=t_f} L(\vec{z}(\tau) d\tau \tag{2}$$

for the dynamic system model with the initial state

$$\vec{x}(t) = \vec{f}(\vec{x}(t), \vec{u}(t), \vec{d}(t)), \vec{x}(t_0) = \vec{x}_0 \tag{3}$$

with output accessible for observation

$$\vec{y}(t) = \vec{g}(\vec{x}(t), \vec{u}(t), \vec{d}(t)) \tag{4}$$

And additional output representing auxiliary variables of interest

$$\vec{z}(t) = \vec{h}(\vec{x}(t), \vec{u}(t), \vec{d}(t)) \tag{5}$$

Where d(t) represents external input variables which are measurable but not accessible for control.

System complexity and its nonlinear nature caused numerous researches to try and find a way to simplify the control structure in greenhouse production, or at least to automate part of it. The reasons for this approach, is the high cost of integrated control systems and the lack of a general system model that will cover important con-

trolled and disturbance variables. Decoupling of the control system typically includes three different aspects:
Greenhouse atmosphere control includes:

- Inside Air Temperature; Plant Temperature; Substrate Temperature; Relative Atmospheric Humidity (RH); Substrate Moisture content; Carbon Dioxide (CO2)
- Irrigation and Fertilization Control includes: Amount of water per plant per hour; Balancing of macro and micro nutrients; Ph levels; Electric Conductivity (EC); UV Water disinfection (for closed irrigation systems)
- Lightning Control includes: Daylight Intensity; Additional HID Light; Additional LED Light for photosynthesis spectral balancing

It should be noted that this approach gives only partial results because of emphasized subsystems variables interaction. Lighting proportionally affects the plants transpiration, and thus the amount of irrigation water; Leaf water transpiration increase RH in the atmosphere; The RH is inversely dependent on the temperature; it is meaningless to activate CO2 dosing system when windows are open for ventilation, and etc.
It is important to note that these systems are quite energy demanding, and a particular aspect of control design should be their energy efficiency. Besides the optimal value of the controlled variables, two additional threshold values must be declared, and all controlled variables must stay in these boundaries.

3 CURRENT SYSTEMS FOR AUTOMATIC CONTROL OF GREENHOUSES

All growing phases can be controlled through Control of Air Temperature, Relative Humidity, CO2, Irradiation and Irrigation [20]. Good overview can be found in [18]. Common control systems for automatic control consist of sensor network for data acquisition connected to the central computer system through adequate communication protocols. Based on obtained data, and adequate algorithms, different actuators (motors, heat pumps, coolers, HID lights, etc.) can be activated in order to keep the measured variables in optimal range. Also, data from sensors and actuators is recorded in log files. Usually, GUI is used to display this data and to provide more optimal control of measured variables.
There are different approaches in designing control systems according to their complexity, control algorithms used, and number of controlled parameters.

Timing Control: The simplest system used today is “Timing Control” system, where simple timers are used to manage actuators. This is open loop control system and requires high level of expert kno-wledge from the growers. Also, possibility for mistake is very high and requires continuous supervision by the grower.

ON/OFF Control: This control design is based on simple feedback loops where the main goal is to keep desired variable in certain limits. The main advantage of this design is simplicity, they are inexpen-sive and reliable. But, this control strategy does not encompass strong interaction between variables (for example, influence of fogging over temperature drop, or air heating over the drop of RH).

PID Control: PID Control systems overcome some of the disadvantages that the ON/OFF control has, but adjusting the parameters (P-proportional, I-Integrative and D-derivative) is based on system transfer function, which represents a problem with this type of control systems. Currently, PID control is usually applied in systems for nutrient solutions.

4 FUZZY LOGIC BASED CONTROL DESIGN

Fuzzy logic [22], [23], [11], [12] is mathematical theory dealing with uncertainty. This approach is widely used in modelling nonlinear systems with high complexity, plant dynamics is unknown or it can change rapidly. This approach is intuitive, input and output variables are linguistically described, and design of control algorithm is primarily based on if-then-else rules.

Fuzzy Logic Controllers are widely used in different engineering areas [21], [8], [9], [10], [7] including AI, Expert systems, Robotics and Biotechnology. There are few researches in applying this promising method into control of greenhouses [2], [5], [14].

The main unit of the Fuzzy Logic Controller is Fuzzy Inference system (FIS). The FIS consist of five processing parts:

- Fuzzification interface which generates linguistic variables based on crisp data inputs from sensor subsystem
- Defuzzification interface which generates crisp control output to the actuators
- Decision making unit, based on predefined control logic, generates inference operations
- Database process provides the fuzzy sets and membership functions used in fuzzy rules
- Rule base unit consisting of an adequate number of fuzzy rules

In the presented system input (sensor) variables are: Indoor Air temperature (IAT) in °C. , Relative Humidity (RH) presented in %, level of Carbon Dioxide (CO2) inside the greenhouse presented in ppm, stage of plant growth in days, and Light Intensity (Lux).

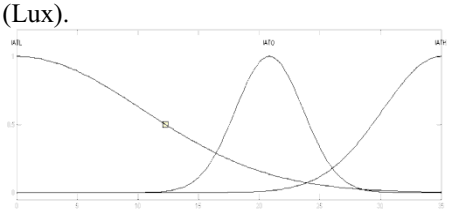


Fig.1Temp.Membership function (ºC)

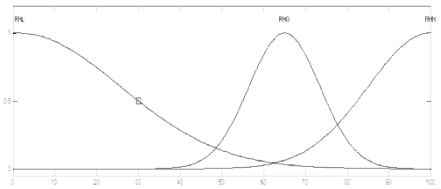


Fig.2 RH Membership function (%)

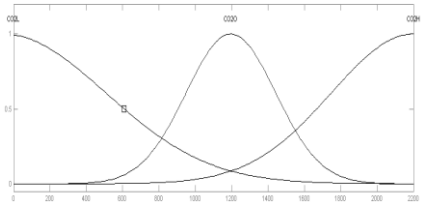


Fig.3 CO2 Membership function (ppm)

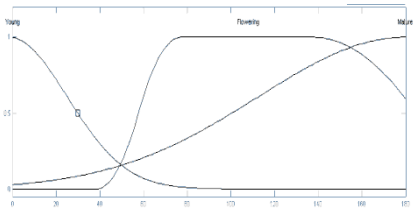


Fig.4 Plant Growth Stage (days)

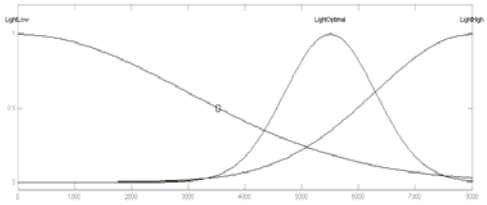


Fig.5 Light Intensity (Lux)

Output (Actuator) variables are: Heating system which can be activated in linear working regime from 0 to 100 %,, Windows position on the top of the greenhouse (closed - 0% full open -100%), CO2 dosing system (0-100%) and Irrigation system with irrigation time of 0 to 200 sec. (This assumption is made for 4L/hour drip irrigation system which is equal to 1.1 mL/sec)

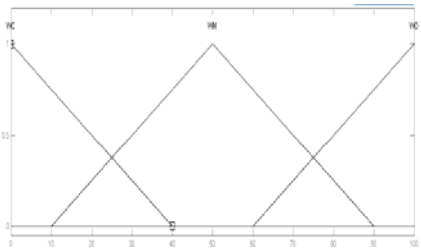


Fig.6. Heating, Window and CO2 membership function (%)

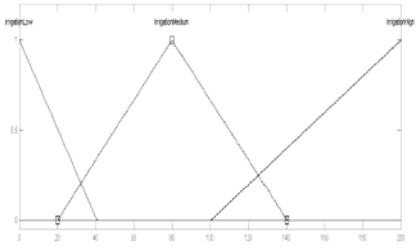


Fig.7. Irrigation membership function (sec.)

Presented FIS for Greenhouse control executes three actions, First process of fuzzification (conversion of crisp values from sensors) into linguistic variables within predefined fuzzy sets. Then, rule base unit based from the knowledge database generates control strategy, and third action is defuzzification where crisp output are generated for actuators.

On Fig.8 is presented the membership function of the Ventilation subsystem activity depending on measured Air Temperature and RH. This is 2-dimensional function which define the status of windows on the greenhouse roof (0%- closed 100% full open) as a membership function depending of air temperature inside the greenhouse and Relative Humidity.

On Fig.9 is presented membership function of CO2 Dosing subsystem depending of the measured level of CO2 and temperature in the greenhouse.

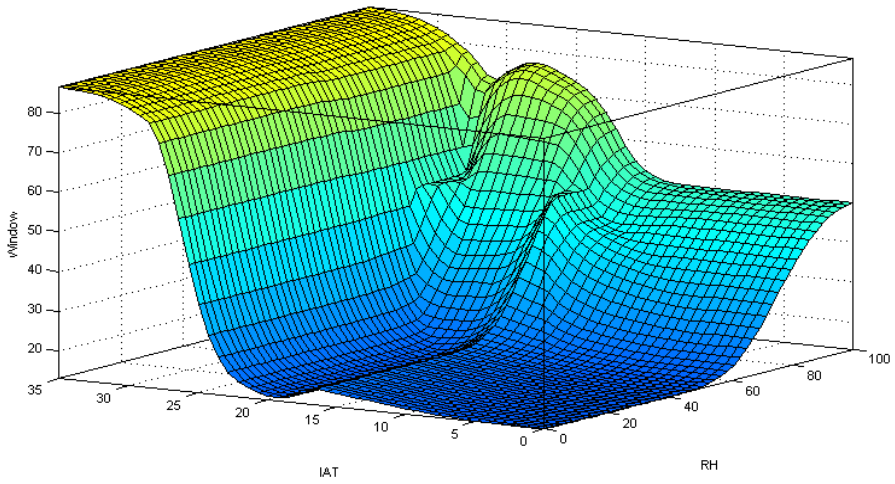


Fig.8 Ventilation subsystem activity depending on measured Air Temperature and RH

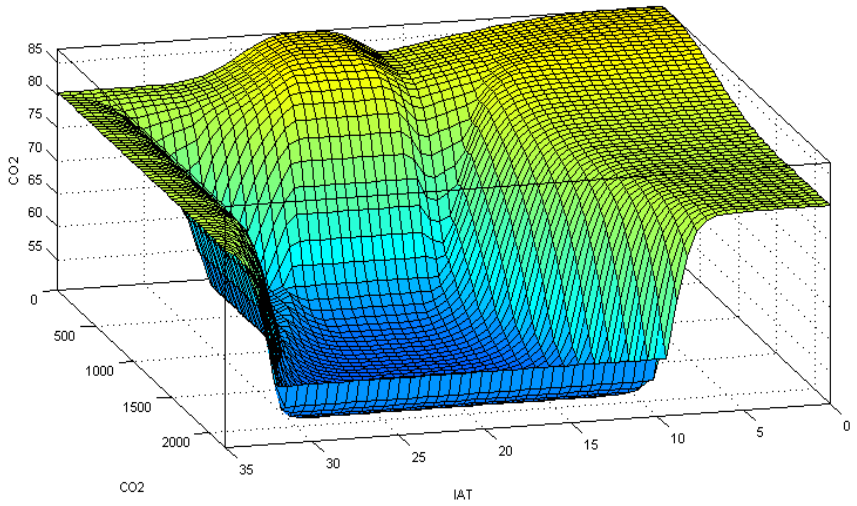


Fig.9 CO2 subsystem activity depending on measured Air Temperature and current level of CO2

On Fig.10 is presented membership function of Irrigation subsystem activity depending on measured Light Level (PAR) and Growth stage of the plant.

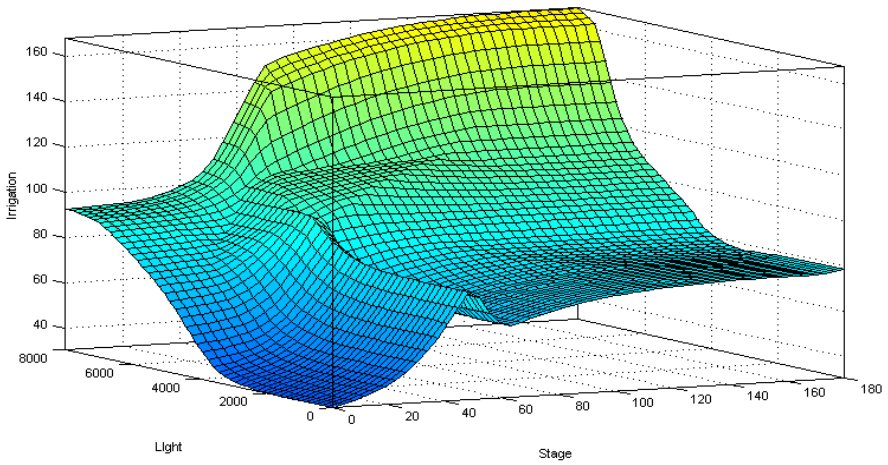


Fig.10 Irrigation subsystem activity depending on measured Light Level (PAR) and Growth stage of the plant.

Next step in the designing process is simulation of the obtained FIS. Obtained crisp outputs for actuator devices from the simulation has been studied, analyzed and compared with the previously collected data from the real system.

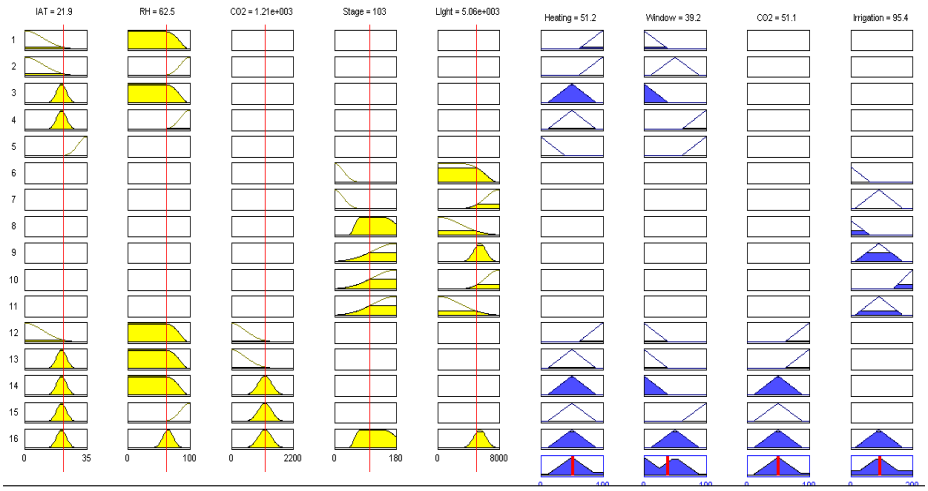


Fig.11 Defuzzified (crisp) actuators output for heating, ventilation, CO2 Dosing system and Irrigation system, depending of the measured values of Air Temperature, RH, CO2 and Growth stage of the Plant crisp inputs.

On Fig. 11 is presented Fuzzy Logic control strategy based on measured input variables. In this case for input temperature of 21.9 C, RH of 82.5 % level of CO2 of 1200 ppm, Plants old 103 days, and Light level of 5.000 Lux, values of actuators should be positioned at 51.2 % of Heating system, ventilation windows should be positioned on 39.2 % of maximal capacity, CO2 system should work on 51,5 % and

Irrigation system should work 95, second in every irrigation cycle. This is real assumption regarding that the temperature is near to the optimum, but high level of RH should be decreased by opening of ventilation windows. Also relatively old plants (113 days) will produce additional increase of RH, and light level of 5000 lux assume clear sun and in this case 91.6 sec x 1.1 ml x10 cycles = 1007 ml of water per plant per day is quite real assumption for this growing stage of plants.

5 CONCLUSION

Presented design of Fuzzy logic based controller for integrated control of Greenhouse generate control strategies based on linguistic variables. This approach allows for human expert knowledge to be incorporated into computer based control. Furthermore, number of different expert based strategies can be simulated and analyzed and compared. The further research will be in area of optimizing energy and water consumption, in order to obtain optimal control of greenhouse systems.

6 REFERENCES

1. Arvantis, K.G., P.N. Paraskevopoulos and A.A. Vernados,2000. Multirate adaptative tempe-rature control of greenhouses.Comp.Electronics Agric.,26:303-320.
2. Baturone, I., F.J.Moreno-Velo, S.Sanchez-Solano, V. Blanco and J.Ferruz,2005. Embedded fuzzy controllers on standard DSPs. Proceedings of the IEEE International Symposium on Industrial Electronics, June 20-23, Dubrovnik, Croatia, pp: 1-7.
3. Bennis, N., J. Duplaix, G. Enea, M. Haloua and H. Youlal, 2005. An advanced control of greenhouse climate. Proceedings of the 33rd International Symposium Actual Tasks on Agricultural Engineering, Feb. 21-25, Croatia, pp: 265-277.
4. Castaneda-Miranda,R., E.Ventura-Ramos, R.R. Peniche-Vera and G.Herrera-Ruiz,2006. Fuzzy greenhouse climate control system based on a field programmable gate array. Biosyst. Eng., 94: 165-177.
5. Dozier, G., A. Homaifar, E. Tunstel and D. Battle, 2001. An Introduction to Evolutionary Computation. In: Intelligent Control Systems using Soft Computing Methodologies, Zilouchian, A. and M. Jamshidi (Eds.). CRC Press, Boca Raton, FL.
6. Hanan J.J.(2001), Greenhouses Advanced Technology for protected Horticulture CRC Press
7. Horiuchi, J.I., 2002. Fuzzy modelling and control of biological processes. J. Biosci. Bioeng., 94: 574-578.
8. Iliev,O.L., N.E.Gough, G.K.Stojanov, G.M.Dimirovski and A.Zakeri (1996), “Obstacle avoidance for intelligent AGVs based on fuzzy control and expectation”. 13th IFAC World Congress (J.B.Cruz, Jr., IPC Chairman), San Francisco CA (USA), Volume Q, pp.441-447
9. Iliev O.L.,G.Dimirovski, Z.M.Gacovski, N.E.Gough and I.Griffith (1996),“Discrete event object oriented modelling of intelligent communication protocols”. World Automation Congress (WAC’96), Intelligent Automation and Control M.Jamshidi (Ed.) Vol.4 pp 325-330, TSI Press (*Selected Papers*), Albuquerque NM (USA)
10. Iliev O.L., B.Percinkova, N.E.Gough, and A. Zakeri (1999), “Two Level Control System for FMS – An Object Oriented Approach”, 14th World Congress of International Federation of Automatic Control, Beijing . P.R. China (5-9) July

11. Jamshidi,M.2003. Tools for intelligent control: Fuzzy controllers, neural networks and genetic algorithms. R. Soc. London Trans. Series A, 361: 1781-1808
12. Kovacic, Z. and S. Bogdan, 2006. Fuzzy Controllers Design Theory and Appliations. CRC Taylor and Francis Group, Boca Raton, FL., ISBN: 0-8493-3747-X, pp: 397
13. Lafont, F. and J.F. Balmat, 2002. Optimized fuzzy control of a greenhouse. Fuzzy Sets Syst., 128: 47-59.
14. Mitra, A.K., S. Nath and A.K. Sharma, 2008. Fog forecasting using rule-based fuzzy inference system. J. Indian Soc. Remote Sens., 36: 243-253
15. Monteith J.L. 1981 Evaporation and surface temperature. Quart. J.Royal Metor. Sci. 107:1-27
16. Pinon,S.M,E.F.Camacho and F.K.Uchen,2000.Constrained Predictive Control of Greenhouse. Elsevier Science Ltd., Spain, ISBN: 0-0-8044184-X.
17. Rodriguez, F., J.L. Guzman, M. Berenguel and M.R. Arahal, 2008. Adaptive hierarchical control of greenhouse crop production. Int. J. Adap. Cont. Signal Process, 22: 180-197.
18. Soto-Zarazua G.M,B.A.Romero-Archuleta, A.Merca-do-Luna, M. Toledano-Ayala, E. Rico-Garcia, R.R. Peniche-Vera and G. Herrera-Ruiz, 2011. Trends in Automated Systems Development for Greenhouse Horticulture. *Int. Journal of Agricultural Research, 6: 1-9*
19. Tanner. C.B. 1966. Comparison of energy balance and mass transport methods for measuring evaporation. Proc. Evarotranspiration and its role in water resources management ASAE, St. JOspeh MI.45-48.
20. Van Straten, G., E.J. van Henten, L.G. van Willigen-burg and R.C.van Ooteghem, 2010. Optimal Control of Greenhouse Cultivation. CRC Press, NewYork, Yager, R.R. and L.A. Zadeh, 1992. An introduction to fuzzy logic applications in intelligent systems. 1st Edn., Springer, New York
21. Zadeh, L.A., 1993. The role of fuzzy logic and soft computing in the conception and design of intelligent systems. Proceeding of the 8th Austrian Artificial Intelligence Conference on Fuzzy Logic in Artificial Intelligence, (FLAI'93), Springer-Verlag, London, UK, pp: 1-1.
22. Zimmermann, H.J., 1991. Fuzzy Set Theory and its Applications. 2nd Ed., Kluwer Academic Publishers, Boston, MA.

Analyzing E-commerce Multi-Agent systems using hierarchical Colored Petri nets

Meriem Taibi[1], Malika Ioualalen[1], and Nabila Salmi[1]

LSI - USTHB - BP 32, El-Alia, Bab-Ezzouar, 16111 - Alger, Algérie
taibi,ioualalen,salmi@lsi-usthb.dz

Abstract. E-commerce systems are important systems widely used by internauts. To automate most of commerce time-consuming stages of the buying process, software agent technologies proved to be efficient when employed in different e-commerce transaction stages. Furthermore, e-commerce systems should ensure correctness properties and a given level of quality of service that meets users expectancy. Therefore, we need to analyze such systems before implementation. This paper presents a high level Petri net modeling and analysis approach for e-commerce systems based on multi-agent technologies.
Keywords: E-commerce, Modeling, Multi-agent system, Hierarchical colored Petri net.

1 Introduction

Nowadays, traditional business trading has evolved to a more flexible technological mean, relying on electronic exchanges between customers and sellers. This is the essence of e-commerce systems, where a great number of participants communicate to achieve selling transactions. In such systems, main characteristics are exponential growing of information, information transparency, information overloading and different negotiation scenariis. To sustain these characteristics, multi-agent systems have been proposed to design e-commerce systems, for their ability to deal with complex properties.

A Multi-Agent System (MAS)[1]consists of a set of agents interacting with each other to achieve a common goal. Generally, MAS are known to work properly in a dynamic large-scale complex environment (open environment), thanks to autonomy, adaptability, robustness and flexibility.

In the context of e-commerce, several multi-agent systems were developed for the mediation of e-commerce, such as Kasbah[2] , Amazon [3] and eBay [4]. Several interaction patterns were defined, ranging from collaboration between agents, to competition for resources, requiring some negotiation level in e-commerce multi-agent system [5].

Although multi-agent technologies improve e-commerce services when used in design and implementation of e-commerce platforms, however, companies engaging in e-commerce are faced to a wide variety of challenges: how should the company be structured to profit in the marketplace, what are main required

S. Markovski, M. Gusev (Editors): ICT Innovations 2012, Web Proceedings, ISSN 1857-7288

correctness and safety properties, what is the quality of service which should be offered to satisfy users. As a result, a qualitative and quantitative analysis of such systems must be undertaken to build solutions adequate to targeted challenges, or to improve solutions that are already in place. In the field of MAS analysis, several studies have been proposed for modeling these systems by Petri nets. In [6], a model was proposed for a promotional game of viral marketing on the Internet. Specifically, authors used stochastic Petri nets for modeling a multi-agent wish list. As well, [7] used colored Petri nets (CPN) as a formal method to model a containerized transport system, then simulate and solve the storage problem. [8] applied a multiagent model formalization using CPN, to study a hunting management system. Elfallah-Segrouchni, Haddad and Mazouzi in [9] also proposed to use the CPN formalism to model interaction protocols. They described in [10] transcriptions of AUML diagrams into CPN models.

Besides, [11] analyzed e-commerce systems based on agents technology by deriving statistical results from interviews and questionnaires. In the context of pure e-commerce systems analysis, most proposed works [12][13][14], relied on simulations, testing and benchmarking to analyze e-commerce systems performances . However, these techniques usually require long time periods to get performance results. Moreover, the e-commerce system should be in operation in the two last methods to be able to test it, and important bugs cannot be easily corrected in those cases. It would be interesting to predict performances of e-commerce systems a priori during the design phase. Among proposals having based their work on formal methods, [15]and [16]having used timed automatas and PNs. The two works allow the modeling of the different interactions and documentation circulating but without focusing on the buyer and seller entities, which are an important part in market establishing.

Our primary objective is define a new e-commerce MAS based on three agents types, and to highlight the interest and principle of using PN in the context of e-commerce analysis. Therefore, we propose, in this paper, a formal methodology for modeling and analysis e-commerce MAS, using high level Petri nets. This methodology starts by distinguishing different agents interacting. Then, we model the different agent interactions using hierarchical CPN and finally, we check a number of qualitative properties.

This document is organized as follows. Section II gives a brief introduction to MAS as well as a description of interactions in e-commerce MAS. Section III recalls hierarchical CPN concepts and details our methodology for modeling of e-commerce multi-agent systems. In section IV we present our results and finally section V, discusses the obtained results and presents future work.

2 Multi-agent systems and E-commerce

2.1 Multi-agent systems (MAS)

There are two main aspects in MAS framework: agents architectures and agents interactions. Most implemented MASs will have a set of software components that provide:

• A communication language, like KQML (Knowledge Query Manipulation Language) and/or ACL (agent communication language);
• a syntax for the protocol (e.g. FIPA Contract-Net Interaction Protocol);

In order to support the communication language, a MAS have at least one communication channel (i.e. TCP/IP or CORBA, ...).

2.2 E-commerce system

According to Vladimir Zwass [17], electronic commerce is synonym of sharing business information, maintaining business relationships and conducting business transactions by means of telecommunications networks. With the dynamics of Internet, electronic commerce has been re-defined[17]. The wide range of business activities related to e-commerce brought a new terms to describe the Internet phenomenon in business sectors. Some of these focus on purchasing from on-line stores on the Internet. Since transactions go through the Internet and the Web, the terms I-commerce (Internet commerce) and even Web-commerce have been suggested but are now very rarely used.

These novel e-commerce characteristics changed the functioning of business enterprises majority, even purchasing and selling habits, during the last decades. To get much success in their business, enterprises are interested in ensuring correctness and offer users expected performances and quality of service.

In this objective, our interest is focused on multiagent e-commerce systems, and particularly, on modeling and analysis of such systems.

3 Using CPNs to Model E-commerce multi agent systems

Petri Nets (PN) and Colored Petri Nets (CPN) provide a framework for modeling and analyzing distributed and concurrent systems.

3.1 CPNs

Definition : A CPN is defined by a 9-tuple ($\sum$, P, T, A, N, C, G, E, I), where $\sum$ a set of non-empty types (called Colored sets); P a set of Places; T a set of Transitions; A a set of Arcs; N a node function; C a color function; G a guard function; E an arc expression function; and I an initialization function.

In contrast to ordinary PNs, CPNs have data associated with tokens. A token color is a schema or specification. Places in CPNs contain multi-sets of tokens. Arcs exiting and entering a place can have an associated constraint function to determine which multi-set elements are to be removed or held. Transitions of CPNs are associated with guard functions enforcing constraints on tokens.

For our modeling, we choose to use hierarchical CPNs[18], which are CPNs, where special transitions may represent sub-models given separately. Each sub-model models a sub-component of the studied system. The interest in such CPNs is to give a compact manageable not cumbersome global model. To go into details, the special transitions are replaced by their corresponding sub-models.

3.2 E-commerce MAS Modeling

E-commerce MAS description An e-commerce MAS is a MAS that connects multiple sellers and buyers agents on a single electronic marketplace called E-marketplace, where many interactions take place [19]. Agents involved are cognitive agents, able to communicate intentionally:
• **Buyer Agent (BA)**: The BA interfaces the customer with other agents in the system: It connects to the system, understands the customer behavior, then launches its requests. For that purpose, the agent allows the customer to send his requests and see results after processing and filtering them. This agent is seen as somehow the Front-office system.
• **Seller Agent (SA)**: It models a merchant who has information that should be offered or announced to buyers agents.
• **Mall Agent (MA)**: It is responsible for responding to BA requests, receive and record customer orders. The MA provides a list of BAs corresponding most to the customer's request, by comparing the customer's request with product features contained in the common database of the MA.

System environment: Our e-commerce MAS environment contains:
• The common compagny database (internal product database and catalog).
• The local databases of customers and sellers.

E-commerce MAS system organization When BA wants to find a product, it sends a request to the MA, which responds by sending the corresponding sellers agents list. The BA can then establish a process of direct negotiation with the selected seller agents.

To distinguish between the different agents and the different exchanged messages (which vary depending on the purpose of agents communication), we choose for *Colored Petri nets* to model the system described above.

CPN Modeling We model our e-commerce MAS with an hierarchically CPN model, to get a compact view of the global system and model subcomponents by separate sub-models for readability. So, we begin modeling by defining main parameters characterizing our CPN model: the structural representation, tokens coloration, the global model and its different sub-models.

Structural representation In our modeling, we consider:

- Places represent either agents states (before and after sending or receiving operations), or resources (ie the database, the buyer agent cart,etc).
- Transitions model sending and receiving actions or some processing actions.
- Tokens express the different agents and the various exchanged messages.
- Incoming arcs labels specify data required for firing the associated transition.
- Outgoing arcs labels specify data produced by a firing.
- Italic symbols above places (i.e.$AGENT$,INT, etc) indicate the color (or type) of tokens in these places.
- The symbols IN/OUT indicate that the place is as an input or output port.
- The symbol Fusion in places indicates that marks in these places are merged.

Token Coloration To differentiate tokens, we use color sets: we associate with each place the set of colors tokens that can mark it, and with each transition the color sets for which it is fired.

- The color agent, noted *AGENT* contains information record about an agent:
  1. The agent identifier denoted *ID*.
  2. The agent state denoted *STATE* (Waiting, Ready, blocked).
  3. The agent type denoted *TYPE* (Mall (M) Seller (S), Buyer (B)).

  So, an *AGENT* color is defined by the triplet (ID, STATE, TYPE). An example is (1, B, Ready).
- The color denoted by *MSG* contains information about exchanged messages:
  1. The message sender and receiver represented by their identifiers,
  2. The message type which may be a query response (Accept, Reject), an offer, a contract or a call for Proposals.

  Indeed, a *MSG* color is defined by the triplet (Rec, Sen, Type). As example is (id1, id2, CFP).

The global model of an e-commerce MAS The global e-commerce MAS model (top-level page of our model) is shown in figure 1. The overall CPN model consists of sub-models refereing to: agents creation, registration of sales agents, product research, negotiation, and purchasing. These sub-models are supervised by the main model.

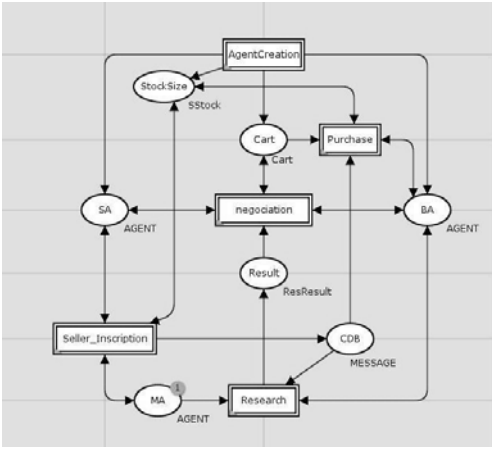


Fig. 1. The global model

Agent creation model When a client initiates the system to make a purchase, the transition *Connection*(figure 2, left.) is fired, the place *BA* receives a new color token, which involves the creation of BA. When SA offers a new products, the place *SA* receives a new color token (SA creation).

Products Registration model Figure 2, right, depicts the agent registration. The registration of products, offered by the seller agent in the common database (CDB), is represented by the transition *Seller_Inscription*, which requires the availability of seller and Mall agents. Products registration increases the SA storage size.

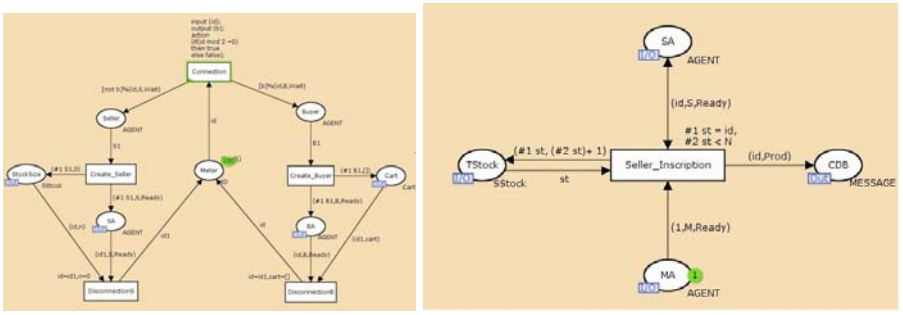


Fig. 2. Agents creation (left) and Registration (right)

Products Search model We assume that communication between agents during a product search conforms to the FIPA request protocol [20]. This protocol allows an agent to ask another agent to perform an action. In our context, BA asks the MA to do a search request. This query is modeled by the transition *Generate_Query*(figure 3). Then, the request will be sent to the Mall by the transition *Send_Req*. The availability of the common database is necessary to search product. If the products exist, the MA returns a list of seller agents *Send_Res* adapted to the products.

Negotiation model E-commerce MAS are becoming more and more autonomous to make decisions. They are based on the level cooperative protocol `FIPA Contract Net Interaction Protocol` [21], which allows negotiation and decision-making between two or more agents, each of which tries to achieve its objective. In our case, the requesting agent is the BA, which negotiates for products based on certain criteria, and the contractors are seller agents.

Consequently, we represent negotiation with the model of figure 4: So, the places *PP1*, *PP2*, *PP3*, *PP4*, *PP5*, *PP6*, *PP7* and *PP8* model shared spaces that describe the sending of messages. Whereas, all the places (*SA*, *SA_Wait*, *BA*, *BA_Wait*) represent the BA and seller agents states:

• BA announces a "call for proposal" on a seller agents network.
• Seller Agents who receive the announcement can answer by either an *offer*(via the transition *Send_Proposal.*), a *reject*(via the transition *Send_reject*), or an *not understood* response (via the transition *Send_notU*), indicating they did not understand the announcement.

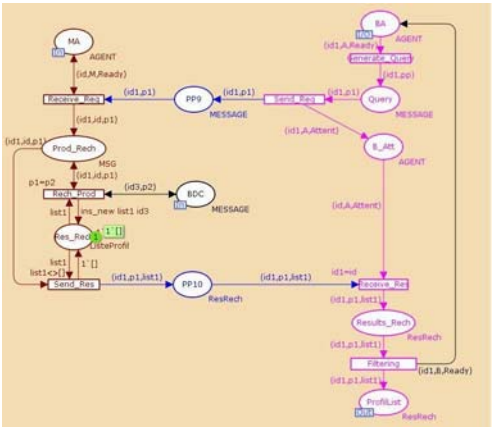


Fig. 3. Research Products

• BA receives and evaluates proposals, sends an *Accept* to seller agents whose proposals are accepted (hrough the transition *send_contract*)and a *Refuse* to other agents (via the transition *Send_Refuse*).
• At the end of interaction, the seller agent sends to the buyer agent, an *Inform message* to confirm the action achieving, or a *failure message* in a failure case.

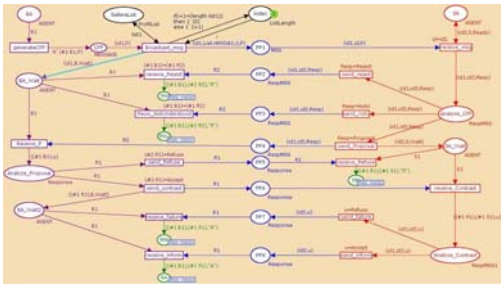


Fig. 4. Negotiation model

Model of adding products to the cart The cart is the set of products selected by the BA for possible order. Modeling the adding products to the cart gives the model of figure5, left. Adding products to the cart depends on the negotiation outcome saved in *(Res_nego)* places: if the negotiation result is positive, BA adds the product to the cart. Otherwise, there is no buying, leading to firing the

transition *NoPurchase*. At the end of negotiation, the BA is activated (its state becomes ready).

Purchase model The figure 5, right, shows the purchase model. After adding products in the cart, BA can then validate the cart through the transition *Validate_Cart* leading to the purchase.

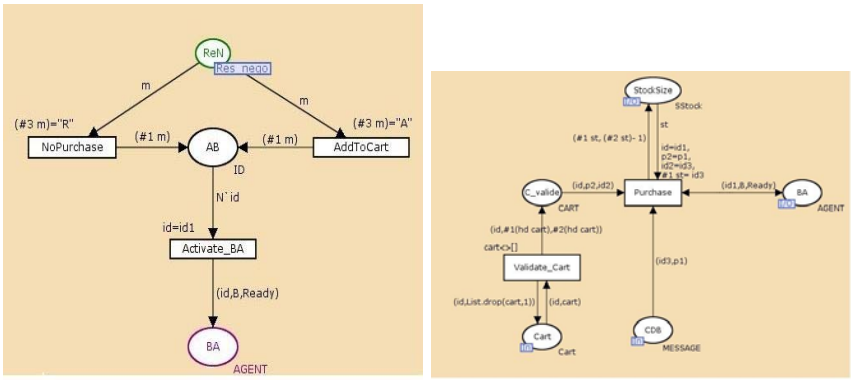


Fig. 5. Adding product (left) and Purchase model (right)

4 Experimental results

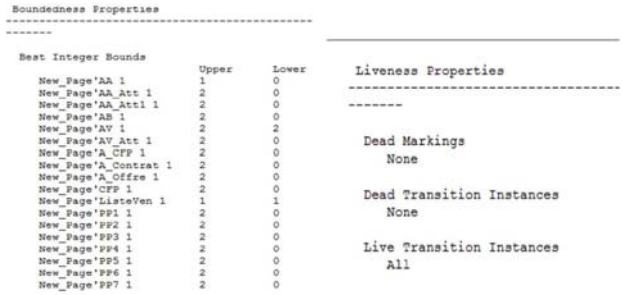
To experiment our models, we designed a website representing the main e-commerce functionalities.

4.1 Editing models with the CPN TOOl

We use CPN Tools software [22], to create and analyze our models. and we connect our MAS website with CPN tool using the library COMMS/CPN, developed to enable communication between CPN Tools and Java applications

4.2 Analyzing models

• Boundedness: The generated reports show that all places are bounded (see figure 6). This can be explained by the fact that the number of requests and messages generated by agents are processed consistently.
• Liveness and dead marking: The generated reports show a non-live model and existing dead markings. After analysis, we found that blocking appears in two cases: no proposal made by sellers or all proposals rejected by buyers. To address this, we added the assumption that: at least one proposal is always accepted by buyer agents. Therefore, we get the report of figure 6.



Boundedness Properties

Best Integer Bounds

	Upper	Lower
New_Page'AA 1	1	0
New_Page'AA_Att 1	2	0
New_Page'AA_Att1 1	2	0
New_Page'AB 1	2	0
New_Page'AV 1	2	2
New_Page'AV_Att 1	2	0
New_Page'A_CFP 1	2	0
New_Page'A_Contrat 1	2	0
New_Page'A_Offre 1	2	0
New_Page'CFP 1	2	0
New_Page'ListeVen 1	1	1
New_Page'PP1 1	2	0
New_Page'PP2 1	2	0
New_Page'PP3 1	2	0
New_Page'PP4 1	2	0
New_Page'PP5 1	2	0
New_Page'PP6 1	2	0
New_Page'PP7 1	2	0

Liveness Properties

Dead Markings
None

Dead Transition Instances
None

Live Transition Instances
All

Fig. 6. Boundeness(left) and Liveness (right) properties report

5 Conclusion and future works

In this paper, we have presented a first modeling proposal for e-commerce MAS, using colored PNs. The long-term goal is to allow analyze such systems, to ensure correctness and performances expected by users. The models were connected with CPN Tools to visualize agents behavior and analyze the system.

Modeling e-commerce functionalities within the MAS field highlights the fact that all underlying concepts make it a very rich behavioral approach. Moreover, it provides e-commerce enterprises a new tool to analyze their systems.

However, still more research work is required in several directions. We target mainly to extend our modeling and analysis by introducing the temporal dimension for being able to perform a quantitative analysis and compute e-commerce system performances(average waiting time, average connection number, etc). We plan also to analyze other e-commerce systems such as auction systems, which are more elaborated and complex than market place that we modeled.

References

[1] Ferber, J.: Les systèmes multi-agents vers une intelligence collective. Inter-Editions (1995)
[2] Chavez, A., Maes, P.: Kasbah : An agent marketplace for buying and selling goods. In: The First International Conference on The Practical Application of Intelligent Agents and Multi-Agent Technology. (1999)
[3] : Amazon. `http://amazon.co.uk`
[4] : Ebay. `http://www.ebay.com`
[5] Helmy., T.: Collaborative multiagentbased ecommerce framework. Int. J. Comput. Syst. Signal (12 2007)
[6] Balague, C.: Les Systèmes multi-agents en marketing : Modélisation par les réseaux de Petri. PhD thesis, École des Hautes études Commerciales (2005)
[7] Gazdare, M.K.: Optimisation Heuristique Distribuée du Problème de Stockage de Conteneurs dans un Port. PhD thesis, ECOLE CENTRALE DE LILLE (2008)

[8] Bakam, I.: Des systèmes multi-agents aux réseaux de pétri pour la gestion des ressources naturelles : Le cas de la faune dans l'est cameroun. PhD thesis, University of Yaoundé 1 (2003)
[9] Amal El Fallah-Seghrouchni, Serge Haddad, H.M.: Protocol engineering for multi-agent interaction. In: International Workshop on Modeling Autonomous Agents in a Multi-Agent World (MAAMAW), Valencia, Spain. (1999)
[10] Hamza Mazouzi, Amal El Fallah-Seghrouchni, S.H.: Open protocol design for complex interactions in multi-agent systems. In: International Conference on Autonomous Agents and Multi- Agents Systems (AAMAS), Bologna, Italy,. (2002)
[11] Devaux, L., Parashiv, C.: Le réseau des agents intelligents sur l'internet:révolution ou évolution commerciale? `http://www.cairn.info/load_pdf.php?ID_ARTICLE=RFG_152_0007` (2004)
[12] McLean, R., Blackie, N.M.: Customer and company voices in e-commerce:a qualitative analysis. Qualitative Market Research **7**(4) (2004) 243–249
[13] McLean, R., Blackie, N.M.: Customer and company voices in e-commerce:a qualitative analysis. Qualitative Market Research (August 2004)
[14] Stefani, A., Xenos, M.: A model for assessing the quality of e-commerce systems. In: Proceedings of the PC-HCI 2001 Conference on Human Computer Interaction, Patras. (2001) 105–109
[15] Jin, X., Ma, H.: An approach to formally modeling the component-based e-commerce system. In: Proceedings of the IEEE International Workshop. SOSE '05, USA, IEEE Computer Society (2005) 15–22
[16] Weitz, W.: Workflow modeling for internet-based commerce: An approach based on high-level petri nets. In: Proc. of the Int. IFIP/GI Working Conference on Trends in Distributed Systems for Electronic Commerce, UK (1998) 166–178
[17] Zwass, V.: Electronic commerces : Structures and issues. (1996)
[18] P Huber, K Jensen, R.M.S.: Hierarchies in coloured petri nets. In: The 10th Int. Conf. on Applications and Theory of PNs Advances in PNs 1990. (1991)
[19] Helmy, T.: Collaborative multi-agent-based e-commerce framework. Int. J. Comput. Syst. Signal **8**(1) (2007) 2–12
[20] : Fipa request interaction protocol. `http://www.fipa.org/specs/fipa00026`
[21] : Fipa contract net interaction protocol. `http://www.fipa.org/specs/fipa00029`
[22] : Cpn tools. `http://cpntools.org`

On the General Principles of Human-Computer Information Retrieval

Vesna Gega[1] and Pece Mitrevski[2]

[1]University for Information Science and Technology "St. Paul the Apostle" Ohrid
vesna.gega@uist.edu.mk
[2] St. Clement Ohridski University, Faculty of Technical Sciences, Ivo Lola Ribar bb
7000 Bitola, Republic of Macedonia
pece.mitrevski@uklo.edu.mk

Abstract. Nowadays, human-computer information retrieval (HCIR) is one of the most researched approaches, which applies human computer interaction (HCI) in information retrieval (IR), especially for text retrieval. The involvement of human activity (i.e. intelligence) in the process of retrieval is an interesting and challenging aspect of HCIR, thus it takes an interactive character with the purpose to improve user's needs. A number of researchers and research groups have conducted different experiments in this area. Innovative techniques for determining relevant documents and navigating in a large and complex set of full document results have been proposed. A novel approach based on focused retrieval for interactive identification of relevant document's parts has brought many improvements in this field. A special emphasis was placed on retrieving passages and XML elements from structured documents. Numerous techniques for focused retrieval based on snippets, facets and relevance feedback have been presented recently, bringing advantages and a more realistic scenario for text retrieval, but also a lot of open issues that need to be explored. This paper surveys the fundamental characteristics of the HCIR, as well as the focused HCIR, such as using passages and XML elements as retrievable units from structured or pure textual documents, thus providing a basis for development of novel HCIR approaches for interactive text retrieval, as our main goal in future research.

Keywords: HCIR, information retrieval, passage, XML element, evaluation

1 Introduction

Information retrieval is researched paradigm more than a half century and it is related to searching information and calculating how those are likely to be relevant for a user's asked query [4], [21]. It is an automated process that includes: indexing, searching, ranking, retrieval and presentation of the relevant information, as shown in Fig. 1. Usually, the searching is realized from very long and complex electronically stored sets of information (digital knowledge), such as World Wide Web, document collections, libraries, and relational databases. The results in different format, ranked by

their degree of relevance are presented to the user. The term results refer to documents containing text, images, videos, audios and their parts. The degree of relevance is obtained as a result of a process implementing variety of probabilistic and statistical functions [19], [21], [33-35], [39].

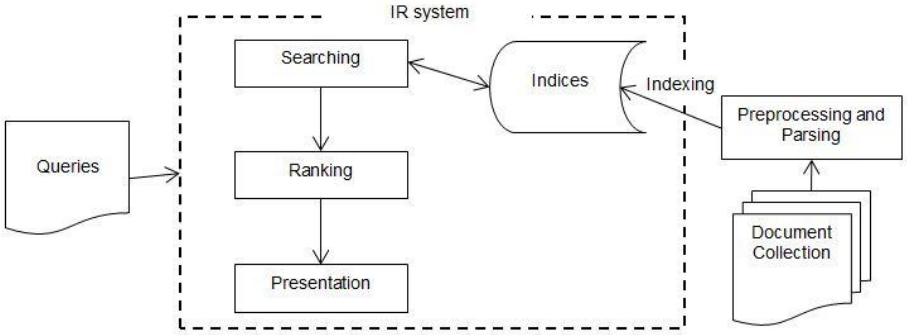


Fig. 1. IR process (inspired by: G. Marchionini, Toward Human Computer Interaction 2005 Lazerow Lecture, University of Washington)

Innovative techniques for information retrieval were developed through the years, tested on different large document collections, as part of English and non-English evaluation forums. One of the most challenging branches in the modern information retrieval is human-computer information retrieval (HCIR). In this paper we describe the progress of HCIR techniques, as well as focused HCIR techniques, for text retrieval in detail. We present the state of the art researches and explain future directions. This paper is organized as follows.

In Section 2, we survey the fundamental characteristics of HCIR, as well as various proposed techniques. The evaluation metrics used to measure effectiveness of HCIR systems are described in Section 3. Section 4 concludes the paper and shows future research directions.

2 Human-Computer Information Retrieval

An interesting and challenging aspect of HCIR is the involvement of human activity (HCI) in the process of retrieval (IR), thus it takes an interactive character and brings real scenario of searching, as shown in Fig. 2. With purpose to improve user's needs, the interaction continues until the user finds the right information of his/her interest by refining the originally proposed query [15]. HCIR concept has been introduced by *Marchionini G.* [24] for the first time between 2004 and 2006. Nowadays, it has evolved to stage where the user with action, perception and reflection is more involved in the relevant information seeking process [23], going beyond static queries and standard representation of results.

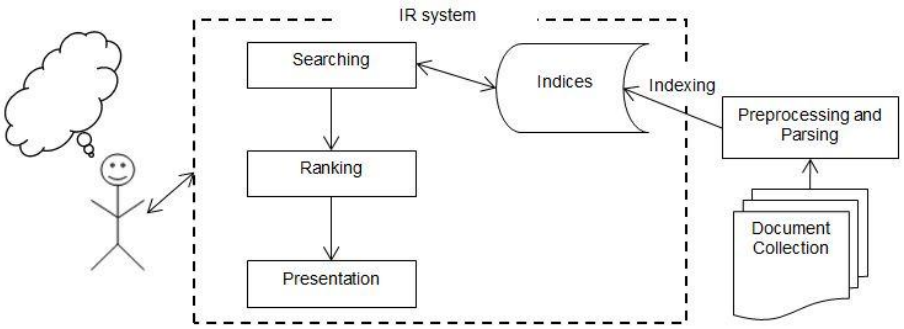


Fig. 2. HCIR process (inspired by: G. Marchionini, Toward Human Computer Interaction 2005 Lazerow Lecture, University of Washington)

2.1 HCIR Techniques

Intersection between HCI and IR is subject of many researchers and research groups, such as Workshop on Human-Computer Interaction and Information Retrieval [1], INEX[2], CLEF[3] and TREC[4]. New and powerful techniques for determining relevant documents and navigating in large and complex set of full document results have been proposed [22], [40]. A novelty approach based on interactive identification of relevant document's parts has brought many improvements in this field. A special emphasis was placed on retrieving passages and XML elements from structured documents, based on snippets, facets and relevance feedback/query expansion [5], [12]. All these techniques aim to effective and efficient interaction process without unnecessary costs of time, minimal mouse clicks or scrolling and context changing.

Faceted retrieval.

Given a query containing one or more terms, the information retrieval system returns an ordered list of ranked relevant results. Usually, that list is very long and user finds difficulties to navigate through it. Faceted search provides interactive hierarchical search and enables users to quickly determine the desired results. It means that user can browse in the categorized list of relevant results, refine the query and narrow down new different relevant results proposed by the system. The categories are also known as facets and the values they contain are facet-values [21]. The advantage of faceted search is the opportunity users to choose relevant results from those categories that are closest to their needs and way of thinking [3-5], [12]. The research challenge is to investigate, introduce and evaluate new effective and efficient approaches for facets and facet-values formulation.

[1] http://www.hcir.info/
[2] https://inex.mmci.uni-saarland.de/data/links.jsp
[3] www.clef-campaign.org
[4] trec.nist.gov

The Relational Browser called RB was for the first time introduced *Marchionini G. et al.* [22]. It is an approach that provides a list of relationships among attributes in an information space, which helps users in full text information seeking. For better and effective browsing, simple interactive information manipulation is implemented. It underwent several changes [40]. The authors have developed a novel interface, called RB++, where the faceted search plays the main role, allowing users explore relationships among facets and refine the original query by simple mouse actions on the facets and facet-values.

Ramirez G. [5] has experimented with large and highly structured document collection of IMDB[5] documents on Indri[6] search engine. Fixed set of facets were used, according to the collection DTD[7]. The facet-values were presented on hierarchical and non-hierarchical way. A non-hierarchical presentation means that all the facet-values are at the same level. As opposite, a hierarchical presentation is related to granularity, thus there are nested facet-values in three hierarchical levels.

Schuth A. et al. [5] also used the large and complex IMDB document collection. Limited numbers of facets covering not unique facet-values from a single document were defined. Each facet has name and an XPath[8] expression by which the facet-values can be retrieved. Two strategies for facets formulation were developed. The ranking of the facet-values in the both cases is based on hits that would be the result if the facet-values were selected.

There have been a number of studies that have examined faceted search on mobile web devices, such as that described by *Capra, R. et al.* [7]. Also different design patterns for mobile faceted search have been proposed[9]. Still there is an open question if faceted navigation is well-suited for the small screen.

Snippet retrieval.

Among the list of relevant returned results, there are also some irrelevant results that don't satisfy user's needs. Showing short summary of the result is another way for helping users to easily decide whether or not the result is relevant, without needing to view the result itself, because the summary is in more human-consumable form. The summary is called snippet. Usually, snippets are automatically generated from the result. They can be extracted as a block of text containing one or more sentences [6], [8] and [13], and a paragraph or several successive paragraphs from the result. In terms of size, the snippet can have fixed or variable length where the length is expressed as number of characters, words, lines or bytes. If the results have logical structure, the snippets can be formed with different kind of grouping elements. There are two types of snippet generation, static and dynamic [21]. The first one is not related to the query. The second one is query dependent and follows the users' needs [36], [38]. Also, metadata associated with the document as result can be used for snippet

[5] http://imdb.com
[6] http://www.lemurproject.org/indri/
[7] http://www.w3schools.com/dtd/default.asp
[8] http://www.w3schools.com/xpath/
[9] http://www.uxmatters.com/mt/archives/2010/04/

formulation. The research question is how to generate snippets to provide sufficient information for maximizing effectiveness and efficiency of the search process.

The preliminary results in [5] showed that poor snippets cause the users to miss over than 50% of relevant results. Also, the percentage of non-relevant documents that are correctly assessed was very high, indicating that snippets are useful in non-relevant documents determination.

The evaluation in [36] and [38] have shown that use of query dependent snippets brought many improvements in terms of speed, precision and recall than the use of query independent snippets.

Efficiency was main goal in [37], where also query dependent snippets were generated from results pages presented by search engines. The authors have proposed document compression method that significantly speeds up the snippets generation. Also, it was shown that snippets are generated faster if documents were previously cashed in the RAM.

Huang Y. et al. [14] presented eXtract, a system that efficiently generates snippets from XML structured documents. They identified that a good snippet should be a self-contained meaningful information unit of a small size that effectively summarizes the query result and differentiates it from others, according to which users can quickly assess the relevance of the query result. For that purpose, they have developed a new effective and efficient algorithm.

Leal L. et al. [5] experimented on Wikipedia[10] document collection and snippet generation was based on selecting highly ranked sentences, ranked according to the frequency of query terms. Also the title of the document was concatenated to the sentences in the snippets. Snippets had limited length of maximum 300 characters, depending to the document length. They expanded the original queries with query expansion technique, and main conclusion is that a large number of extra terms negatively influences on sentences selection.

Rongmei Li et al. [5] ranked the Wikipedia documents with the Language Model [39] and snippets are generated from the ranked list of relevant results. The snippets were term-extracted. Each term was weighed according its relative occurrence in the document and in the entire document collection. The top K scoring terms in two different clusters were chosen to form the snippet. One is a cluster containing only terms (words) that don't have any relationship between each other. Another is a cluster that contains semi-sentences related to the query.

Wang S. et al. [5] used Vector Space Model [19] for searching and ranking Wikipedia documents, based on structure and content. The authors also expand the query with query expansion techniques. For snippet generation they used query relevance, significant words, title/section-title relevance and tag weight to evaluate the relevance between sentences and query. The sentences with higher score formed the snippets.

Relevance feedback.

Query refinement is core process in relevance feedback task, realized in order to improve retrieval performances. Information retrieval system fully automatically or

[10] http://www.mpi-inf.mpg.de/departments/d5/software/inex/

with user's help can refine originally proposed query, depending of ranked results proposed by the system [1], [30]. There are three types of feedback: explicit, implicit and blind or "pseudo" feedback [21]. The explicit feedback is characterized by user's involvement in the retrieval process. It is an interactive process where user marks relevant and non-relevant returned results for his/her asked query and the system re-ranks unseen results and returns new list of results, closer to his/her interests. The implicit feedback is related to user's behavior, thus results that are more often viewed, scrolled, clicked, saved and printed are considered to be more relevant [18], [25]. The main difference between implicit and explicit feedback is that the user doesn't asses the results and give any addition effort, but he/she on indirect passive way influences on the feedback. The blind or "pseudo" feedback is a fully automated approach where information retrieval system simulates user's activity and does standard information retrieval. It retrieves a number of relevant results, and supposes that top K results of them are most relevant to the user's asked query, influencing on the feedback. Mechanisms for improving relevance feedback retrieval are interesting to explore.

Rocchio algorithm is method that implements relevance feedback using Vector Space Model [19], proposed by *Rocchio, J. J.* [29], around 1970. By query refinement an arbitrary percentage of relevant and non-relevant results tends to increase the information retrieval system's recall, and possibly precision. This scenario is optimal when the angle between the refined query vector and result vector is smaller, thus the similarity is maximized with relevant documents and minimized with non-relevant documents [21].

Kelly D. et al. [18] have continued the work of *Morita M. et al.* [25], experimenting with implicit feedback, in context of these hypotheses: Users will spend more time reading those documents that they find relevant, Users will scroll more often within those documents that they find relevant, and Users will interact more with those documents that they find relevant.

Lv Y., et al. [20] have proposed new effective pseudo relevance feedback approach, named positional relevance feedback, based on positions of query terms in feedback documents. Their experiments were conducted using full document feedback and passage feedback.

Allan J. [2] has developed a hybrid approach for relevance feedback retrieval from large documents. The idea was to use passages in relevance feedback task only where necessary. It means that full short documents in combination with passages from long documents in relevance feedback, produce for more effective retrieval.

Query reformulation may be a difficult and time consuming activity [21], thus fully automatic relevance feedback information systems were proposed and evaluated on [5], [21]. They have developed simulated exhaustive incremental user feedback [1]. Many proposed algorithms for focused retrieval based on passages and XML elements feedbacks went far beyond Rocchio feedback algorithm [5], [12].

3 On the Methodologies for HCIR System Evaluation: Issues and Prospects

Usually, the metrics for evaluating the effectiveness of a standard information retrieval system [27] use the widely accepted Cranfield methodology [10], including document collection, queries and relevance judgments in the evaluation process [21]. They are based on a set of measures, such as precision, precision @ K, recall, F-Measure, Mean Average Precision (MAP) and Discounted Cumulated Gain (DCG) [4], [9], [27].

In faceted search, the relevance of the data covered by facet-value can be measured with Normalized Discounted Cumulated Gain (NDCG) as a DCG variant for calculating relevance for all queries [5]. For the first time, NDCG was introduced by *Järvelin K. et al.* [16], and modified in two variants: normalized NDCG and recursive NDCG by *Schuth A. et al.* [32]. Also, Interaction cost is used to measure effectiveness of a faceted retrieval system, based on user-computer interaction evaluation in information seeking process [5]. It is measured as the number of results and facets that the user examined before he/she encounter the first relevant results. Usually, the interaction process is simulated in order to avoid expensive user study and make user study repeatable [17, 32].

There is a pallet of measures for snippet retrieval evaluation. *Overwijk A. et al.* [26] have employed measures based on the standard precision and recall, modified in terms on spans and nuggets. Comparison of relevance assessments based on whole documents vs. short snippets was introduced by *Bellot P. et al.* [5]. They have proposed a measure for calculating how effective snippets were in the retrieval process satisfying user's needs, called Mean Prediction Accuracy (MPA). Its averaged variant over all topics is called Mean normalized prediction accuracy (MNPA). Negative recall (NR) is the percentage of irrelevant documents correctly assessed, averaged over all topics. Also, *Bellot P. et al.* [5] have used Geometric Mean of recall and negative recall (GM) for snippet retrieval evaluation. Different approaches for evaluation of web search snippets at Yandex, a powerful Russian web search engine were surveyed and experimented by *Savenkov D. et al.* [31].

Ruthven I. et al. [30] gave a review of evaluation methods for relevance feedback, intended to measure the effect of feedback on the unseen relevant documents. The first one is Residual ranking, where the documents used in relevance feedback system, are removed from the document collection before evaluation. That includes the relevant and some non-relevant documents. Precision and recall are calculated on the remaining (residual) document collection. The second one is Freezing, where the initially top ranked documents are "frozen", the ones used to modify the query. The relevance feedback system performs ranking again the remaining documents, and the precision/recall evaluation is conducted on the entire document collection. The third one is Test and control groups, where the document collection is randomly partitioned into two equal groups, the first used to train the relevance feedback system and the second used to evaluate the system. With simulated exhaustive incremental user feedback proposed by *Bellot P. et al.* [5] and *Geva S. et al.* [12], the approach for evaluation was extended in several ways, relative to traditional ad-hoc evaluation with

standard precision/recall measures based on Cranfield methodology. It gives exhaustiveness on the evaluation process, thus it becomes reliable and less dependent on the specific search engine in use. Here, there is an evaluation platform where evaluation is performed over executable implementations of relevance feedback algorithms rather than being performed over result submissions.

With HCIR popularization the need of new and different measures has occurred. Despite the information retrieval evaluation, user's effort in the interaction process should be considered and evaluated. Our point of view is that, despite described automated measures, the complex process of HCIR evaluation could be solved with a range of different methods based on user behavior, such as eye-tracking [11, 31], manual evaluation [31], and click-through mining [28]. The aim of all the developed and effectively evaluated approaches in the HCIR field is the improvement of searching quality. Nevertheless, we note that they can be used in the process of discovering how people search, thus different information retrieval systems should be built.

4 Conclusion

The major work realized so far in the field of HCIR is investigated in this paper. We have surveyed and explained several techniques that involve human effort in the searching process, such as facets, snippets and relevance feedback. Existing approaches for effective and efficient retrieval have been summarized and various research challenges examined in detail, such as retrieval of document's parts rather than whole documents. Thus, our contribution is to present altogether the problems and solutions in the domain of the HCIR paradigm as a base for further research steps, such as development of novel improved HCIR approaches for interactive text retrieval and experimenting with them on both English and Macedonian document collection.

References

1. Aalbersberg, I.J.: Incremental Relevance Feedback. In: Proceedings of SIGIR, pp.11-22 (1992)
2. Allan, J.: Relevance Feedback With Too Much Data. In: University of Massachusetts Amherst, MA, USA (1995)
3. Athenikos, S.J., Lin, X.: Search As You Think AND Think As You Search: Semantic Search Interface for Entity/Fact Retrieval". Presented at the Fifth Workshop on Human-Computer Interaction and Information Retrieval (HCIR) (2011)
4. Baeza-Yates R., Ribeiro-Neto B.: Modern Information Retrieval New York. In: ACM Press (1999)
5. Bellot, P., Chappell, T., Doucet, A., Geva, S., Kamps, J., Kazai, G., Koolen, M., Landoni, M., Marx, M., Moriceau, V., Mothe, J., Ramírez, G., Sanderson, M., SanJuan, E., Scholer, F., Tannier, X., Theobald, M., Trappett, M., Trotman, A., Wang, Q.: Report on INEX 2011. In: SIGIR Forum 46(1), pp. 33-42 (2012)
6. Brandow, R., Mitze, K., Rau, L.F.: Automatic condensation of electronic publications by sentence selection. In: Information Processing & Management 31(5), pp.675–685 (1995)

7. Capra, R., Raitz, J.: Diamond Browser: Faceted Search on Mobile Devices. Presented at the Fifth Workshop on Human-Computer Interaction and Information Retrieval (HCIR) (2011)
8. Chuang, W.T., Yang, J.: Extracting Sentence Segments for Text Summarization: A Machine Learning Approach. In: SIGIR '00 Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval (2000)
9. Clark, C.L.A, Kolla, M., Cormack, G.V., Vechtomova. O., Ashkan, A., Büttcher, S., MacKinnon, I.: Novelty and diversity in information retrieval evaluation. In: SIGIR 2008, pp.659-666 (2008)
10. Cleverdon, C.W.: The Cranfield tests on index language devices. In: Aslib Proceedings, 19, pp.173–192 (1967)
11. Cutrell, E., Guan, Zh.: What Are You Looking For? An Eye-tracking Study of Information Usage in Web Search. In: CHI'07 (2007)
12. Geva, S., Kamps, J., Schenkel, R., Trotman, A.: Comparative Evaluation of Focused Retrieval. In: 9th International Workshop of the Inititative for the Evaluation of XML Retrieval, INEX 2010, Vugh, The Netherlands. December 13-15, (2010). Revised Selected Papers Springer (2011)
13. Goldsteiny, J., Kantrowitz, M., Mittal, V., Carbonelly, J.: Summarizing Text Documents: Sentence Selection and Evaluation Metrics. In: Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval. (1999)
14. Huang, Y., Liu, Z., Chen,Y.: eXtract: a snippet generation system for XML search. In: PVLDB 1(2), pp.1392-1395 (2008)
15. Ingwersen, P.: Information Retrieval Interaction. Taylor Graham (1992)
16. Järvelin. K., Kekäläinen, J.: Cumulated gain-based evaluation of IR techniques. In: ACM Trans. Inf. Syst., Vol. 20, No. 4., pp. 422-446 (2002)
17. Kashyap, A., Hristidis, V., Petropoulos, M., FACeTOR: Cost-Driven Exploration of Faceted Query Results. In: CIKM (2010)
18. Kelly, D., Belkin, N.J.: Reading time, scrolling and interaction: exploring implicit sources of user preferences for relevance feedback. In: Proceeding SIGIR '01 Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval, pp.408 - 409, NY, USA (2001)
19. Lee, D., Chuang, H. & Seamons, K.: Document ranking and Vector space model. In: IEEE Software, 14(2), pp.67–75 (1997)
20. Lv, Y., Zhai, C.: Positional relevance model for pseudo-relevance feedback. In: Proceeding SIGIR '10 Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval, pp.579-586, ACM New York, NY, USA (2010)
21. Manning, C.D., Raghavan, P., and Schütze, H.: Introduction to Information Retrieval. Cambridge, UK: Cambridge University Press (2008)
22. Marchionini, G., Brunk, B.: Towards a General Relation Browser: A GUI for Information Architects. In: J. Digit. Inf. 4(1) (2003)
23. Marchionini, G.: From Information Retrieval to Information Interaction. In: ECIR 2004: 1-11 (2004)
24. Marchionini, G.: Toward Human-Computer Information Retrieval Bulletin. In: June/July 2006 Bulletin of the American Society for Information Science (2006)
25. Morita, M., Shinoda, Y.: Information Filtering Based on User Behaviour Analysis and Best Match Text Retrieval. In: SIGIR 1994, pp.272-281 (1994)
26. Overwijk, A., Nguyen, D., Hauff, C., Trieschnigg, D., Hiemstra, D., Jong, F.D.: On the evaluation of snippet selection for WebCLEF. In: Proceeding CLEF'08 Proceedings of the

9th Cross-language evaluation forum conference on Evaluating systems for multilingual and multimodal information access, pp.794-797 Springer-Verlag Berlin, Heidelberg. (2009)
27. Pehcevski, J., Piwowarski, B.: Evaluation Metrics for Semi-Structured Text Retrieval. In: Encyclopedia of Database Systems, Editors-in-chief: Liu, Ling; Özsu, M. Tamer, Springer (print and online) (2009)
28. Radlinski, F., Kurup, M., Joachims, T.: How Does Clickthrough Data Reflect Retrieval Quality? In: CIKM'08 (2008)
29. Rocchio, J. J.: Relevance feedback in information retrieval. In: G. Salton, editor, The SMART Retrieval System: Experiments in Automatic Document Process-ing, Prentice-Hall Series in Automatic Computation, chapter 14, pp.313–323. Prentice-Hall, Englewood Cliffs NJ (1971)
30. Ruthven, I., Lalmas, M.,: A survey on the use of relevance feedback for information access systems. In: Knowl. Eng. Rev., 18(2), pp.95-145 (2003)
31. Savenkov, D., Braslavski, P., Lebedev, M.: Search Snippet Evaluation at Yandex: Lessons Learned and Future Directions. In: CLEF, Vol. 6941 Springer, pp. 14-25 (2011)
32. Schuth, A., Marx. M.J.: Evaluation Methods for Rankings of Facetvalues for Faceted Search. In:Proceedings of the Conference on Multilingual and Multimodal Information Access Evaluation (2011)
33. Singhal, A., Buckley, C. & Mitra, M.: Pivoted Document Length Normalization. In: ACM SIGIR'96 (1996)
34. Sparck Jones, K., Walker, S. & Robertson, S.E., 2000a.: A Probabilistic model of information retrieval: development and comparative experiments. In: Part 1. In Information Processing and Management 36, pp.779-808 (2000)
35. Sparck Jones, K., Walker, S. & Robertson, S.E., 2000b.: A Probabilistic model of information retrieval: development and comparative experiments. In: Part 2. In Information Processing and Management 36, pp.809-40 (2000)
36. Tombros, A., Sanderson, M.: Advantages of query biased summaries in information retrieval. In Proceeding SIGIR '98 Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval Pages 2 – 10 ACM New York, NY, USA (1998)
37. Turpin, A., Tsegay, Y., Hawking, D., Williams, H.E.: Fast generation of result snippets in web search. In: SIGIR 2007, pp.127-134 (2007)
38. White, R.W., Jose, J.M., Ruthven, I.: A task-oriented study on the influencing effects of query-biased summarisation in web searching. In: Information Processing & Management 39(5), pp.707–733 (2003)
39. Zhai, C., Lafferty, J.: A Study of Smoothing Methods for Language Models Applied to Information Retrieval. In: ACM Transactions on Information Systems (TOIS), Volume 22, Issue 2 (2004)
40. Zhang, J., Marchionini, G.: Coupling browse and search in highly interactive user interfaces: a study of the relation browser++. In: JCDL, pp.384 (2004)

Architecture of an electronic student services system and its implementation

Ivan Chorbev[1], Marjan Gusev[1], Dejan Gjorgjevikj[1], Ana Madevska-Bogdanova[1]

[1] Faculty of computer science and engineering, University of Ss Cyril and Methodius, "Rugjer Boshkovikj" 16, P.O. Box 393, 1000 Skopje,R. of Macedonia
{ivan.chorbev, marjan.gushev, dejan.gjorgjevikj, ana.madevska.bogdanova}@finki.ukim.mk

Abstract. A new university information system that provides electronic services for both university management and students was developed for iKnow project and implemented at Universities in Macedonia. It is an eStudent system enabling complete electronic functioning of a University, rendering paper documents obsolete. The system is web based and implements state of the art modular service oriented technologies. This paper presents the modules and functionalities, software and database architecture.

Keywords: iKnow, electronic student services, university information system, student information systems, LMS, eStudent, MVP, MVC.

1 Introduction

The eStudent Information System iKnow is designed to store and administer student's records and personal files, as well as related university data. It is developed using innovative approach and knowledge management techniques. The system provides exchange of electronic information among all stakeholders: students, professors, administration, university management and the Ministry of Education.

The system consists of two main components, each consisting of modules: the new students Enrolment Student Services (ESS) and the Core Student Services (CSS). Special software modules have been developed for data migration from legacy software applications into iKnow using Excel template files.

The Enrolment component consists of the candidate's enrolment wizard, the enrolment forms for manual entry of candidate's data, candidate's data processing, ranking module and the enrolment results publishing.

The Core student services component includes:

- Module for administration (university and faculty), which encompasses users management, exams sessions, semesters and the integration with the learning management system (LMS) Moodle;
- Module for study programs and schedules, study programs management, courses management, equivalency of courses, connections between courses and programs

S. Markovski, M. Gusev (Editors): ICT Innovations 2012, Web Proceedings, ISSN 1857-7288

- Student activities module, the students online services, with its components: semester enrolment, exam application, courses selection, documents and certificates request, grades overview, diploma thesis management, the student services department for overall student data processing;
- Module for personal identification and access control
- Module for electronic payment and use of resources
- Migration of old data, interfaces to other systems

This paper is organized as follows: Chapter 2 gives an overview of other systems similar to the one presented in this paper. Details of the implementation of the system are presented in chapter 3, including the software architecture, database specifics and performance tests. Technical data about the hardware server infrastructure and security issues can also be found here. The customization and fine tuning of the user interface along with evaluation of the functionalities implemented in the system are presented in the chapter 4. Evaluation analyses of the software implementation are presented in chapter 5, followed by the concluding remarks in the chapter 6.

2 Related work

Information systems used for storing and organizing data about students and related university concepts and activities have been around ever since computers became available in the university environment. With the advancement and increased availability of computers as well as software platforms, such systems grew in complexity and features offered. Lately, with the rise of the Internet and its ubiquitous presence these systems became ever more student oriented transferring increasing number of tasks as well as opportunities to the students themselves. Operating such systems becomes increasingly the student's job offering them control over their educational choices as well as timely and transparent information they ought to have.

Today's members of e-Society expect each service they need to be available online. Young, IT era educated students are the potential primary users of such online services, rather than an exception to the rule. The availability of omnipresent Internet service through mobile devices anytime and everywhere increases expectations for higher education institutions to transform into 24 / 7 service providers. These expectations can only be met by deploying state of the art online student services that provide plethora of correct and timely information and enable instant feedback from students into the systems. Students expect to be allowed to apply for exams, read their exam results, choose elective courses or enroll in a semester by only scrolling their fingers on the touch screen of their mobile device anytime, everywhere. In the same time, the stored information must be safe, guarded from privacy breaches, consistent, readily available on demand and with permission only. On the other hand, educational institutions are expected to derive evermore complex statistical reports and archive data, provide more complex combinations of elective study programs, making their information systems increasingly strained and in need of features like extensibility, flexibility, modularity.

Currently there are multiple student information systems in use in universities across the world [15], [16], ranging from in-house developed solutions, up to from-

market of-the-shelf products sometimes adapted to the specific needs of the university in question. Also, they provide a variable range of online services for students, but very few offer complete electronic functioning and eliminated paperwork. In some instances CRM systems are adapted to serve as academic student info systems [17],[18],[19]. Also ERP solutions are sometime adapted for roles of student data storage [20], [21]. Systems use both open source and licensed databases with corresponding interfaces.

3 Implementation

3.1 Server infrastructure and security

In order to increase the performance and secure the stability of the system in face of the increasing number of concurrent users, a reorganization of the server architecture was performed engaging virtualization. Two servers and an independent storage system were employed at University Sts Cyril and Methodius, Macedonia. The first server is used as SQL server with Microsoft SQL Server installed and the second is used as an application server. The system is upgraded by installing 4 virtual Windows servers using VMware. Both servers have an intel xeon e5630 2.56 GHz cpu and 16GB of RAM, 1 TB storage.

Significant improvement of the responsiveness is achieved since virtualization and the multiple Internet Information Servers (IIS) that resulted avoid deadlocks and blockings often experienced in single, overused IISs.

Additional backup Internet link was installed in order to improve system availability in case of failures of the primary link. A third server on a remote location is installed that ought to be fully synchronized and take over in case of failure of the primary server farm or its Internet links.

The SSL protocol is used when accessing all system functionalities; both web forms and web services. A digital certificate was installed to encrypt all the communication.

3.2 Database architecture

The system is based on an SQL Server 2010 R2 database. Other database platforms are also integrated because of usage of external Learning Management Systems (LMS) like Moodle. The SQL database contains 145 tables so far, including several ASP.NET membership tables. The tables are connected with relations following the rules of a normal database form. However, certain redundancies were allowed in order to optimize query speeds and reporting efficiencies. Although the student's grade average can be calculated in real time, to avoid reporting delays and increased the efficiencies, the pre-calculated averages are stored for each student. There are also

other pre-calculated redundant data if they are often retrieved. Special attention is set to keep the consistency of these pre-calculated values. Recalculations are performed whenever the possibility arises for change of a pre-calculated value, and therefore the data consistency is guaranteed in the same time severely increasing efficiency. The database also contains 203 stored procedures used to increase efficiency and code transparency. **Fig. 1** gives an illustrative part of the database diagram.

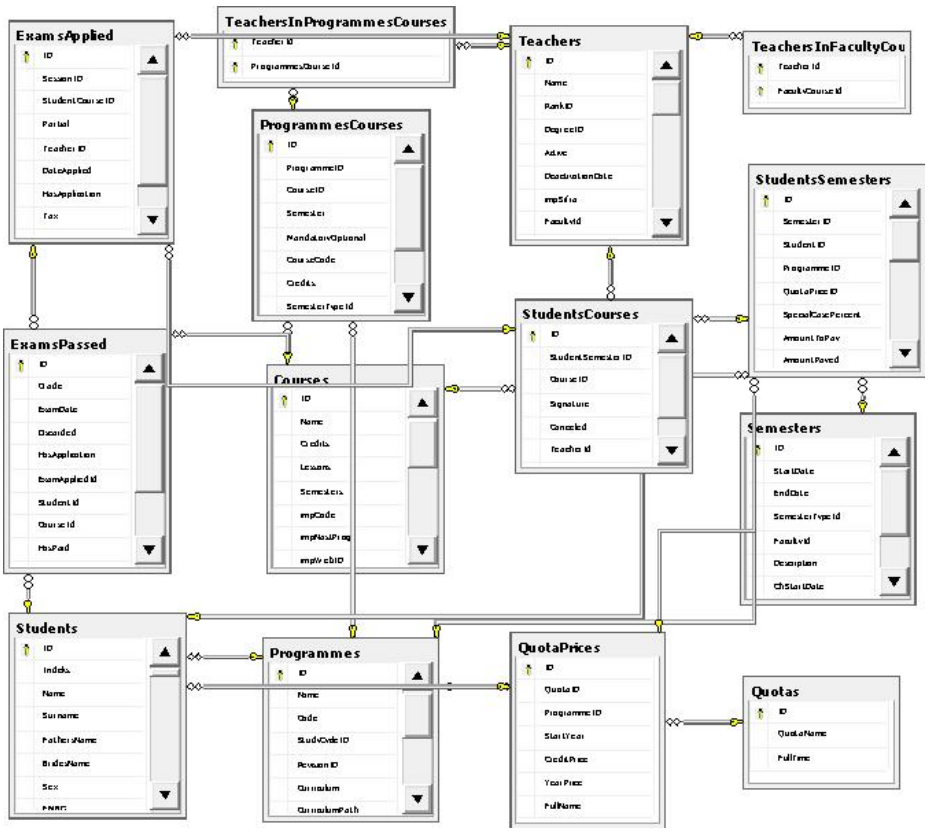


Fig. 1 An illustrative part of the database diagram

3.3. The MVP pattern used

The software architecture of the system strictly follows the principles of the MVP software design pattern.

Both the Model-View-Controller (MVC) and Model-View-Presenter (MVP) patterns have been use for several years [13], [14]. They both address the key principal of separation of concerns between the User Interface and the business layers.

There are several frameworks that are based on these patterns including: JAVA Struts, several PHP libraries, ROR, Microsoft Smart Client Software Factory (CAB),

Microsoft Web Client Software Factory, ASP.Net MVC framework etc. The ASP.NET Web Forms MVP [12] implementation of the pattern was used for iKnow.

The MVC pattern aims at separating the user interface - UI (View) from its business layer (Model). The MVP is a presentation pattern based on the concepts of the MVC pattern. In the iKnow pattern implementation of MVP the presenter interacts with a service (controller) layer to retrieve/update the model. One of the major advantages of the MVP pattern is the easy integration of unit tests. In the development of iKnow the advantage was used developing several unit tests in the view interface and service layer for testing the presenter and the model.

The clear separation of concerns and responsibilities between layers and components imposed by the pattern in the development of the iKnow system benefited the project in many ways. The major benefit was in the testing phase which showed that software bugs were relatively low in number. Also, the bugs were rather easy to correct since functionalities were separated and the errors were easy to locate.

Another benefit of using the pattern was the increased productivity due to code reuse. The project development speed increased as the system grew bigger since each new form or functionality was composed of controls that were previously developed for the preceding modules. Similarly, the flexibility of the used data layer enabled easily integrating different interfaces, ranging from the web forms for users to the web services for other third party systems.

One of the drawbacks of using the MVP pattern is the added complexity of the project. Since iKnow is a relatively big project, the complexity of the used pattern added at the initial stages resulted in an eventually cleaner, easily maintainable and upgradable project in the later stages. While unnecessary in smaller application, MVP is useful in bigger projects producing a high quality code structure that is relatively bug-proof if the pattern is strictly followed. The complexity of the pattern paid out later in the agile development phase when multiple functionalities changes were easily implemented without introducing instabilities in the system.

Another drawback of the pattern is the longer learning period even for relatively experienced developers that have not previously faced this architecture. This resulted in harder recruitment of developers in the team and inability to easily enlarge the team in face of incoming deadlines. However, after a learning period, different in length for every developer, the benefits of the high quality code produced compensate for the late engagement.

The ASP.NET Web Forms MVP [12] implementation of the pattern was used for iKnow. WebFormsMVP is an open source project.

The framework supports AJAX. Our experiments show that the speed and responsiveness of the application is more than satisfactory.

Performance tests were performed over the system using a tool capable to simulate multiple users accessing the software (WAPT 7.1 http://www.softlogica.com/). **Fig. 2** shows the average number of http errors depending on the number of consecutive users using the system. It is obvious that the system is stable in the sense that the number of errors is constant and independent of the number of users. The test lasted for 25 minutes, starting from 10 up to 90 users, increasing the number of users every 5 minutes with a rate of 20. The right axis and the black lines of the graph show the number of users, and the horizontal axis shows time. The red line shows the average number of errors with labels shown on the left axis.

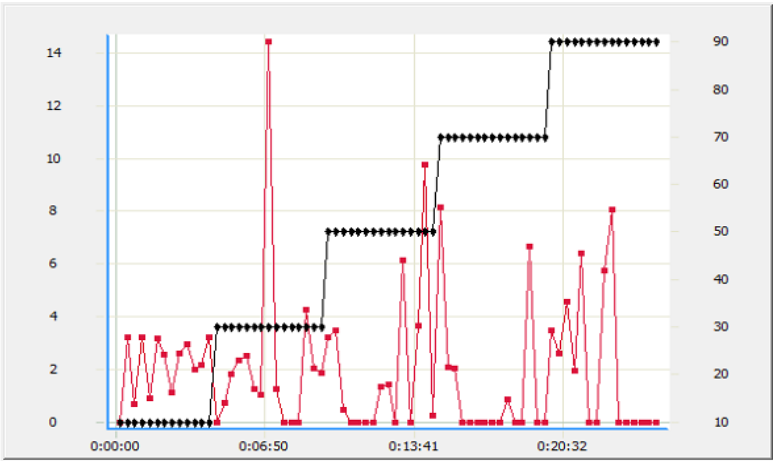


Fig. 2 HTTP errors in percent with increasing number of users

3.4 Software project architecture

All logical components in the system comply with the NTier concept, meaning division of the application in Layers and Tiers. The architecture of the system is presented in **Fig. 3**.

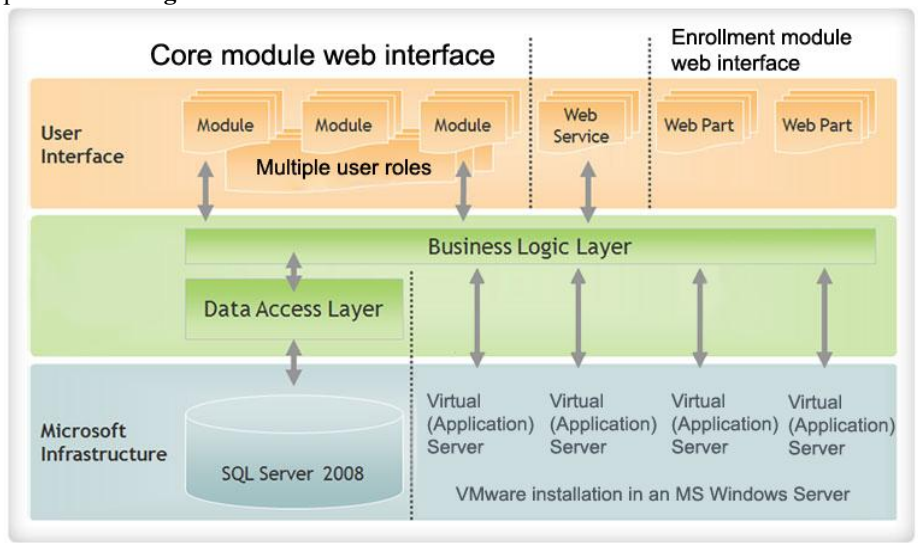


Fig. 3 Software architecture of the system

Several of the projects included in the complete iKnow solution are shortly described in the remaining of this chapter:

UniSS.DataModel is a Class library project containing the Entity framework model of the data. The model maps all tables, procedures and functions in classes and class methods. The model takes care of the overall communications and data flow from the database to the application and vice versa.

The audit log component is implemented in this project as well, logging all transactions from and to the database. The implementation is placed in the UniversityEntitiesCreator.cs class. Because the entity framework model is a partial class, and it is an extension to the model, it makes an addition to the OnSave event, by processing the XML of the changed data based on the old and new state of an entity.

UniSS.Repositories is a Class library project storing the classes for implementation of the Repository template. The classes of the implementation are in the subdomain (folder) Repos. The refactored entities that exist only in the business layer are in the subdomain (folder) Domain. For instance, if a list of students containing their index number, name and surname is to be rendered on the interface, it is not the complete Student entity with 50 columns that is retrieved, but rather a subdomain named StudentsRefactored, containing the aforementioned three columns. It is an optimization that decreases the operational memory used by the application as well as the time needed to process the data.

UniSS.Logic is a Class library project containing the implementation of the MVP template. The subdomain Models contains the models, Views contains the interfaces, and Presenters contains the presenters.

UniSS.Logic is a Web Forms application project in which all forms are separated according to their functionalities. The structure is based on the following hierarchy: Master Page, Web Page and UserControl. The user controls usually contain the entire business logic of the form and a Web Page implements the appropriate control. The Master contains the main layout.

The form named StudentsListReport.aspx is an example taken for illustrative presentation in this paper

Every form has a distinct Model, View and Presenter. All data that the form can manipulate is stored in the StudentsListReportModel model. The data consists of all control variables, the records needed for the form filters retrieved from the database, results of the search etc.

The IStudentsListReportView View component implements all the events in the form, for instance, when changing certain filter values the appropriate filtered data are retrieved and shown. When the selected value in a DropDownList component is changed, an event is triggered that refreshes the data in the grid. When clicking the Export button, the appropriate event is triggered, etc.

The presenter makes a connection of the view and the model, making an implementation of the event handlers of the view. When an event is fired, the appropriate repository calls a certain data retrieval function. The data is stored in an appropriate model variable

The function StudentsListReport defind in the repository is called, using LinqToEntities to query the database and retrieve data that is placed in the variable Students of the model.

In the user control named StudentsListReportControl in the UniSS.WebSite project, the appropriate events of the View are triggered, and after the successful

execution, the variables in the model are filled with data and bonded to the DataSource of the grid.

4 The User Interface

4.1 User interface specifics and implementation details

Systems offering electronic student services are complex information systems. They encompass multitude of functionalities provided to numerous users with different access privileges. Also, the users have different levels of familiarity to the system and the underlying processes. The user interface and it's performance is essential for the system's adoption as much as the systems reliability and scope of functionalities. [5],[6].

Agile development of a user interface is necessary due to multiple reasons:

- changing requirements [4]
- user satisfaction (constructive remarks by experienced users, habits gained by using legacy applications, trial period experiences , upgrade of technologies during development.

Users demanded several important features :

- grouping of functionalities on fewer forms for easy access and decreased need for navigation (**Fig. 4**))
- Screen size and resolution (conflicting) [2], [3]
- Simplicity and impossible to make an error [1]
- Automation of processes

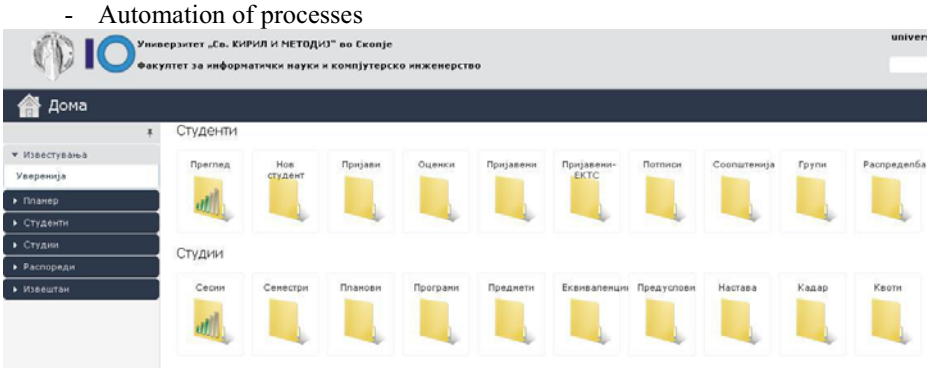


Fig. 4 Homepage – navigation through standard top and side menus or main part with folders

The system was developed as a web application with multiple user profiles and privileges. Users tend to avoid leaving the current page that they are working on, due to time-consuming navigation. Therefore the forms were heavily enriched with popups that could present additional information or provide means for inserting/updating data [7].

Standard web page popups were avoided as deprecated, and instead, hidden forms existing on the same web page were used (**Fig 6**, **Fig. 7**). Although effective in terms of work efficiency and user satisfaction, such forms tend to exponentially grow in complexity and size. The side-effects of such policies are complex forms prone to higher number of errors and extreme difficulty to test and validate.

Server side preparation of such forms presents a serious load on both the application and database server. Additionally, the size of the content itself is a burden to the Internet link since it generates a significant traffic. Since the demands are conflicting, a balance had to be reached for effective, but still lightweight forms that could be easily navigable. The advantages of AJAX and caching were heavily used to achieve the desired goals. For example, **Fig. 5** shows a form in which students choose the courses they will enroll during the semester. Multiple calculations are performed in real time generating the multiple selection list containing only courses that the student is allowed to take, the courses the student must repeat, the ECTS credit limitations in the semester, the financial implications of the selections, etc.

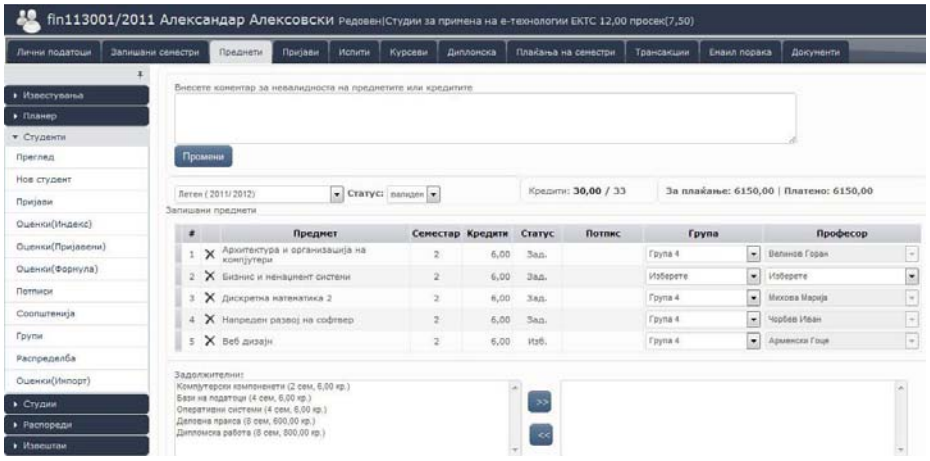


Fig. 5 Courses selection per semester, a backend complex form rendered easy to use

The screen size of the monitors used by the employees in the student services department varies from 15" up to 22". Therefore an adaptive interface had to be developed that is capable of using the benefits of large screens and resolutions, in the same time avoiding rendering small monitors useless. The content had to be visible and the interface usable in all screen sizes. Additionally, management used pads mainly for reporting, adding another layer of complexity, making the interface workable on touch screens and appropriate screen resolutions. Similarly students use Tablet computers or smartphones with variable, usually small resolution, demanding the interface to be robust and adaptable.

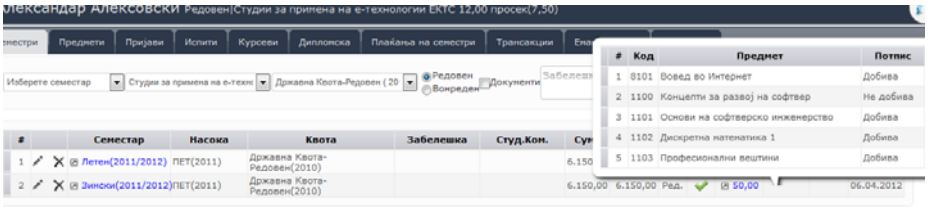


Fig 6 Overview of student semesters, expandable as needed with modal popups with additional data for SMS payments, courses enrolled in the semester, etc.

Although training is always thorough and unavoidable when deploying a complex information system such as this one, users tend to learn the system by "trial and error" and intuition. The user interface had to provide clear and short labels and messages, prevent from accidental deletions, and lead the user through the business processes. We have followed the design principle that simplicity is the key issue (**Fig. 7**).

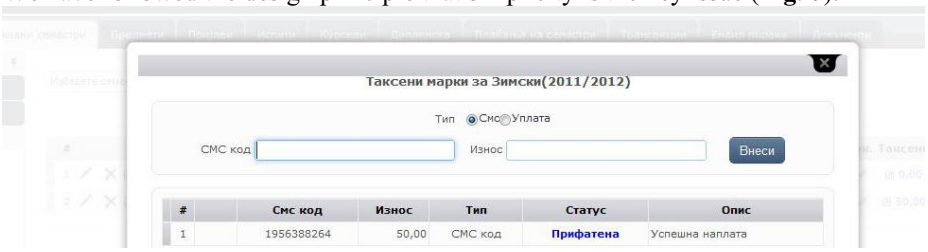


Fig. 7 Modal popup for SMS payment automation

Efficient use of the workforce in the student services department is only possible if the processes are as automated as possible, therefore bulk importing and fast input of grades for exams passed was developed.

4.2 User interface evaluation

In order to evaluate the usability and quality of the system, a questionnaire was developed and given to the end users to assess their opinions. The questionnaire consisted of 33 questions. Each question was answered by the users with a grade of 1-5. The user could answer with a 0 if he/she had no opinion on the matter.

The questions were organized in seven parts including: Accessibility; Layout; Navigation; Exception and status handling; User guidelines and online help; and Learning; and Content and Efficiency.

The results of the evaluation questionnaire were statistically analyzed and average grades were calculated for each of the questions. The users that participated belonged to 5 different faculties in the university bringing diversity of user backgrounds. The number of users questioned was 12. Out of the users questioned, 83.3% were female. Religious and cultural background of the users was also diverse. The answers provided were mostly positive.

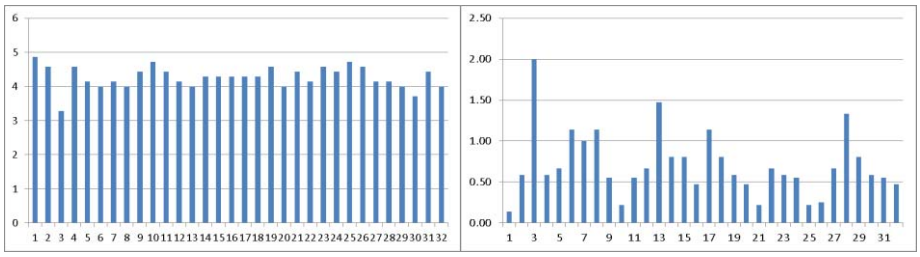


Fig. 8 a) Average grades given by the users in the questionnaire, b) Variance of average grades given by the users in the questionnaire

Fig. 8a presents the average grade awarded to each question by the users, while **Fig. 8**b presents the variance of the average grades.

The user satisfaction is evidently high, due to their participation in the user interface fine tuning and the efficient interface that resulted.

The Core component has generally been accepted as quite usable, easy to learn and work on. User's requirements have been implemented using agile development. Procedures and the user interface have been remodeled according to interviews with the initial users of the system. Adaptations of the interface have been made based on feedback from both teachers and student services employees. Students have confirmed the high usability of the application when asked.

The enrollment component is in the process of adaptation and customization based on the experiences during the first use of the system in August and September 2011. Problems in its use were noted, passed to the developers and the updated version is expected for the next term of enrollments (August 2012).

5 Evaluation of the software implementation

5.1 Detailed evaluation of the enrolment component

For the enrolment term in August-September 2011 the enrolment module was implemented with all planned functionalities except for the automatic transfer of the data for the enrolled students in the core module. Some faculties at the university completely relied on the iKnow system for enrolment (FCSE), while others used it in parallel with legacy systems, achieving identical results (Faculty of Natural Sciences, etc.) The faculties successfully completed the enrolment process using the iKnow system. Some smaller bugs noted during the enrolment process were addressed promptly by the developer that provided very agile response during the enrolment period. **Fig. 9** gives detailed statistical overview of the number of bugs in the Enrollment component.

Status	Total
Closed	333
In Progress	5
New	17
Pending	14
Rejected	20
Resolved	2
TOTAL	391

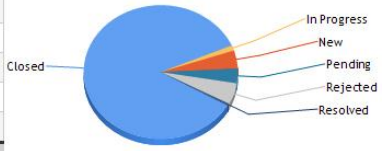


Fig. 9 Bug statistics of the Enrollment component

Student-candidates are generally satisfied with the solution. It's use is straightforward resulting in very few mails sent from students to the support team. The mails were often targeting data inconsistencies entered by the system administrators, and almost no mails were targeting issues of the functional system features and usability.

The processing of candidate applications by the enrollment committee was optimized for performance and accuracy. Two interfaces were developed to process applications: a wizard similar to the candidates interface for detailed analysis, and a short form for quick processing and updating. The later short form enhanced usability and significantly reduced the applications processing time for the enrollment committee. The experiences from the initial use are being implemented into the new version aiming at further enhancement of usability, reduction of complexity, discarding unnecessary steps and stages of each application. Automation of certain checks are also implemented (client side validators, consistencies of specific data types)

The main ranking of candidates for enrolment in different study programs was very rigorously tested prior to system roll-out, therefore no inconsistencies were detected.

At the focus of our interest in the nonfunctional requirements were the usability and the response times. The response time was measured during the testing phase and several optimizations were demanded. By the time the enrolment process started, the application had satisfactory response times. The response time averaged below 5 seconds for the most complex forms for fewer than 30 simultaneous requests, and below 10 seconds for the same forms for fewer than 50 simultaneous requests. The most frequently used forms by the students and the enrolment committee were optimized for its best usability. The main focus of optimization and design was aimed at the interface for student candidates, since the number of such users was to be measured in thousands. The strain of the servers was reduced to minimum, and the user satisfaction had to be high. Also, these users could receive no training; therefore they ought to face a trivially simple, yet fully functional error proof interface. Last minute optimizations were also made in the interface for the enrollment committees. Although of secondary importance, the speed of operation and the user satisfaction among the enrollment personnel was essential for complete adoption of the system in the entire university.

The measure of success of the implemented system is a very important issue in order to improve its functionalities in the future. We are using social media as Facebook and twitter to reach to the end users and gather their experience of their

eventual difficulties when using the system as well as gathering ideas for further improvements. In order to better explain the enrollment process to the candidates, social networks were used as a communication channel [23]. The channels were:

- Facebook page (Statistics shown on **Fig. 10**, **Fig. 11**)
- Twitter profile
- YouTube channel (**Fig. 12**)
- website (FAQ, Contact & Facebook live-chat)

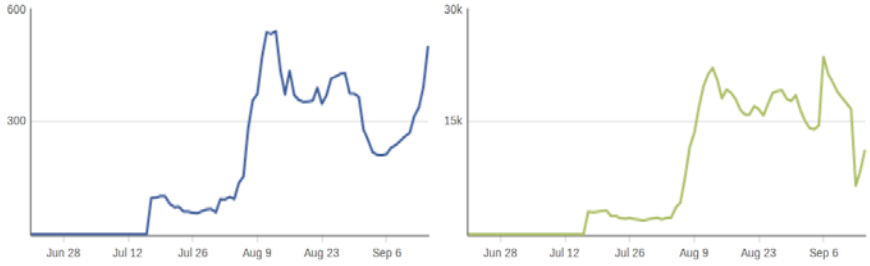


Fig. 10 Facebook - All stories, talking about stories (left) and viral reach (right) [23]

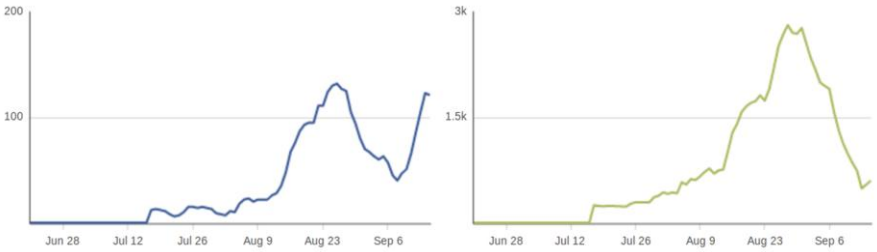


Fig. 11 Facebook – post by others, talking about posts (left) and viral reach (right) [23]

Fig. 12 Enrollment training video seen 803 times on the Youtube channel [23]

5.2 Detailed evaluation of the Core Student Services component

Most of the functionalities in the e-student services are fully implemented. Few of them are dropped out when implementing the system as unnecessary burden in this phase and several are pending - mostly reporting modules.

According to the specification document [9], there are 76 functionalities. Fully implemented are 46 functionalities, 24 are going to be implemented in the near future, 5 are not implemented in this phase, because other external systems should be established first, and 1 of them is discarded as irrelevant and an unnecessary burden to the system (administration of seminar thesis and reports, because mandatory seminar thesis are treated as any other course in the e-learning system). **Fig. 13** gives detailed statistical overview of the number of bugs in the Core component. The average time spent working on resolving software bugs is presented in **Fig. 14**.

Status	Total
Closed	774
In Progress	111
New	51
Pending	4
Rejected	2
Resolved	207
TOTAL	1149

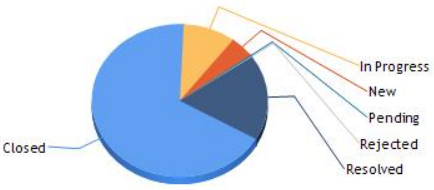


Fig. 13 Bug statistics of the Core component

AVERAGE TIME

Opening reported bugs	27.65 hours
Closing resolved bugs	23.71 hours
Resolving reported bugs	25.96 hours
Resolving rejected bugs	25.72 hours
Completing reported bugs	103.04 hours

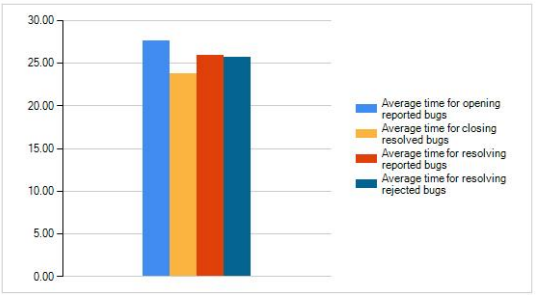


Fig. 14 Average time working od software bugs

Module for study programs and schedules

The question of status and quality of the system's functionality regarding the module for study programs and schedules in particular defining student programs, courses, prerequisites and rules for studies, is very important. These functionalities are implemented with special concern focused on making this interface easy to use with minimal steps in achieving results. Excel import with simple predefined templates was provided. Also, there are multiple different forms with the same functionality, offering options to various users to insert the data using several different methods, depending on the structure of the available data, user habits, the amount of data etc.

The functionalities of mapping of faculty staff to courses are also implemented. Teachers can be mapped in the courses menu, one teacher for all occurrences of the course in programs, or in the programs menu, with a different teacher for each course-

program link. The functionality of connecting equivalent courses is also implemented. Courses with different names or from different programs can be connected as equivalent so that students that change study programs during their studies can still keep their record of appropriate exams passed.

Student activities module

System's functionality regarding the student activities module with respect to the functionality of enrolment in a semester and selection of courses is implemented. Both students can do it by themselves or the student services employees can do that for them. The functioning is successful for both first time students that started using the system from the first moment of enrollment in the university as well as students that were migrated from legacy applications.

Forming groups at the beginning of each semester is a very important functionality in order to begin the course activities on time. This functionality is implemented. Students can be assigned to groups for each course they have selected, both individually, or multiple student at once, using excel import or multiple selection.

Module for administration

The system's functionality regarding the module of administration with respect to the administration of faculties and accredited study programs is implemented. Administration of members of the faculty is also implemented. Creation of users in the system is implemented in parallel. The administration of classrooms, rooms and laboratories is implemented. The administration of exams is part of the LMS, but the administration of earned ECTS credits and grades from exams passed is fully implemented. The functionality for storing and administration of personal records for students is implemented

Module for personal identification and access control

This module is implemented regarding the access control - currently, teachers can record the attendance of each student in the system. MS .NET membership access control is implemented. Based on special demands by the Faculty of computer Science and Engineering (FCSE), authentication was integrated with the Central Authentication Service used only by the FCSE users (both students and staff).

Module for presence monitoring and student activities is planned. The university is working in parallel on a distinct system for attendance control using RFID cards. Integration of both systems is planned.

Module for electronic payment and use of resources

Administration of payments by the students is implemented by automatic retrieval of SMS payments for administrative tax for students. Other payments are entered manually by student services department so far.

The administration of the use of learning management systems (LMS) is implemented by Moodle integration. In the iKnow system, teachers can demand creation of a Moodle course, and all students enrolled in the course in iKnow are automatically enrolled in the Moodle course as well.

Migration of old data, interfaces to other systems

The system supports easy migration of legacy data using imports of Excel templates filled with the old data. The imports are limited to students, exams passed, courses, study programs etc. However, FSCE being the first faculty that uses the system and having access to the database structure, achieved more thorough migration of enrolled semesters for legacy students and the courses taken in each semester.

The system incorporates web services that enable integration with external Learning Management Systems, Ministry of Education high school graduation storage systems, Biro of statistics. Also web services are available for developing external modules (mobile apps, etc).

6 Conclusion

This paper presents the details of the implementation of an advanced student oriented student services system that provides extensive list of functionalities to both students and staff using state of the art web software technologies. The integrated systems enables university level real time reporting and management. Students are provided with timely and secure information and opportunities to change information and their choices in the system by themselves anytime, anywhere. An interactive and intuitive user interface guaranties ease of use, reduced help and support effort and prolonged life of the system.

Systems like the one presented never stop to grow and change. Legislation changes along with new ideas to improve high education and provide students with new services and choices. Therefore the modularity and extensibility of the system in question is a key issue that ought to guarantee the systems future and growth. The build-in web services that await their use in native mobile applications are an example of such forward thinking. Every important piece of information in the system can be retrieved or changed using the secured web services that can easily be integrated in Android, iPhone or Windows Mobile apps. Such examples are being developed and will be readily available. Similar web services are developed to enable integration with external Learning Management Systems, Ministry of Education high school graduation storage systems, Biro of statistics etc.

Another forward thinking feature of the system is the idea of providing software as a service. In the era of evermore growing popularity of the concept of cloud computing, this architecture hosted on a cloud can be adapted with relatively minor effort to server multiple universities bringing all the known benefits.

References

1. Anthony Vance, Braden Molyneux,Paul Benjamin Lowry , "Reducing Unauthorized Access by Insiders through User Interface Design: Making End Users Accountable", Hawaii International Conference on System Sciences, January 2012, pp. 4623-4632
2. Jörg H. Mayer,Timm Weitzel , "Appropriate Interface Designs for Mobile End-User Devices--Up Close and Personalized Executive Information Systems as an Example", Hawaii International Conference on System Sciences, January 2012, pp. 1677-1686
3. Gang Huang,Daimeng Wang, "Adapting user interface of service-oriented rich client to mobile phones", Service-Oriented System Engineering, IEEE International Workshop on December 2011, pp. 140-145
4. Lu Xudong, Wan Jiancheng, "User Interface Design Model", Software Engineering, Artificial Intelligence, Networking, and Parallel/Distributed Computing, Eighth ACIS International Conference on, July 30 2007-Aug. 1 2007, Volume: 3, Page(s): 538- 543
5. Massila Kamalrudin,John Grundy, Generating essential user interface prototypes to validate requirements, "Automated Software Engineering, International Conference on", November 2011, pp. 564-567
6. Shinichi Inenaga,Kaoru Sugita,Tetsushi Oka,Masao Yokota, "Performance Evaluation of User Interfaces According to User Computer Skill and Computer Specifications", Intelligent Networking and Collaborative Systems, International Conference on, December 2011, pp. 446-449
7. Shinichi Inenaga,Kaoru Sugita,Tetsushi Oka,Masao Yokota , "A Preliminary Evaluation for User Interfaces According to User Computer Skill and Computer Specifications", P2P, Parallel, Grid, Cloud, and Internet Computing, International Conference on ,October 2011, pp. 295-298
8. David R. Karger, "Creating user interfaces that entice people to manage better information", Proceedings of the 20th ACM international conference on Information and knowledge management (CIKM '11), October 2011, pp. 1-2
9. Project Tempus JPGR 511342 – iKnow http://iknow.ii.edu.mk
10. Ivan Chorbev, Marjan Gusev, 'iKnow Student services" System documentation of Project Tempus JPGR 511342, 2010-2011, http://iknow.ii.edu.mk
11. Dejan Gjorgjevik, Marjan Gusev, "iKnow Enrolment module", documentation of Project Tempus JPGR 511342, 2010 – 2011, http://iknow.ii.edu.mk
12. http://wiki.webformsmvp.com/index.php?title=Main_Page (1.8.2012)
13. http://www.infragistics.com/community/blogs/todd_snyder/archive/2007/10/17/mvc-or-mvp-pattern-whats-the-difference.aspx (1.8.2012)
14. http://msdn.microsoft.com/en-us/magazine/cc188690.aspx (1.8.2012)
15. http://schoolcomputing.wikia.com/wiki/Student_Information_Systems (1.8.2012)
16. http://www.dmoz.org/Computers/Software/Educational/Administration_and_School_Management/ (1.8.2012)
17. http://www.microsoft.com/australia/education/schools/solutions-for-schools/customer-relationship-management.aspx (3.8.2012)
18. http://crm.dynamics.com/en-us/education (3.8.2012)
19. http://civicrm.org/node/586 (3.8.2012)
20. http://www.jenzabar.com/products.aspx?id=102 (3.8.2012)

21. http://www.sap.com/industries/higher-education-research/index.epx (3.8.2012)
22. Milos Jovanovik, Vladimir Zdraveski and Marjan Gusev, Social Networks for Customer Relationship Management, Molika, Bitola, CIIT 2012
23. Dejan Gjorgjevikj, Ana Madevska Bogdanova, Ivan Chorbev, Marjan Gusev, Implementation of electronic student services at UKIM, Molika, Bitola, CIIT 2012
24. Ivan Chorbev, Marjan Gusev, Dejan Gjorgjevikj, Sasko Ristov, User interface agile development and evaluation of the iKnow student services system, Molika, Bitola, CIIT 2012

Appendix A: MVP and MVC pattern detailed elaboration

There are several frameworks that are based on these patterns including: JAVA Struts, several PHP libraries, ROR, Microsoft Smart Client Software Factory (CAB), Microsoft Web Client Software Factory, ASP.Net MVC framework etc.

The MVC pattern aims at separating the user interface - UI (View) from its business layer (Model). The pattern separates responsibilities within three components:

- the view is responsible for rending UI elements,
- the controller is responsible for responding to UI actions,
- the model is responsible for business behaviors and state management.

In most implementation of the pattern, all three components can directly interact amongst each other. In some implementations the controller is even responsible for determining which view should be displayed (Front Controller Pattern).

The MVP is a presentation pattern based on the concepts of the MVC pattern. The pattern separates responsibilities within four components:

- the view is responsible for rending UI elements,
- the view interface is used to loosely couple the presenter from its view,
- the presenter is responsible for interacting between the view/model,
- the model is responsible for business behaviors and state management.

In some pattern implementations the presenter interacts with a service (controller) layer to retrieve/update the model. The view interface and service layer are often used to write unit tests for the presenter and the model.

There are a lot of benefits as well as drawbacks when using either the MVC or MVP pattern. The biggest drawbacks include the additional complexity and learning curve. While the patterns may not be appropriate for simple projects; advanced software systems can greatly benefit from using a pattern.

Several major benefits from using design patterns are presented in the list below:

- The presenter or controller serves as an intermediary between the UI code and the model allowing the view and the model to evolve independently of each other.
- Clear separation of concerns and responsibilities
- Testing - by isolating each major component it is easier to write unit tests. This is especially true when using the MVP pattern which only interacts with the view using an interface.

- Code Reuse - especially true when using a complete domain model and keeping all the business logic and state management procedures in appropriate modules.
- Hide Data Access - Separating and distancing from the data access enables replacing the database platform in future implementations and other independent changes in the database.
- Flexibility and Adaptability - A properly design solution using MVC or MVP can support multi UI and data access technologies at the same time.

The key advantages of the MVP pattern include:

- The view is more loosely connected to the model. The presenter is concerned with binding the model to the view.
- Easier to unit test because the interaction with the view is channeled through an interface
- Usually one view is mapped to one presenter; however, complex views may have multiple presenters.

The ASP.NET Web Forms MVP [12] implementation of the pattern was used. WebFormsMVP is an open source project created by Tatham Oddie and Damian Edwards.

The WebFormsMvp open source framework utilizes the following components [12]:

- Presenter: This is a class that inherits from Presenter or Presenter<T>.
- View interface: This is an interface that inherits from IView or IView<T>.
- View Implementation: The page or user control that implements the view interface.
- Event Arguments: A custom event argument class is the way to pass data from the view to the presenter.
- Model: This is a class that contains the data needed

Appendix B: Questionnaire

The questions were as follows:

Accessibility

1. How do you judge the information of the launch and the training?
2. How do you access the process of registration?
3. Does your browser display all information correctly?
4. Is site load time appropriate to content and response?

Layout

5. Text-to-background contrast
6. Is the font size and style easy to read?
7. Does the site have a consistent look and feel?
8. If you have a disability regarding your eyesight: Is the content readable?
9. Is the label location and format consistent?

Navigation

10. Are the major parts/menus of the site directly accessible from the main page

11. Are the navigation labels clear and descriptive?
12. Is the workflow navigation consistent and easy to identify?
13. Is the respective location within the process (site) transparent?
14. Is the site search easy to access?
15. Is the exit point clear on each page?
16. Does it require minimal steps in sequential menu selection?

Exception and status handling

17. Are the messages regarding status clear and descriptive?
18. Are the messages regarding exceptions/errors clear and descriptive?
19. Position of messages on screen is good

User guidelines and online help

20. Is the site designed to require minimal help and instructions?
21. Is the help and instruction information easily accessible?
22. Is there an easy channel available to communicate with an administrator?

Learning

23. Easy to learn to operate the system
24. Easy to explore new features by trial and error
25. Easy to remember names and use of commands

Content and Efficiency (refers to all information i.e. field explanations)

26. Is the content understandable?
27. Is the content well-structured and correlates to your requirements?
28. How do you evaluate the support of the system?
29. Do you observe an increase in efficiency?
30. Does the system provide a sufficient number and quality of reports?
31. It is easy to use.
32. What is your overall evaluation of the system?

Comparative Analysis of the Government ICT Projects in Macedonia, Estonia and Slovenia

Smilka Janeska-Sarkanjac

Ss Cyril and Methodius University in Skopje, Macedonia
Faculty of Computer Science and Engineering
smilka.janeska.sarkanjac@finki.ukim.mk

Abstract. This comparative analysis of the government ICT projects of Macedonia, Estonia and Slovenia faces and compares the projects in the three countries. All three countries have similar size, similar historical and socio-economic background, but Estonia and Slovenia are significantly ahead in the development of information society and in economic development in comparison to Macedonia. Our thesis is that part of this successful economic development is due to the government initiative and decisions regarding the choice and implementation of ICT projects (infrastructural and e-government). We have found six crucial preconditions for development of ICT that have tangible impact on economic growth: government is the leader in introducing information society; government is early adopter of ICT; government is proactive; government recognizes ICT as growth tool; government provides the funding; government provides quality project management.

Keywords: ICT for development, ICT application, ICT projects, public sector, Macedonia, Estonia, Slovenia

1 Introduction

The proper application of ICT should lead to the people's wellbeing and raising of the level of empowerment and participation. Technology must be applied with sensitivity to social, economic, and political contexts [8].

The question whether ICT investments contribute to the economic growth had a vague answer for a long time. The well-known Solow's productivity paradox 'You can see the computer age everywhere but in the productivity statistics' [18] started to resolve in the early 2000s. There have been a number of macroeconomics studies demonstrating the positive impact of ICT on economic growth and development, conducted on the data from developed countries [19, 4, 14]. On the other hand, most of the authors arrived at negative or mixed results regarding economic growth or returns to capital for developing countries [5, 16, 17]. More recent research started to indicate positive economic returns to ICT investment for developing countries also [9] and ICT driven returns to productivity gains momentum [15].

For example, Hosman, Fife and Armey have found that a 100% increase of ICT expenditure per capita produces an additional 9% increase in the growth rate [9]. An-

S. Markovski, M. Gusev (Editors): ICT Innovations 2012, Web Proceedings, ISSN 1857-7288

other important ICT segment, especially for underdeveloped countries, penetration of mobile phones, has contributed significantly to economic growth. Fuss, Meschi and Waverman [19], looked at 92 countries, both developed and developing, to estimate the impact of mobile phones on economic growth for the period 1980 to 2003. They found that a 10% difference in mobile penetration levels over the entire sample period implies a 0.6% difference in growth rates between otherwise identical developing nations.

These recent macroeconomics researches do prove that ICT investments lead to economic growth, but to gain insight into how ICT can contribute to the development of a country, we need a better understanding on the way ICT is adopted, how ICT funds are employed, or distinguish the characteristics of specific projects that make them likely to succeed, and make them catalysts for economic growth.

There is strong evidence that e-government reforms are more likely to be a top-down process rather than a result of citizen demand, even in highly developed countries as USA [21]. In the case of developing countries, this is even more emphasized, because of the modest number of e-services offered by the private companies, and underdeveloped awareness of their advantages. This puts even more responsibilities in the hands of governments of developing countries.

Important obstacle in introducing ICT in a developing country is the resistance of political or business elites against new technologies. It is argued that new technologies could be extremely disruptive. They sweep aside old business models and make existing skills and organizations obsolete. They redistribute not just income and wealth but also political power [1]. The political-business nexus of the socialist regime elite that held power in the transition considered the introduction of new technology a threat. New technology brings transparency and empowerment to the citizens, which is not acceptable.

There are lots of questions concerning the challenges a government confronts in its effort to find the proper ICT strategy. One of the fundamental questions in formulating an ICT strategy for development in a country is what kind of project a government should choose to develop. We will try to offer some of the answers by comparing the most important ICT projects in Macedonia, Estonia and Slovenia. Besides that, we will consider nine horizontal enablers defined by EU Directorate General for Information Society and Media as infrastructural elements that provide foundations for robust, streamlined and sustainable e-government services [6].

Our thesis is that the factors that led to the differences in economic development of the three countries are twofold: (1) historical legacy and geo-political circumstances (path-dependence), including relatively different social, economic and political context and (2) "subjective" factors that largely depend on the decisions made by the governments. We will focus on the government decisions regarding the ICT projects (infrastructural and e-government) run by the governments in the three countries, and the other factors we will hold constant for the purpose of this research.

2 Basic Information

Macedonia, Estonia and Slovenia are of similar size and have a similar history – all of them were established as independent states after the fall of communism: all of them

were subject to a socialist planned state economy, more or less lacked private initiative, and experienced the transition from socialism to capitalism.

Slovenia and Estonia, unlike Macedonia, were the most economically developed regions in their former federal states (although Slovenia was at a considerably higher level in this regard), with geographical proximity to Western Europe: Slovenia borders Austria and Italy while Estonia has a maritime border with Finland [2]. Macedonia was one of the least developed regions in former Yugoslavia, and its neighboring countries were also not as developed. Karch [11] states that geographical proximity has played a traditional role in explaining policy diffusion, at least until recently.

Some of the basic data for the three countries are shown in Table 1.

Table 1. Estonia, Slovenia and Macedonia, basic data, 2010

Criteria	Estonia	Slovenia	Macedonia
Area	45,000 km2	20,273 km2	25,713 km2
Ethnic groups	Estonian 69% Russian 25,5% Others 5,5%	Slovenes 83,1% Serbs 2,0% Others 14,9%	Macedonian 64,2% Albanian 25,2% Others 10,6%
Population	1,333,000	2,018,000	2,056,000
GDP	$19,083 billion (2010) $3,965 billion (1995)	$49,158 billion (2010) $14,386 billion (1995)	$9.300 billion (2010) $3,400 billion (1995)
GDP growth	3,1%	1,2%	0,7%
GDP per capita (PPP)	$18,518	$28,030	$9,727
Unemployment rate	5% (2008) 16,9% (2010) 11% (2011)	7,5% (2008) 10% (2010) 12% (2011)	35% (2008) 32% (2010) 31,3% (2011)
Education expenditure	4,9%	4,95%	4,42%
Expenditure on R&D	1,44%	1,86%	0,2%

Estonia, in less than 20 years, has become one of the leading countries in Eastern Europe. The Estonian economy has experienced almost double-digit growth for years (i.e. 11.74% in 1997, 9.974% in 2000, 7.516% in 2001, 10.562% in 2006), which applies that it was one of the fastest growing economies in the world, until the world crisis in 2008. Today it successfully copes with the effects of the crisis and achieves positive growth rates again. Much of this success of Estonia is due to the application of ICT, which plays an important role in the country.

Slovenia, in comparison to Estonia, experienced modest growth rates, but the starting position, at the beginning of 1990s was much better then ones of the other two countries. The highest growth rate was in 2007, 6,873% and the lowest was in 2009, -8,129%. During 1990s and 2000s, the average growth rate was 3,44%. Because of its good starting position (and traditional openness and trade with the West countries), Slovenian economic, social and political developments have been stable and relatively successful (its exports amount 60% of GDP), maintaining the biggest GDP in the region - despite having lower economic growth. Slovenians speak many languages (proficiency in English is one of the highest in the EU) and they are also prone to learning new technologies [3].

Macedonia, in comparison to Estonia, has experienced modest growth rates in the past 20 years, from negative ones in the beginning of the 1990s (i.e. -7,5% in 1993), to the highest one, 5,9% in 2007. In comparison to Slovenia, the starting position of

Macedonia in 1990s, with GDP about four times lower then Slovenia's was also rather unfavorable.

According to the research undertaken by the World Economic Forum on the use of information technology in 142 countries [20], Estonia ranks 24th in the Networked Readiness Index and it is the highest-ranking Central and Eastern European country. Slovenia takes 37th place, and Macedonia takes 66th place.

3 ICT projects in Estonia

At the beginning of the 1990s, when Estonia regained its independence, it was relatively a technologically backward country. The industrial machinery from the Soviet era was outdated, and state infrastructure in terms of institutions and people had to be built up from scratch. However, foreign direct investments started coming to Estonia. Krull [12] states that crucial factors supporting the development of Estonian information society and growth were:

- Building modern telecommunication infrastructure;
- The Tiger's Leap project and Estonian Educational and Research Network, back in 1993, provided schools with computers and Internet, from which generations of advanced ICT users rose, bringing their knowledge and their habits to their families, therefore producing a spillover effect;
- Early adoption of regulation related to information society;
- Introduction and the raising of public awareness for government programmes such as e-government, Village Road, x-Road, ID cards, etc.;
- The collaboration among the government, private companies and NGOs for various ICT programmes, such as the Tiger Leap programme.

According to many analyses, the government was the leader in introducing information society in Estonia, together with a pro-active ICT sector and advanced ICT user population. The business sector and NGO followed when they found their own interest in ICT projects run by the government. The most successful projects run by the government are the following [7]:

- Electronic ID card, which is used by almost 90% of the population. It serves as an identity document and as a travel document within the EU. It serves as a pass to almost every e-service in Estonia, to e-banking, e-elections, buying transportation tickets, e-taxes, e-education, e-health, etc.
- Mobile phone applications – m-parking, m-ticket for public transport, m-banking.
- E-taxes – Estonia's tax board offers online a pre-completed tax form, which enables easy and fast submission of taxes by citizens and companies. In 2011, over 93% of the income tax declarations were presented via e-tax system.
- E-elections – since 2005, Estonians, among the first countries, have been given the opportunity to vote via Internet, using the ID card or mobile phone as identification. In the 2011 parliamentary elections 24.3% of the people who voted used the e-voting system.
- E-business registration – full e-service for a registration of a new company.

- E-banking – bank sector was a fast follower to the introduction of information society by the Estonian government with the e-banking project. It was widely accepted, and currently 98% of the bank transactions are received through e-banking.
- E-ticket for the public transport is paid via Internet and is registered on the citizen's personal ID card.
- Digital prescriptions – integral information system that keeps record of the medical prescriptions in a central database, and enables patients to get their prescribed medicine in a pharmacy only with an ID card. It was launched in January 2010.
- E-health record – medical information system started in 2010, which contains information on diagnoses, doctor's visits, tests, treatments, prescribed medications etc. There is a patient portal that can be accessed with patient's ID card.
- E-school – since 2003, parent-teacher communication is facilitated with the portal e-School for all Estonian schools. The grades of the students can be tracked, their absence from classes, the content of their lessons etc.
- University via Internet – the results from the state exams are kept in an information system, together with the high school grades. Students may submit applications to universities via the state's internet-based application system, using the former data.

Important factor in developing e-services was early introduction of X-Road (2001), the data exchange layer of the state information system, which included a complex security solution: authentication, multi-level authorization, a high-level log processing system, encrypted data traffic with time stamps, a warning system for servers against cyber attacks etc. An important principle applied from the very start of the X-Road is its service-oriented architecture. Besides basic 20 e-services, in 2008 Estonia had over 800 e-government services for citizen and companies, being second after Austria in EU in terms of fully electronic services, according to the Capgemini survey [10]. All the common horizontal enablers, according to EU Directorate General for Information Society and Media, are available in Estonia, as one of six leading EU countries.

4 ICT projects in Slovenia

With 95% full online availability of e-government services Slovenia is above the EU average of 82% [6]. Slovenia develops its e-government services according to EU recommendations, focusing on 12 services for citizens and 8 for businesses, which are defined as priority ones:

- Income taxes: declaration, notification of assessment – there is prefilled tax declaration for taxpayers as in Estonia, but they cannot use their ID card for authentication and authorization, but a qualified certificate issued by any registered certification authority in the country.
- Job search services by labor offices – there are two online job search services, by Employment Service of Slovenia and by the Ministry of Public Administration.
- Social security benefits –online processing of unemployment benefits, child allowances, medical costs (reimbursement or direct settlement), student grants. Not all of the social benefits are available online.

- Personal documents: passport and driver's license – there is information on the application process and email reminders on the expiration date of passports; otherwise the process is conducted in traditional manner. Renewal of a driver's license is a full online service that includes paying online, and receiving the new driving license by post. This e-service holds one of the lowest grades regarding online sophistication scores in the country - 50 out of 100 in 2010.
- Car registration (new, used, imported cars) – full online service.
- Application for building permission– full online service.
- Police report (e.g. in case of theft) - Since 2004, citizens can report crimes to the police online. Authentication with a qualified digital certificate is required. Since 2009, an anonymous denunciation of corruption to the Police has been enabled.
- Public libraries (availability of catalogues, search tools) - full online service that contains over 3 million bibliographic records, with a booking system.
- Certificates (birth and marriage): request and delivery - full online service that can be used by all residents equipped with qualified digital certificates.
- Enrolment in higher education/university – online application.
- Announcement of moving (change of address) - users need to send the electronically signed application form together with requested enclosed documents.
- Health related services (interactive advice on the availability of services in different hospitals; appointments for hospitals) – information services. Lowest ranking grades regarding online sophistication scores in the country – 32/100 in 2010.
- Social contributions for employees – full online service.
- Corporate tax: declaration, notification – full online service, same as for citizens.
- VAT: declaration, notification – full online service.
- Registration of a new company – full online service.
- Submission of data to statistical offices – full online service.
- Customs declarations – full online service.
- Environment-related permits (incl. reporting) – mainly informational web sites, some electronic services for obtaining environment-related permits.
- Public procurement - portal was established in 2007. It supports prior information notice, contract notice and contract award notice, as well as their amendments, tender documentation and relevant questions, answers and explanations. In 2009 the portal was upgraded with an additional platform for e-Submission, e-Tender evaluation and e-Auctions.

Besides these 20 e-government services, there are about 600 additional e-services in the country. Out of 9 measured horizontal enablers, 6 are available in Slovenia. These are: E-ID, Authentic Sources, Secure e-Delivery, Architecture Guidelines, Catalogue of Horizontal Enablers and E-Payment. The following enablers are not yet in place: Single Sign on, E-Safe and Open Specifications.

5 ICT projects in Macedonia

In 2005, the Republic of Macedonia has established the portal Uslugi.gov.mk as the single point of access to information and services of the government. The portal is a

result of the government's efforts to create a more efficient and transparent administration by presenting all available services for both citizens and businesses to the public. The assessment of the 10 basic services recommended by EU, and offered by the Macedonian government is as follows [13]:

- Income taxes: declaration, notification of assessment – there is no prefilled personal tax declaration for tax payers as in Estonia and Slovenia. Companies may pay the taxes online. Since 2012 there is also an online form of e-taxes for citizen. The percentage of tax declarations presented via e-tax system is very small, because of the small number of owners of the qualified certificate issued by registered certification authority in the country.
- Job search services by labor offices – there are two institutions in this field, Employment Service Agency and Administration Agency. The former one offers online submission of the data of new employments and the termination of employments, and the latter one offers online information on job openings and online job application. No other online services are offered.
- Social security benefits – other than online application for student dormitories, loans and grants, there is no other e-service regarding social security benefits. At best, there are information and downloadable forms for the social benefit.
- Public libraries (availability of catalogues, search tools) - online service that contains over 500 000 bibliographic records on book and non-book materials in Macedonian libraries.
- Health related services – G2B information on required documents or downloadable forms, no G2C online services. Information system on e-health record is under construction, health e-cards are being distributed to patients currently.
- Social contributions for employees – full online service.
- Corporate tax: declaration, notification – full online service.
- Registration of a new company – full online service.
- Customs declarations – Single Window for Export/Import licenses and tariff quotas system (EXIM) is online. Construction of the integral Customs information system with online services in progress.
- Public procurement - Public procurement is one of the most advanced part of e-services in Macedonia. Portal was established in 2007, and it is a one-stop-shop for public procurement in the country, which streamlines complex procedures and facilitates interaction between businesses and government institutions. It is also compliant with the European Union Directives and supports all forms of public procurement, including electronic auctions. Public institutions and businesses register with the system and obtain the obligatory digital certificates for posting tenders or sending bids. Since January 2008, e-Auctions have also become part of it.

The other 10 basic e-services recommended by EU, and listed in the previous chapter are not available at present. Most of them offer static information about the service to the citizen and businesses, or downloadable application forms.

Other significant projects regarding ICT in the public sector are: a computer for every student project; document management system for all the ministries; electronic land registry system, launched by Real Estate Cadastre Agency (May 2010), currently available for the city of Skopje and few other towns, and working on including entire

country; electronic grade-book, mandatory for all schools since 2012/2013; web publishing of school textbooks (April 2010); m-service which enables the payments of administrative fees by mobile phone (July 2011); m-parking; 680 Internet kiosks with additional wireless Internet in rural areas (January 2010); free of charge ICT courses; free of charge internet clubs; system for ranking of applicants for 'social apartments'; system for ranking of applicants for international cargo transport licenses (2007); budget planning system for budget users (www.e-budget.gov.mk); e-democracy – integral system for parliament document and process management, etc.

Out of 9 measured horizontal enablers, only 2 are available in Macedonia: e-ID and e-Payment. The other 7 are either not yet in place or are used only within the individual projects.

According to World Economic Forum in 2012 [20], the Government Online Service Index that assesses the quality of government's delivery of online services on a 0-to-1 scale, Macedonia is ranked 69th out of 142 countries with index 0.32.

6 Discussion

Given the experiences regarding the role of ICT in the public sector in Estonia, Slovenia and Macedonia, we can make several conclusions about the role of ICT in the development growth paths of the three countries. As mentioned before, we focused solely on the comparison of selected ICT projects by the governments of the three countries, and not on the analysis of historical, political, social, and economic context of each country that requires more comprehensive approach.

Our insight shows that all the contries put the accent on services. Roughly said, ICT in the eyes of the users equals e-services of the government. However, the differences between the three countries are more than obvious. To put it more clearly, in Macedonia 10 basic e-services recommended by EU are not available, unlike in Slovenia or Estonia. In Estonia all horizontal enablers are functional, in Slovenia 6 out of 9 measured horizontal enablers are available, but only 2 in Macedonia.

Why is that so? Estonia has focused on ICT since its independence, and has been an early adopter with a consistent policy of promoting ICT use and investing in ICT infrastructure, as it has realized the potential benefits of ICT in enabling economic growth and development of the country. We have to mention again the proximity of technologically advanced Finland and Sweden, and population with high level of technical education, as a fertile ground for wide ICT adoption.

We can say that Slovenian leaders consider ICT as one of the growth tools, but not the main one. They lack the Estonian proactive behavior, they follow the EU recommendations, and that works well for Slovenia, because of its good starting position. The differences in ICT adoption in Estonia and Slovenia follow the differences in political reforms after gaining independence: Estonia directed radical changes in the sense of the liberalization of society, while Slovenia was oriented toward a so-called "gradualism" [2].

By contrast, Macedonia has been a late adopter of ICT (it lags at least 10 years behind Estonia). Notwithstanding the historical and geo-political differences between Macedonia and the other two countries, this has been, in our view, the main difference between Macedonia, Slovenia and Estonia respectively, as to ICT. Therefore, it is

very important to emphasize that ICT initiatives had not been undertaken by the Macedonian government until 2005. Starting 2006, a minister without portfolio in a newly formed Macedonian government was assigned to an information society development, and in 2010 a Ministry of Information Society was established.

Despite of the initiatives of the government, it appears that ICT has still not been recognized as a powerful growth tool, as it is in Estonia and Slovenia. Macedonian investment in ICT is significantly below Estonian and EU average (2,9% in Estonia, 2,7% in average in EU27, and 0,84% in Macedonia in 2009). Admittedly, individual, often successful ICT projects are undertaken, but there is a general lack of prerequisites: solid infrastructure, common registries, common data exchange layer, interoperability elements, most of the horizontal enablers, to number just a few. The ICT policies of the country are, though, of high quality. Laws regarding information society are also adopted to a large extent. On the level of written strategies and documents Macedonia is not behind the developed EU countries.

One of the first actions that the government with a vision of developed ICT should undertake is a proper prioritization of ICT projects for development; that is to say, the government should make more serious efforts to provide funds for the projects, technical expertise, governance structures, proper management of the projects, and regulation. The proactive behavior of the government will stay a set of wishes if the projects are not funded, if they are not backed by expertise and management.

7 Conclusion

ICT4D sets complex goals ahead, both as outcomes and as impact to a society. The effects for the economy appear where ICT users utilize the technology - create new knowledge, save time, create new jobs, new livelihoods, redesign business processes. Our research has analyzed the effects of ICT adoption and use in three countries with different approaches to ICT. Estonia was seen to be an early and intensive adopter of ICT, then Slovenia as first follower, whereas Macedonia lagged in the adoption of ICT and then did not adopt ICT as intensively.

Our results suggest that Estonia's approach was the most effective in creating economic growth; Slovenia has successfully developed its e-government services, but is behind Estonia. As to Macedonia's macroeconomic results that could be connected to development of ICT, we can say that they are still unsatisfactory.

These outcomes from ICT adoption as an impact to an economic growth were result of government's decisions in all three countries to invest in ICT. What makes the differences between them is the careful and elaborate selection of ICT projects and, which is, perhaps, most important, their successful execution.

We can pick out six crucial preconditions for development of ICT that will have tangible impact on economic growth: government is the leader in introducing information society; government is early adopter of ICT; government is proactive; government recognizes ICT as growth tool; government provides the funding; government provides quality project management.

The proactive government of Republic of Macedonia recognizes ICT as growth tool, but has funded poorly the ICT projects. In this respect, much more is needed to be done since Macedonia has to compensate the starting position of a late adopter.

References

1. Acemoglu, D., Robinson J.: 10 Reasons Countries Fall Apart. Foreign Policy july/august 2012, http://www.foreignpolicy.com/articles/2012/06/18/10_reasons_countries_fall_apart
2. Adam, F., Tomšič, M., Kristian, P.: Political Elite, Civil Society, and Type of Capitalism. Estonia and Slovenia. East European Quarterly, Vol. 42, Issue 1 (Spring), 43--67 (2008)
3. Ala-Mutka K., Gaspar P., Kismihok G., Suurna M., Vehovar V.: Status and Developments of eLearning in the EU10 Member States: the cases of Estonia, Hungary and Slovenia. European Journal of Education, Vol. 45, No. 3, 2010, Part II, 494--513 (2010)
4. Cronin, F. J., Colleran, E. K., Parker, E. B., Dollery, B.: Telecommunications infrastructure investment and economic development. Telecom. Policy, 17(6), 415--430 (1993)
5. Dewan, S., Kraemer, K. L.: Information technology and productivity: Preliminary evidence from country-level data. Management Science, 46, 548--562 (2000).
6. Directorate General for Information Society and Media: Digitizing Public Services in Europe (2010)
7. Estonian Ministry of Foreign Affairs: Fact Sheet 2012 – Estonia Today, Tallinn (2012)
8. Heeks, R.: ICT4D 2.0 The Next Phase of Applying ICTs 4 International Development. Computer, June 2008 (Vol. 41 #6), http://www.lirne.net/2008/07/ict4d-2/
9. Hosman, L., Fife, E., Armey, L.E.: The case for a multi-methodological, cross-disciplinary approach to the analysis of ICT investment and projects in the developing world. Information Technology for Development, 14, 308--327 (2008)
10. Kalja, A., Robal, T., Vallner, U.: Towards information society: Estonian case study. In: Proceedings of PICMET '09: Technology Management in the Age of Fundamental Change, 3218--3225. Oregon, USA (2009)
11. Karch, A.: Emerging issues and future directions in state policy diffusion research. State Politics and Policy Quarterly, vol. 7, no. 1, 54--80 (2007)
12. Krull A.: ICT Infrastructure and E-Readiness Assessment Report: ESTONIA. PRAXIS Center for policy studies, Tallinn (2003)
13. Ministry of Information Society and Administration of Macedonia, http://www.mioa.gov.mk
14. Norton, S.: Transaction costs, telecommunications, and the microeconomics of macroeconomic growth. Economic Development and Cultural Change, 41(1), 175--96 (1992)
15. Papaioannou, S., Dimelis, S.: Information technology as a factor of economic development: Evidence from developed and developing countries. Economic Innovation and New Technology, 16(3), 179--194 (2007)
16. Pohjola, M.: Information technology and economic growth: A cross country analysis. In M. Pohjola (Ed.), Information technology and economic development. Oxford University Press, Oxford (2001).
17. Seo, H. J., Lee, Y. S.: Contribution of information and communication technology to total factor productivity and externalities effects. Information Technology for Development, 12(2), 159--173 (2006)
18. Solow R.: We'd better watch out. New York Times Book Review, July 12, 36 (1987)
19. Waverman, L., Meschi, M., Fuss, M.: The impact of telecoms on economic growth in developing countries. Vodafone Policy Paper Series, Number 2, 10--23 (2005)
20. World Economic Forum: The Global Information Technology Report 2012 – Living in a Hyperconnected World. Geneva (2012)
21. Yun, H., J., Opheim, C.: Building on Success: The Diffusion of e- Government in the American States. Electronic Journal of e-Government Volume 8 Issue 1, 71--82 (2010)

Quasigroup-based Hybrid of a Code and a Cipher

Victor A. Shcherbacov

Institute of Mathematics and Computer Science
of the Academy of Sciences of Moldova, Academiei str. 5,
MD−2028 Chişinău, Moldova

Abstract. We construct quasigroup-based hybrid of a code and a cipher.
2000 Mathematics Subject Classification: 94A60, 20N05, 20N15.

Keywords: cipher, code, quasigroup, T-quasigroup, orthogonality, n-ary groupoid, system of orthogonal n-ary groupoids

1 Introduction

We construct quasigroup-based hybrid of a code and a cipher and give an algorithm that describes this construction. Some results presented in this paper are taken from [18].

Hybrid idea is sufficiently known, see, for example, [16], [17]. Following Markovski, Gligoroski, and Kocarev [9], [10], we name such hybrid as a cryptcode.

Author chooses "example" style for this paper in order to make it accessible for engineers and students.

Definition 1. *A T-quasigroup (Q, A) is a quasigroup of the form $A(x, y) = \varphi x+\psi y+c$, where $(Q, +)$ is an abelian group, φ, ψ are some fixed automorphisms of this group, c is a fixed element of the set Q [8], [15].*

Theorem 1. *A T-quasigroup $(Q, \cdot)$ of the form $x \cdot y = \alpha x + \beta y + c$ and a T-quasigroup $(Q, \circ)$ of the form $x \circ y = \gamma x + \delta y + d$, both over a group $(Q, +)$, are orthogonal if and only if the map $\alpha^{-1}\beta - \gamma^{-1}\delta$ is an automorphism of the group $(Q, +)$ [14].*

Denote elements of the group $Z_2 \oplus Z_2$ as follows: $\{(0; 0), (1; 0), (0; 1), (1; 1)\}$. The group $Aut(Z_2 \oplus Z_2)$ consists of the following automorphisms:

$$\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix}, \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}, \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix}, \begin{pmatrix} 0 & 1 \\ 1 & 1 \end{pmatrix}$$

Denote these automorphisms by the letters $\varepsilon, \varphi_2, \varphi_3, \varphi_4, \varphi_5, \varphi_6$, respectively.

Notice that $\varphi_2^2 = \varphi_3^2 = \varphi_4^2 = \varepsilon, \varphi_5^2 = \varphi_6, \varphi_6^2 = \varphi_5$. It is known that $Aut(Z_2 \oplus Z_2) \cong S_3$ [6], [7].

For convenience we give Cayley table of the group $Aut(Z_2 \oplus Z_2)$.

$\cdot$	ε	φ_2	φ_3	φ_4	φ_5	φ_6
ε	ε	φ_2	φ_3	φ_4	φ_5	φ_6
φ_2	φ_2	ε	φ_5	φ_6	φ_3	φ_4
φ_3	φ_3	φ_6	ε	φ_5	φ_4	φ_2
φ_4	φ_4	φ_5	φ_6	ε	φ_2	φ_3
φ_5	φ_5	φ_4	φ_2	φ_3	φ_6	ε
φ_6	φ_6	φ_3	φ_4	φ_2	ε	φ_5

Information on codes can be found in [4].

2 Construction

Code part. We shall use a code given in [13, Example 19]. Let's suppose that the symbols x, y are informational symbols, and the symbol z is a check symbol. Remember $x, y, z \in (Z_2 \oplus Z_2)$. We propose the following check equation $x + \varphi_5 y + \varphi_6 z = (0; 0)$, i.e., we set the following formula to find the element z:

$$z = \varphi_5 x + \varphi_6 y \tag{1}$$

Recall, statistical investigations of J. Verhoeff [19] and D.F. Beckley [2] have shown that the most frequent errors made by human operators during data transmission are single errors (i.e. errors in exactly one component), adjacent transpositions (in other words errors made by interchanging adjacent digits, i.e. errors of the form $ab \to ba$), and insertion or deletion errors. If all codewords are of equal length, insertion and deletion errors can be detected easily.

Twin error is an error of the form $(aa \to bb)$. In [13] it is proved the following

Theorem 2. *Any $(n-1)$-T-quasigroup code (Q, g) with check equation*

$$d(x_1^n) = \alpha_1 x_1 + \alpha_2 x_2 + \cdots + \alpha_n x_n = 0$$

detects:

- *any transposition error on the place $(i, i+k)$, $(i \in \overline{1, n-k}$, $k \in \overline{1, n-i}$, $i + k \leq n)$ if and only if the mapping $\alpha_i - \alpha_{i+k}$ is an automorphism of the group $(Q, +)$;*
- *any twin error on the place $(i, i+k)$, $(i \in \overline{1, n-k}$, $k \in \overline{1, n-i}$, $i + k \leq n)$ if and only if the mapping $\alpha_i + \alpha_{i+k}$ is an automorphism of the group $(Q, +)$.*

From Theorem 2 follows that the proposed code detects any transposition and twin error. The proposed code is quasigroup code, therefore it detects any single error [12], [13].

Suppose we have a word of the form ab, $a, b \in Z_2 \oplus Z_2$. There exist $3 \cdot 3 = 9$ double errors that can be done in this word. It is easy to see that given code detects 6 errors and it cannot detect 3 double errors.

Thus this code detects 12 from theoretically possible 15 errors in any word of the form ab, $a, b \in Z_2 \oplus Z_2$, i.e., it detects 80% errors in information symbols by supposition that the check symbol was transmitted without error.

Cryptographical part. We construct cryptographical part of the proposed cryptcode. For this aim we take three T-quasigroups over the group $Z_2 \oplus Z_2$:
$(Z_2 \oplus Z_2, D)$ with the form $D(x, y) = \varphi_3 x + \varphi_6 y + a_1$;
$(Z_2 \oplus Z_2, E)$ with the form $E(x, y) = \varphi_2 x + \varphi_5 y + a_2$;
$(Z_2 \oplus Z_2, F)$ with the form $F(x, y) = \varphi_3 x + \varphi_5 y + a_3$.

Lemma 1. *The quasigroups $(Z_2 \oplus Z_2, D)$, $(Z_2 \oplus Z_2, E)$, and $(Z_2 \oplus Z_2, F)$ are orthogonal in pairs.*

Proof. We can use Theorem 1 and Cayley table of the group $Aut(Z_2 \oplus Z_2)$.

Define three ternary operations:

$$K_1(D(x, y), z) = D(x, y) + z$$
$$K_2(E(x, y), z) = E(x, y) + z$$
$$K_3(F(x, y), z) = F(x, y) + z$$

It is clear that these operations can be replaced by a more complex system of operations.

Lemma 2. *The triple of ternary operations $K_1(x, y, z), K_2(x, y, z), K_3(x, y, z)$ forms an orthogonal system of operation.*

Proof. We solve the following system of equations

$$\begin{cases} \varphi_3 x + \varphi_6 y + a_1 + z = b_1 \\ \varphi_2 x + \varphi_5 y + a_2 + z = b_2 \\ \varphi_3 x + \varphi_5 y + a_3 + z = b_3 \end{cases} \tag{2}$$

where b_1, b_2, b_3 are fixed elements of the set $Z_2 \oplus Z_2$.

We use properties of the groups $(Z_2 \oplus Z_2)$ and $Aut(Z_2 \oplus Z_2)$:

$$\begin{cases} \varphi_3 x + \varphi_6 y + z = b_1 + a_1 \\ \varphi_2 x + \varphi_5 y + z = b_2 + a_2 \\ \varphi_3 x + \varphi_5 y + z = b_3 + a_3 \end{cases} \tag{3}$$

We do the following transformations of the system (3): (first row + third row) $\rightarrow$ first row; (second row + third row) $\rightarrow$ second row; and obtain the system:

$$\begin{cases} y = b_1 + a_1 + b_3 + a_3 \\ x = \varphi_4(b_2 + a_2 + b_3 + b_4) \\ \varphi_3 x + \varphi_5 y + z = b_3 + a_3 \end{cases} \tag{4}$$

In the third equation of the system (4) we replace x by $\varphi_4(b_2 + a_2 + b_3 + b_4)$ and y by $b_1 + a_1 + b_3 + a_3$, obtaining:

$$\begin{cases} x = \varphi_4(b_2 + a_2 + b_3 + a_3) \\ y = b_1 + a_1 + b_3 + a_3 \\ z = b_3 + a_3 + \varphi_5(b_1 + a_1 + b_2 + a_2) \end{cases} \tag{5}$$

Therefore, the system (2) has a unique solution for any fixed elements $b_1, b_2, b_3 \in (Z_2 \oplus Z_2)$, operations $K_1(x,y,z), K_2(x,y,z), K_3(x,y,z)$ are orthogonal.

Triplets of orthogonal operations $K_1(x,y,z), K_2(x,y,z), K_3(x,y,z)$ (by $a_1 = a_2 = a_3 = (0;0)$) define on the set Q^3 permutation with the following cycle type: $1^2 2^1 4^1 7^2 14^1 28^1$, i.e., this permutation contains two cycles of order 1, one cycle of order 2, and so on. Denote this permutation by the letter K.

The order of permutation K is equal to 28. Notice that using isotopy [3], [11] or generalized isotopy [14] it is possible to change the order of permutation K.

We shall use the system of three ternary orthogonal groupoids (Q, A), (Q, B), (Q, C) of order 4 from [5].

In these tables $A(0,1,2) = A_0(1,2) = 3$, $C(2,3,2) = C_2(3,2) = 2$, and so on.

A_0	0	1	2	3
0	0	1	2	3
1	1	2	3	0
2	2	3	0	1
3	3	0	1	2

A_1	0	1	2	3
0	1	0	3	2
1	0	1	2	3
2	3	2	1	0
3	2	3	0	1

A_2	0	1	2	3
0	2	3	0	1
1	3	0	1	2
2	0	1	2	3
3	1	2	3	0

A_3	0	1	2	3
0	3	2	1	0
1	2	3	0	1
2	1	0	3	2
3	0	1	2	3

B_0	0	1	2	3
0	3	0	1	3
1	0	2	3	0
2	1	2	1	3
3	1	1	2	2

B_1	0	1	2	3
0	2	1	1	0
1	2	3	3	0
2	0	2	1	3
3	0	0	3	1

B_2	0	1	2	3
0	1	2	0	0
1	2	0	3	1
2	0	2	3	2
3	3	2	1	1

B_3	0	1	2	3
0	3	3	2	2
1	0	1	2	1
2	0	2	0	3
3	3	1	0	3

C_0	0	1	2	3
0	3	1	2	0
1	2	1	1	2
2	0	1	0	1
3	3	1	2	3

C_1	0	1	2	3
0	1	2	1	3
1	1	2	3	1
2	0	2	2	0
3	1	3	1	1

C_2	0	1	2	3
0	3	3	0	0
1	2	1	0	1
2	3	3	2	0
3	3	0	2	3

C_3	0	1	2	3
0	2	1	0	0
1	2	0	2	3
2	3	3	2	0
3	2	0	0	3

Denote permutation that defines this system of three ternary orthogonal groupoids by the letter M, $M = M(A(x,y,z), B(x,y,z), C(x,y,z))$. This permutation has the following cycle type: $1^1 17^1 20^1 26^1$. The order of this permutation is equal to $17 \cdot 20 \cdot 13 = 4420$.

In order to use the system of orthogonal groupoids and the system of orthogonal T-quasigroups simultaneously we redefine the basic set of the T-quasigroups in the following (non-unique) way: $(0;0) \to 0$, $(1;0) \to 1$, $(0;1) \to 2$, $(1;1) \to 3$.

We propose the following cryptographical term (a cryptographical primitive):

$$H(x,y,z) = M^k(K^l(x,y,z)), k, l \in \mathbb{Z} \tag{6}$$

The transformation H is a permutation of the set Q^3. Indeed, this transformation is a composition of two permutations: K^l and M^k.

Remark 1. It is possible to use the following cryptographical procedure:

$$H_1(x,y,z) = K^t(M^k(K^l(x,y,z))), t, k, l \in \mathbb{Z},$$

and so on.

3 Algorithm

We propose the following

Algorithm 1 *1. Take a pair of information symbols $a, b \in (Z_2 \oplus Z_2)$;*
2. using formula (1) (or its analogue), find value of the check symbol c;
3. apply the cryptographical term H to the triple (a, b, c);
4. therefore, we obtain first three elements of the cipher-text;
5. take a pair of information symbols $d, e \in (Z_2 \oplus Z_2)$;
6. using formula (1), find value of the check symbol f;
7. change values of the numbers k, l in the cryptographical term H; also it is possible to change the term H by other term of such or other type;
8. apply the cryptographical term H to the triple (d, e, f);
9. we obtain next three elements of the cipher-text;
10. and so on.

Remark 2. At Step 7 of Algorithm 1 it is possible to use ideas of Feistel schema. Namely, it is possible to calculate the numbers k, l using some bijective functions, where the numbers of triplet $H(a, b, c)$ and previous values of k and l are used as arguments.

Decoding. Using permutations K^{-1} and M^{-1}, we can construct corresponding triplets of orthogonal 3-ary groupoids and so on.

Resistance relative to some possible attacks. Taking into consideration Remark 1, we can estimate the number of possible keys in the presented cryptcode. This number is equal to (64!). Length of any key is equal to $64 \cdot 3 \cdot 2 = 384$ bits.

At each step of the proposed algorithm only three symbols (six bits) are ciphered. Moreover, after any step this key can be changed. Therefore, brute-force attack is difficult.

Statistical attack also seems to be difficult. It is possible to present the following argument: the symmetric group S_{64} acts on the set, which consists from 64 triplets 64-transitively [7].

A code-crypt algorithm. Denote the coding procedure from Algorithm 1 as $C(x, y)$ since this procedure is a function of two variables. Therefore, we can describe procedures of coding and enciphering in Algorithm 1 by the following formula:

$$H(x, y, C(x, y)), \tag{7}$$

where H is taken from equation (6). It is possible to construct a code-crypt algorithm by the formula $C_1(H(x, y, z))$ since there exists a possibility to use an analogue of the code C for three information symbols [13, Example 19], i.e., we can transpose the procedures C and H.

Conclusion. Almost all constructions in this paper are performed over the field $GF(2^2)$. An analog of Algorithm 1 can be constructed over a field of the order more than four. Also we can use an alternating more powerful code [1].

Acknowledgment. Author thanks Referees for their helpful comments.

References

1. Bakeva V., Ilievska, N.: A Probabilistic Model of Error-Detecting Codes Based on Quasigroups. Quasigroups Related Systems 17(2), 135–148 (2009).
2. Beckley, D.F.: An Optimum Systems with Modulo 11. The Computer Bulletin 11, 213–215 (1967).
3. Belousov, V.D.: Foundations of the Theory of Quasigroups and Loops. Nauka, Moscow, (1967). (in Russian).
4. Blahut, Richard E.: Theory and Practice of Error Control Codes. Addison-Wesley Publishing Company, Advanced Book Program, Reading (1983).
5. Csorgo, Piroska, Shcherbacov, Victor: On Some Quasigroup Cryptographical Primitives, http://arxiv.org/abs/1110.6591 (2011).
6. Hall, Marshall: The Theory of Groups. The Macmillan Company, New York (1959).
7. Kargapolov, M.I., Merzlyakov, M.Yu.: Foundations of Group Theory. Nauka, Moscow (1977).
8. Kepka T., Němec, P.: T-quasigroups, II. Acta Univ. Carolin. Math. Phys. 12(2), 31–49 (1971).
9. Markovski, S., Gligoroski, D., Kocarev, Lj.: Totally Asynchronous Stream Ciphers + Redundancy = Cryptcoding. In Proceedings of the 2007 International Conference on Security and Management, SAM 2007, June 25-28, 2007, Las Vegas, USA, 446–451, (2007). http://www.informatik.uni-trier.de/ ley/db/conf/ csreaSAM/csreaSAM2007.html/.../GligoroskiMK07.
10. Markovski, S., Gligoroski, D., Kocarev, Lj.: Error Correcting Cryptcodes Based on Quasigroups. NATO ARW, October 6-9, 2008, Veliko Tarnovo, Bulgaria, (2008). https://www.cosic.esat.kuleuven.be/.../Markovski_slides_nato08.ppt.
11. Pflugfelder, H.O.: Quasigroups and Loops: Introduction. Heldermann Verlag, Berlin (1990).
12. Mullen, G.L., Shcherbacov, V.A.: Properties of Codes with One Check Symbol from a Quasigroup Point of View. Bul. Acad. Stiinte Repub. Mold., Mat., 2 (48), 71–86 (2002).
13. Mullen, G.L., Shcherbacov, V.A.: n-T-quasigroup Codes with One Check Symbol and Their Error Detection Capabilities. Comment. Math. Univ. Carolin. 45(2), 321–340 (2004).
14. Mullen, G.L., Shcherbacov, V.A.: On Orthogonality of Binary Operations and Squares. Bul. Acad. Stiinte Repub. Mold., Mat., (2 (48)), 3–42 (2005).
15. Němec, P., Kepka, T.: T-quasigroups, I. Acta Univ. Carolin. Math. Phys. 12(1), 39–49 (1971).
16. Shcherbacov, V.A.: Elements of Quasigroup Theory and Some Its Applications in Code Theory, (2003). urls: www.karlin.mff.cuni.cz/drapal/speccurs.pdf; http://de.wikipedia.org/wiki/Quasigruppe
17. Shcherbacov, V.A.: On Some Known Possible Applications of Quasigroups in Cryptology (2003). www.karlin.mff.cuni.cz/drapal/krypto.pdf
18. Shcherbacov, Victor: Quasigroup Based Crypto-Algorithms. arXiv:1110.6591v1 (2012). http://arxiv.org/pdf/1201.3016v1.

19. Verhoeff, J.: Error Detecting Decimal Codes, volume 29. Math. Centrum, Amsterdam (1969).

S. Markovski, M. Gusev (Editors): ICT Innovations 2012, Web Proceedings, ISSN 1857-7288

S. Markovski, M. Gusev (Editors): ICT Innovations 2012, Web Proceedings, ISSN 1857-7288

Review of Limitations on Namespace Distribution for Cloud Filesystems

Genti Daci[1], Frida Gjermeni[1]

Polytechnic University of Tirana, Faculty of Information Technology, Tirana, Albania
gdaci@abcom.al, frida_gjermeni@hotmail.com

Abstract. There are many challenges today for storing, processing and transferring intensive amounts of data in a distributed, large scale environment like cloud computing systems, where Apache Hadoop is a recent, well-known platform to provide such services. Such platforms use HDSF File System organized on two key components: the Hadoop Distributed File System (HDSF) for file storage and MapReduce, a distributed processing system for intensive cloud data applications. The main features of this structure are scalability, increased fault tolerance, efficiency and high-performance for the whole system. Hadoop supports today scientific applications, like high energy physics, astronomy genetics or even meteorology. Many organizations, including Yahoo! and Facebook have successfully implemented Hadoop as their internal distributed platform. Because HDSF architecture relies on a master/slave model, where a single name-node server is responsible for managing the namespace and all metadata operations in the file system. As a result this poses a limit on its growth and ability to scale due to the amount of RAM available on the single namespace server. A bottleneck resource is another problem that derives from this internal organization. By analyzing the limitations of HDSF architecture and resource distribution challenges, this paper reviews many solutions for addressing such limitations. This is done by discussing and comparing two file systems: Ceph Distributed File System and the Scalable Distributed File System. We will discuss and evaluate their main features, strengths and weaknesses with regards to increased scalability and performance. Also we will analyse and compare from a qualitative approach the algorithms implemented in these schemes such as CRUSH, RADOS algorithm for Ceph and RA, RM algorithms for SDFS.

Keywords: Algorithm, cloud filesystems, Hadoop distributed filesystem, performance, Ceph File System, Scalable Distributed File System

1 Introduction

Apache Hadoop is a famous, suitable platform for storing, processing and transferring large amounts of data sets in a cloud environment. The compact organization

S. Markovski, M. Gusev (Editors): ICT Innovations 2012, Web Proceedings, ISSN 1857-7288

realised by Hadoop Distributed File System [2] for storing data of different complexity and variety and MapReduce [3] for processing intensive distributed data applications, give HDSF a fundamental role in adapting the changing way of storing and processing information today. Projected to run on low cost hardware, HDSF file system is highly fault tolerant due to its replication policies for recovering data blocks in case of hardware failures or data corruption. Efficiency and high performance are provided adding to replication policies the ability to access data files in a stream, write-once-read-many [2] manner. One of the main features of HDSF is its ability to scale through hundreds of nodes in cloud computing systems. This is achieved by decoupling or separating metadata from data operations since metadata operations are usually faster than the data ones. MapReduce exploits the scalability of the file system [7] offering computation movement to interested data for a certain application.

Several important fields benefit from HDSF services such as astronomy, mathematics, genetics [9] or even economy used in online commerce and fraud detection [7], etc. Also important organizations like Yahoo!, Facebook and eBay have successfully implemented Hadoop as their internal distributed platform.

HDSF file system structure organization is based on the master/slave model where only a single NameNode server (master) menages the entire namespace, metadata operations and access to data files from clients of the file system. Data files are split into data blocks that are stored in DataNodes, usually one for every node in a cluster environment. Because of speed increasing, the whole file system metadata is stored in the NameNode single server's main memory. This implies a limit on the capacity of the system to grow and scale into a larger number of nodes than those supported by the available amount of RAM present in the NameNode server. Also the presence of this single server as the only metadata source creates a bottleneck resource and a single point-of-failure if it goes down. The same scenario can happen in the case where the number of contemporary requests for metadata operations that address the namespace server, reaches the limits of its capacities. Also the network bandwidth available for exchanging data between the actors in the file system is a limiting factor regarding scalability and overall performance.

The contributions of this paper are: 1) review of Ceph distributed file system and the new Scalable Distributed File System with regards to growth and scalability limits and 2) compare their features, strengths and weaknesses 3) analyse and compare from a qualitative approach the algorithms implemented in these schemes such as CRUSH, RADOS algorithm for Ceph and RA, RM algorithms for DSFS.

The rest of the paper is organized in the following order: section 3 will describe the limitations and solutions of the existing HDSF architecture, regarding scalability improvement. In section 4, features and comparison of Ceph file system and DSFS file system will be discussed and in section. Also algorithms implemented in Ceph such as CRUSH, RADOS and those implemented in DSFS such as RA, RM will be compared from a qualitative approach.

A lot of work has been made during the last years on the field of distributed networked file systems. Andrew File System (AFS) [15] is a distributed file system that presents several advantages over traditional networked file systems, especially in the fields of security and scalability. It uses a cluster of servers to provide a homogeneous

and location transparent namespace to all the clients who want to interact with it. No matter the fact that it can handle large amounts of file requests it may not be convenient for large scale and scientific operations like those handled by HDSF.

The Networked File System (NFS) [16] is another way to share files on a network. It has the disadvantage that it oversimplifies everything by making a remote file system appear as a local one. In situations like those handled in HDFS it wouldn't be appropriate or even reliable.

Other file systems used in distributed enviroments are Frangipiani [17] and InterMezzo [18]. Frangipiani is a distributed file system where disks on multiple machines are stored and managed in a single shared pool. Due to a simple internal structure system recovery and load balancing are easily managed. In InterMezzo, local file systems serve as server storage and client cashe and make the kernel file system driver a wrapper around the local file system. These file systems do not have a good resouce allocation regarding dynamic link, storage and proccesing capacities.

Another very well-known file system is the Google File System (GFS) [6] which resembles to HDSF and has recently announced a namespace distribution even if no information for the public hasn't been provided yet.

2 Related Work

A lot of work has been made during the last years on the field of distributed networked file systems. Andrew File System (AFS) [15] is a distributed file system that presents several advantages over traditional networked file systems, especially in the fields of security and scalability. It uses a cluster of servers to provide a homogeneous and location transparent namespace to all the clients who want to interact with it. No matter the fact that it can handle large amounts of file requests it may not be convenient for large scale and scientific operations like those handled by HDSF.

The Networked File System (NFS) [16] is another way to share files on a network. It has the disadvantage that it oversimplifies everything by making a remote file system appear as a local one. In situations like those handled in HDFS it wouldn't be appropriate or even reliable.

Other file systems used in distributed enviroments are Frangipiani [17] and InterMezzo [18]. Frangipiani is a distributed file system where disks on multiple machines are stored and managed in a single shared pool. Due to a simple internal structure system recovery and load balancing are easily managed. In InterMezzo, local file systems serve as server storage and client cashe and make the kernel file system driver a wrapper around the local file system. These file systems do not have a good resouce allocation regarding dynamic link, storage and proccesing capacities.

Another very well-known file system is the Google File System (GFS) [6] which resembles to HDSF and has recently announced a namespace distribution even if no information for the public hasn't been provided yet.

3 HDFS and Scalability

The two main actors in HDSF architecture execute different operations over the file system. Operations like open close and rename of the files and directories of the namespace are executed by the NameNode server. Whereas operations like read and write, that in fact are client request operations over the data As mentioned above there are serious limitations regarding scalability, caused by the single NameNode server and its amount of RAM for namespace storage. Furthermore, from works by Shvacko *at al*. [10], 30% of the processing capacity of the name-node capacity is dedicated to internal load, such as processing block reports and heartbeat [2]. The remaining 70% of capacity processing can easily be expired by 10.000 writers on a 10,000 node HDSF cluster as calculated by Shvacko *at al*. [10]. Consequently a bottleneck and name-node saturation would happen. Any intention to increase the number of DataNode servers, files stored in the DataNodes or clients requests beyond the supporting abilities of the single namespace server, causes a single point-of-failure in the file system.

When a client requests access to a file for a read or write operation, he first contacts the namespace server for the location of the data blocks forming the desired file on the datanodes. Than the transfer to or from the DataNodes takes place, as described in Fig.1.

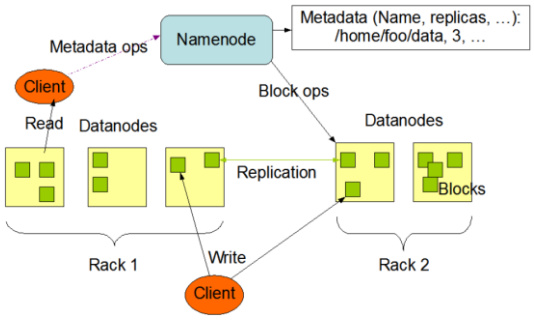


Fig. 1. HDSF architecture by [1]

Eventually the distribution of the namespace server and the decentralisation of the existing master/slave architecture represents an inevitable fact. That said, not all the approaches that attempt to introduce several namespace servers are successful. Such an example is Federation [8], where the number of clients and the amount of data stored increases, due to the presence of several NameNodes. But the static partitioning prevents from redistribution in the case when the growth of a volume reaches the NameNode capacities, so again we arrive at a lack of scalability.

No matter the system's necessity of distributing the namespace server into a cluster of NameNode servers, the costs of money and time (even years for building such a distributed environment) it takes, have to be considered. This means we have to think

about a distributed namespace rather than a distributed environment of Namenode servers. Hbase [1] is such an approach, but even here the algorithms that have to be implemented for the namespace partitioning and the issue of renaming represent hot topics of discussion.

4 Solutions Based on Distributed File Systems

In this section we analyse two file systems that address HDSF limtitations above discussed. Ceph Distributed File System and Scalable Distributed File System (SDSF) main features, strengths and weaknesses are specified regarding namespace distribution, scalability improvement, higher performance and increased fault tolerance.

4.1 Ceph [11]: How and Why

Ceph is an open-source platform [6]. How to address HDSF scalability limitations using Ceph? "Simply" by using Ceph as the file system for Hadoop platform as explained in [6].

Ceph is an object–based parallel file system, which internal organization [6] fulfills HDSF main requirements for playing the role of its file system.

One of the main strengths of Ceph is the distribution of the entire metadata operations and namespace file system into a cluster of metadata servers (MDS). The process of distribution is realised in a dynamic way thanks to the Dynamic Subtree Partitioning algorithm. Each MDS measures the popularity of metadata using counters. Any operation that affects the actual i-node increments the counters and all of it ancestors till the root. This way a weighted MDS regarding the recent local distribution, is provided. The achieved values are periodically compared and as e result appropriately sized subtrees are migrated, to keep the workload evenly distributed.

Besides workload balancing, also the ability to prevent hot spot points for metadata operation requests is achievable. Surely, the scalability and overall performance increase.

When a client wants to interact with the file system, first he requests the MDS as can be seen by Fig.2. As a response the MDS returns the i-node number, replication factor and the striping strategy of the requested file. Than the client can calculate on its own the placement of the data objects and replicas of the requested files.

OSDs (Object Storage Device) are the storage intelligent devices, that store both data and metadata, but due to RADOS (Reliable Autonomic Object Storage) [13] (also discussed in section IV) they are seen as a single logical object store by the clients and MDSs (as explined in Fig.2). RADOS represents a key component for Ceph because it provides every communication iniciatior with a cluster map containing information about OSDs, thier status and CRUSH (Controlled Replication under Scalable Hashing) [12]. CRUSH [12] resolves the problem of data distribution and location. It eliminates allocation tables by giving every party the ability to

calculate the data location.

No matter the strengths that RADOS and CRUSH offer, Ceph still remains in an experimental phase. Besides the advantages in scalability, it introduces a higher degree of complexity to the the file system, since it inserts objects, as storing entities. The abstraction that from it derives, it is not to be let out of attention.

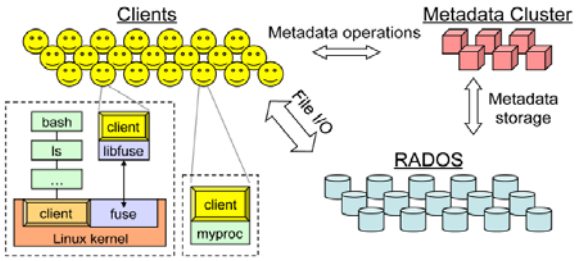


Fig. 2. Internal Architecture of Ceph by [11]

4.2 Scalable Distributed File System: Another Alternative

Scalable Distributed File System (SDSF) [4] is another distributed file system that attempts to resolve the limitations of HDSF platform. One of the main innovations of SDFS is the introduction of a front end light server (FES), as a connection point between the NameNode servers (NNS) and the client requests. The NNS handling the current request is chosen by hashing the request ID. To a certain point of view, this solution distributes metadata operations workload, but how many client requests can handle at once the FES [4]? Wouldn't this be another critical point for the file system overall performance ? No matter the fact that the presence of this front end server does not cause a resource bottleneck, since it is stateless, what about the load of a single NNS [4]? No information is provided about how many contemporary requests can a NNS handle. All the questions above raised represent fragile points that need further development or alternative solutions.

A strength point in SDFS is the introduction of a communication protocol that can achieve a better link utilisation during the operations with the file system. This is realised by the Resource Allocator (RA) and Resource Monitor (RM) algorithms, that will be further discussed in the next section. After a client connects to the FES, it is its rensponsibility to find the appropriate NNS to handle the client's request. As can be seen in Fig.3 the NNS finds the best block server (BS) after consulting the RA and the RM. The RA acts like a software router helped by the RM which gives the rmetric of every BS it is associated to. This scenario provides an increased scalability that derives from full link utilisation and rational resource allocation. By the end of this section we can conclude that namespace distribution is the key to scalability improvement regarding cloud file systems.

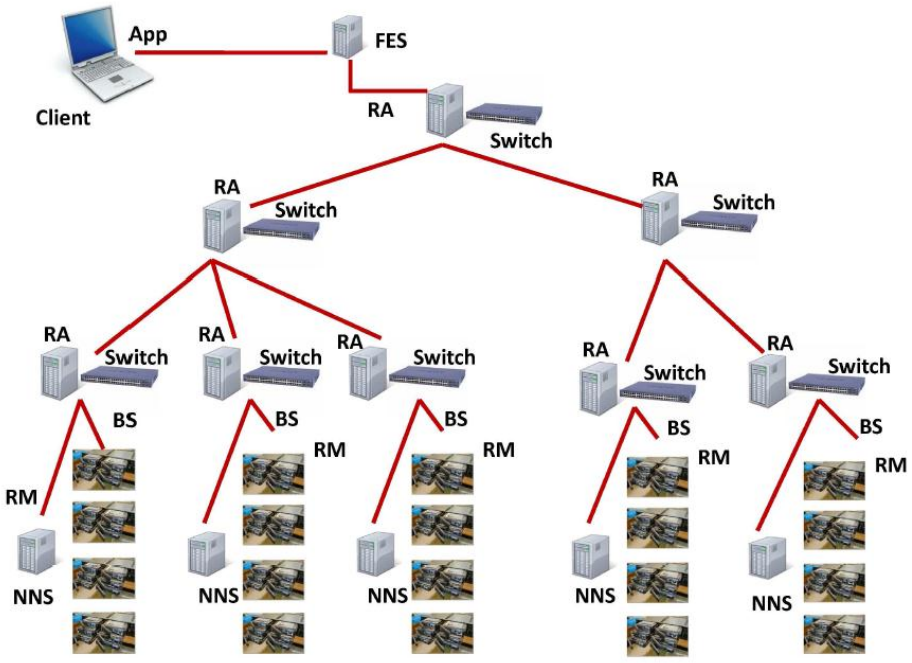


Fig. 3. Overview of SDFS by [4]

5 Comparing File Systems and Internal Algorithms

In this section we compare Ceph and SDSF file systems in their main features and their internal algorithms. While comparing we also try to give conclusions and further improvements for both file systems.

Decoupled data and metadata- Ceph and SDSF both offer separated data management from metadata. Since metadata operations are more complex and represent 30-80% of all file system operations, decoupling data from metadata operations gives the clients of the file system the possibility to a faster access on data manipulation. Therefore this derives to a performance improvement.

Distribution of Metadata Operation- Both file systems offer a distributed cluster of servers for namespace management, providing this way an increase in scalability. SDFS applies the hashing way [14] of spreading metadata workloads through the namenode servers (NNS). To some extend workload is evenly distributed for well behaved hash functions, but still hot spots can be created consisting of individual files located on a single NNS. Besides, the hashing algorithm destroys hierarchical locality and benefits that from it derive. Ceph offers a much better alternative using dynamic subtree partitionig for workload distribution in the MDS Cluster. As above explained this algorithm answers in a more effective way the dynamic needs of such cloud environments and a better chance to scale.

Data storage- Since Ceph is an object-based file system, the key logical unit of the whole file system are objects. CRUSH [6] is the algorithm that distributes data objects in a pseudo-random way through the OSDs. It allows any party to individually calculate the location of data and subtitutes this way, the process of lookup with that of calculation. Also this avoids all the overhead crewated during the lookup of the allocation tables. The abstraction of allocation tables is still present in SDSF where data is stored in the BS as data blocks directed by the NNS. In general NNS use the policies of k-local allocation or even global allocation for the storage of new data blocks. Somehow these algorithms offer a fair distribution but are not able to respond to the contiguous evolution of cloud computing systems.

Client requests- In SDSF client requests are handled at the front end light server that directs them to an appropriate NNS. No matter the fact that this server is stateless and in case of failures bringinig it up is very fast, still it becomes a critical point and compromises performance. The number of requests it can contemporary handle is a finite one. This can be a limiting factor for scalability that in fact we want to increase. On the other side Ceph offers an alternative solution provided by RADOS. RADOS enables every client with a global cluster map to direct requests. It also realises, OSDs to be seen as a single logical store. This way a critical point like the presence of the FES is avoided.

A basical advantage of SDSF over Ceph is taking into consideration full link utilisation for every operation in the file system. The RA and RM algorithms provide that, by periodically monitoring and fairly allocating resources on every request. This provides a higher data rate of exchanging information that brings to an increased performance of the whole file system.

We have to underline the fact that Ceph Distributed File system provides a higher abstraction for the system, but also introduces a more “intelligent” way to store and distribute data than SDFS. Ceph is an object-based file system, while the Scalable Distributed File System still conserves the stored data on blocks, stored in the BS[4]. This results in a higher performance and increased ability to scale.

At the end of this section we can conclude that both couples of algorithms (CRUSH, RADOS in Ceph and RA, RM in SDSF) are implemented in distributed file systems that are highly discussed today to become the file systems serving Hadoop, for scalability and performance increase. Even though there are main differences in the way they behave and the logical data partitions they operate on, a paralel comparision can be realised addressing the issues of scalability and performance.

6 Conclusions and Future Work

This paper reviewed solutions for the issue of scalability and performance increasing, regarding a well-known cloud filesystem like Hadoop Distributed File System. Because of having a single name node server as the source of all metadata operations, all stored in a finite size of RAM, scalability is seriously compromised. Two file systems were analysed and compared in their main fatures and internal algorithms to address HDSF problems: 1) Ceph Distributed File System and 2) Scalable Distributed

File System (SDFS) and their algorithms 1)CRUSH, RADOS for Ceph and 2)RA, RM for SDFS. It is for sure that to provide scalability improvement, the namespace has to be distributed in a cluster of servers where fair policies of metadata workload have to implemented. Ceph seems to answer in a better way these needs with the dynamic subtree partitioning algorithm for handling namespace distribution dynamically. Also CRUSH and RADOS algorithms handle data storage, client requests and file system opearations in a way that fulfills the contiguous evolution of cloud enviroments. Anyway the complexity is high and further developement is needed for providing full compatibily with HDSF platform.

On the other side SDSF preserves a more traditional structure with data stored in blocks and allocation tables used to retrieve them. This brings heavier workloads in the file system and therefore a lower performance. An advantage of SDSF over Ceph is this new protocol realised by the RA and RM algorithms that utilises nearly full link and results in faster operations in the file system. Implementing this new approach in Ceph should be point of further research and devlopement. A weak point in SDSF is the FES server because the amount of contemporary requests it can handle. Alternative solutions have to be provided to handle this scenario.

References

1. Apache, *The Apache HBase Book*, October 2010: http//hbase.apache.org/ book.html.
2. D. Borthakur "The Hadoop Distributed File System: Architecture and Design*"*, *The Apache Software Foundation*, 2007.
3. J. Dean and S. Ghemawat, "MapReduce: Simplified Data Processing on Large Clusters". *In proceedings of the 6th Symposium on Operating systems Design and Implementation* (OSDI 2004), pp.137-150. USENIX Association 2004.
4. D.Fesehaye, R.Malik, K.Nahrstedt "A Scalable Distributed File System for Cloud Computing",Technical Report, March 2010.
5. M.K. McKusick and S. Quinlan, "GFS: Evolution on Fast-forward," *ACM Queue*, vol. 7, no. 7, ACM, New York, NY. August 2009.
6. C. Maltzahn, E. Molina-estolano, A. Khurana, A.J. Nelson, S. Brandt, S. Weil "Ceph as a scalable alternative to the Hadoop Distributed File System", *;login*: vol. 35, no. 4, pp.38-49, August 2010.
7. M. Olson "HADOOP: Scalable,Flexible Data Storage and Analysis", *IQT QUARTERLY/Connecting Innovation and Intelligence*, vol.1, no. 3, pp.14-18, May 2010.
8. S.Radia, S.Srinivas, "Scaling HDFS Cluster Using Namenode Federation," HDFS-1052, August 2010:https://issues.apache.org/jira/secure/attachment 12453067/high-level-design.pdf.
9. K. V. Shvachko, "Apache Hadoop The Scalability Update*", ;login*: vol.36, no.3, pp.7-13, June 2011.
10. K.V.Shvachko, "HDFS Scalability: The Limits to Growth,", *;login:*, vol.35, no.2, pp.6–16, April 2010.
11. Sage A. Weil, Scott A. Brandt, Ethan L. Miller, Darrell D.E. Long, and Carlos Maltzahn, "Ceph: A Scalable, High-Performance Distributed File System," *Proccedings of the 7th Symposium on Operating Systems Design and Implementation (OSDI)*, Seattle, WA, November 2006, pp.307-320.

12. Sage A. Weil, Scott A. Brandt, Ethan L. Miller, and Carlos Maltzahn. "CRUSH: Controlled, Scalable, Decentralized Placement of Replicated Data," *Proc. of the 2006 ACM/IEEE Conf. on Supercomputing (SC '06)*, Tampa, FL, November 2006.
13. Sage A. Weil, Andrew Leung, Scott A. Brandt, and Carlos Maltzahn. "Rados: A Fast, Scalable, and Reliable Storage Service for Petabyte-Scale Storage Clusters," *Proceedings of the 2007 ACM Petascale Data Storage Workshop (PDSW '07)*, Reno, NV, November 2007.
14. Sage A. Weil, Kristal T. Pollack, Scott A. Brandt, and Ethan L. Miller, "Dynamic Metadata Management for Petabyte-Scale File Systems" *Proc. of the 2004 ACM/IEEE Conference on Supercomputing (SC '04)*, Pittsburgh, PA, November 2004.
15. Howard, J., Kazar, M., Menees, S., Nichols, D., Satyanarayanan,M., Sidebotham, R., and West, "M. Scale and performance in a distributed file system". ACM Transactions on Computer Systems (TOCS) 6,1 (1988), 51–81.
16. Shepler, S., Callaghan, B., Robinson, D., Thurlow, R., Beame, C.,Eisler, M., and Noveck, D. Network file system (NFS) version 4 protocol. Request for Comments 3530 (2003).
17. Thekkath, C., Mann, T., and Lee, E. "Frangipani: A scalable distributed file system", ACM SIGOPS Operating System Review 31, 5 (1997), 224–237.
18. Braam, P., Callahan, M., and Schwan, P. "The intermezzo file system. In Proceedings of the 3rd of the Perl Confer-ence,O¨ı£¡Reilly Open Source Convention, Citeseer.

Designing Backend Servers for Mobile Applications in the Industrial Project Management

Jovan Kostovski

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

Ilina Kareva

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

Abstract. This paper describes the design challenges which everyone faces when a backend server for mobile applications is developed. For this particular research a backend server for mobile context-aware services for industrial process management was developed. The context information is gathered from the employees' smart phones and the industrial process management legacy information systems. The goal of the research was to utilize the gathered context information in order to improve the management and monitoring of the industrial processes, make the operations teams more efficient and taking the employee safety to a higher level. As a result of the research a prototype system was build and the defined use cases were tested. The build system is placed on top of the existing process control and management systems, acts like a gateway between these systems and the employees' smart phones and enables complete up to date information about the process state and monitoring of the vital parameters of the employees when they are working on some risky tasks. Based on the tests done, it was concluded that this type of design ensures: prompt information sending to it is users, security, scalability, reliability and enables easy addition of extra system components for gathering process data and content adaptation for mobile devices.

Keywords: backend server, industrial process management, operations and maintenance enterprise mobility, mobile phone application, context-aware services.

adfa, p. 1, 2011.

S. Markovski, M. Gusev (Editors): ICT Innovations 2012, Web Proceedings, ISSN 1857-7288

1 Introduction

To enable continuous production, the production companies must ensure that all of the systems and machines which take part in the production process work properly. Machines' maintenance is very important because production outage means lower incomes or lost of clients due to unmet delivery deadlines. On the other hand if a malfunction or machine defect appears, the machine reparation may cost a lot or the restarting of the production line takes too long.

No matter how careful we are and what kind of precaution measurements we take the defects and production outages appear. If it's not a matter of a human error, the outages may appear because of machine's overloading or the machine parts simply abide due to machine's usage. Because the production outages can't be avoided the companies have various procedures and business processes which tend to make the outages less frequent and define employee's responsibilities in such situations. In case of production outage the most important thing is that every concerned employee, no matter what position he/she holds in the company's hierarchy has to be properly and timely informed about the situation. To ensure fast and quality intervention by the maintenance personnel, the employees who will be sent must be those who are physically closest to the place where the problem happened and they should have the relevant knowledge and experience in solving such problems.

To enable prompt and correct information delivery to all concerned parties in case of an emergency and production outage there must be a system which will monitor the current state of the whole production process and which will "know" the current state of each machine which is used in it. The role of such a system is to timely inform all concerned employees and to send only the information which is relevant to them.

With the development of the technology, the mobile devices and the smartphones are widely spread and they are becoming more and more powerful tools for accessing any kind of information in any format: audio, video, image or text. Because of that, they are perfect platform for displaying different kind of information. People are in constant movement and search for information, so it's very important to have the information available anywhere, anytime. Because of this and the various hardware accessories like: camera, compass, sensors for orientation, temperature, location etc. the smartphones enable us to detect the user's state so that we can send the most appropriate information in that particular moment in time. The other useful things which can help us to determine the state of the user are the way the user interacts with the smartphone and their habits when they are using the phone.

By using smarthones as a platform to which the relevant production process information is sent, the employees can have the exact state of the whole production process anywhere, anytime, so that they can timely react in that particular situation.

The goal of the research described in this paper was to make an overview of the problems which happen in the industrial process management, specifically the maintenance of the production lines and to give some pointers how these processes can be improved with using the modern mobile technologies, mobile devices and the services they offer. The idea was with minimal changes to the existing IT infrastructure to

S. Markovski, M. Gusev (Editors): ICT Innovations 2012, Web Proceedings, ISSN 1857-7288

enable access to the production process and production line data, organize it in a meaningful way, and make it accessible to the employees' mobile devices anywhere, in order get the process maintenance on the higher level.

2 Related Work

2.1 General division of the systems

There are different systems available on the market which partly cover the area discussed in this paper, but there are only a few which make their data available on mobile devices.
These systems can be divided in three general categories:

1. ERP [1] (Enterprise Resource Planning) and BPM (Business Process Modeling) systems. They are used by the management teams to organize the activities and the business processes in the companies;
2. SCADA [2] (Supervisory Control and Data Acquisition) systems. They are used by the process engineers, process operators and the maintenance personnel to monitor the state of the production processes;
3. Various custom designed systems used by the maintenance personnel used in their daily tasks.

The ERP systems are usually used by the managers, but some of them have modules which are used by the warehouse or the maintenance personnel. These modules are usually used for tracking the state of machine parts, whole machines or tools owned by the company as well as for tracking the taken actions on the production line. Example of such module is the Plant Maintenance [3] (PM) module of SAP [4] and OpenERP [5]. These systems are great for tracking the company's processes, but in the field of plant maintenance they don't offer much except tracking the actions taken on the production line and parts replaced. Usually they lack the part which abstracts the current process state and presents data which are relevant to a specific user. Regarding the data access from mobile devices they offer limited or no access, but some of them, for an extra charge, offer APIs so that some third party companies or the users themselves can develop mobile clients.

The SCADA systems are the basic information systems in every control and monitoring of industrial processes. They visually show the process state, the values of the process parameters and alarm the operators in case of some abnormal state. The process state is acquired by reading various sensors or local controlling loops made with PLC (Programmable Logic Controllers). The alarms are usually made with sound (blowing a horn) but there are some systems which can send SMS or email messages. Regarding the mobile data access the things written for ERP systems are also valid. Many times these systems are built to specifically satisfy the needs of a

particular industrial process and they use a specific proprietary hardware which makes upgrades and data retrieval hard.

The third group of systems comprises the systems which help the maintenance personnel. These systems improve the communication between the maintenance personnel, give better overview of the parameters of interest, give overview or the environment in which the work is done or even follow the vital functions of the employees. These systems are usually closely coupled with the monitored industrial process and have different implementations. The common thing about all of them is that they use some specialized hardware which the users are wearing on them, so called wearable computers [6][7][8].

2.2 Similar systems

The systems similar to the one discussed in this paper, usually use location based services, wearable computers and augmented reality.

Google made an interesting system called Google Maps Coordinate [9]. The system is build on top of few Google services like Google Maps and Google Latitude and it's intended to be used for coordination of the field work done by the maintenance teams. The team coordination is done on a central place and the field worker positions and the jobs done are shown on a map in real time. The maintenance personnel can add write messages and update the job status reassign the jobs to some other worker etc. The system is quite new and is build on top of proven Google technologies. The difference between this system and the one discussed in this paper is that the data and the jobs are inserted manually and they are not taken out automatically from some production process.

The other systems are made by some industry giants like SAP and Siemens. SAP has more general ERP solutions which are customized to the needs of a specific implementation. On the other hand Siemens has some solutions which are completely closed and made from their own hardware and software. This is good in terms of system integration, everything is from single vendor and can be easily integrated, but in case some hardware needs to be bought from some other vendor, there can be problems integrating it or maybe some extra charges for customization.

The rest of the systems are developed mostly by some research institutions and are targeting electrical and nuclear plants and waste water facilities [10][11][12][13]. They use location based services for outdoor localizing and QR Codes or barcodes for indoor localizing or object detection. They also use augmented reality to show the users how to get to the place where the problem is. To read some process parameter when the users are standing in front of some machine augmented reality, barcodes and image processing on the mobile device are used. These systems are good, but they are made specifically for some particular plant and a particular process.

3 System Overview

3.1 General overview of backend servers for mobile applications

The general block diagram of a backend server for mobile applications is shown on figure 1. It can be seen that the content which has to be delivered to the mobile devices is retrieved and processed so that it can be manipulated by the rest of the system. After this the contents are adapted so that they are suitable for displaying and delivery to the mobile devices. At the end the content is being delivered to the mobile devices.

Fig. 1. Backend server for mobile applications block diagram

The content retrieval and the initial data processing depends on the content sources and the form and amount of data which they are serving. Depending on the content sources the connectors for retrieval can use different mechanisms for data retrieval like content polling, some notification messages when there is data that the system is interested in etc. The retrieved data can be in various formats as well.

The content adaptation is done depending on the number of different devices to which the data should be delivered and the format in which the data should be delivered. If we have some devices which are connecting via web service we would probably use JSON or SOAP messages and if we are delivering to a mobile phone browser we'll have to adjust the HTML layouts and image sizes so that they will be appropriate for a given device.

The content delivery is made based on the way the content should be delivered. This depends on the way the communication between the backend server and the mobile application goes.

The crucial thing that must be done is to move all the heavy processing on the server side so that the client mobile applications will act just like a thin terminal. The other thing that must be done is to optimize the power consumption on mobile device because the battery life is limited. The whole communication and the mobile device sensor readings have to be carefully designed so that the mobile device can handle all the data sent from the server without losing any of its parts. We should also be careful with the amount of data that the server sends to the mobile phones so that we do not affect the data plan and cause extra expenses for the users. If the mobile application uses the sensors, than we must optimize how we use them so that our app will not kill the battery in a short time interval after we started it.

Speaking in general, the design of the backend server depends from the application in which it is used but we should always have in mind that the mobile devices have limited hardware resources and battery life, we must optimize and secure the data transfer and we should adapt the content so that it suits the best to the mobile device.

3.2 System block diagram and system requirements

Before we set the system requirements for our backend server for mobile applications for the industrial project management we will give a short overview of the context-aware data which can be acquired from the process and from the employee's smartphones. Context is a set of information which describes some object or a person in some particular situation.

We can divide the context information gathered from the process in three general groups:

- The place where the information is gathered;
- The employee profile which need information of a specific type;
- The meaning which the information has to the set production process goals and how it affects them.

The information can be gathered in various places: the current state of the production process, the state of the machines on the production line, the history of failures, etc... This is the information in its raw state and it is used by the process engineers, process operators and the maintenance personnel.

The second group contains the same information as the first one, but it is filtered and grouped so that the each employee gets only the information which is relevant to him/her.

The third group gives some interesting information based on abstraction. The process parameters are analyzed and based on their values or correlation to other parameters a conclusion is made and presented to a particular group of employees. For example if there are few parameters whose values observed together means that some machine is broken, the managers won't get the parameter values, but will get a message saying that the machine is broken.

The employee context information can be the employee's job position, current tasks, latest completed tasks and sensor readings from his /hers mobile device (location, orientation, environmental temperature, vital functions – pulse, blood pressure...). All these information can help us to determine which data is relevant for the employee in a specific situation.

Based on the available context information and goals which are set for the backend server we can set the system requirements:

- Reading of the process parameters in real time from various sources;
- History of the process parameter values fluctuation;
- Statistical analysis of the process parameter values and prediction of possible defects;
- Prompt delivery of the relevant information about the current process state and predicted failures to all concerned parties;
- Support for various mobile devices;
- Remote view and help for the maintenance personnel in the field;

- Tracking of the employee's vital parameters if they are working in some bad conditions;
- Tracking employee's performance on the reparation task and proposing the most appropriate person for a specific problem.

According to the written above, we have designed the system whose block diagram is shown on figure 2.

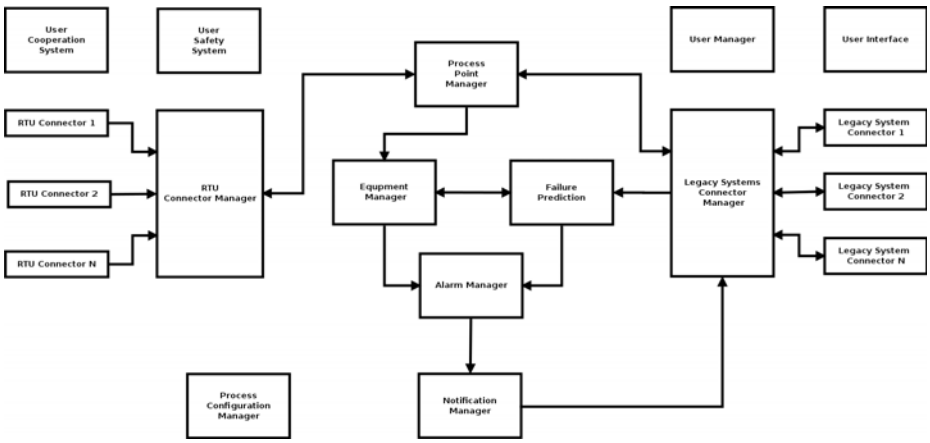


Fig. 2. Block diagram of the designed system

On the block diagram few major groups can be seen:

- Remote terminal unit manager and connectors – these are used to connect to sensors and data sources apart from the legacy systems;
- Legacy systems manager and connectors – used to connect to the existing systems ;
- Process point and notification manager – management of process point parameters and their values as well as definition and sending of notifications;
- User management – management of user rights, grouping the users by their position in the company and data relevant to their working tasks;
- Employee communication system and vital parameters monitoring – a sub-system which enables communication between the maintenance personnel while they are working and monitoring of employee's vital parameters.

The designed system has modular component design which enables easy addition or replacement of various components. It is build on top of the existing IT infrastructure and process control circuits. This way without any problems it can be implemented for various process control applications.

4 Implementation Details

The system is designed to be implemented with standard well proven software technologies. Because the process control IT infrastructure can have many different legacy control systems produced by various manufacturers, we have decided to use messaging bus and interfaces to connect the control systems to it.

The message bus is implemented with 0MQ [14] a high-performance asynchronous messaging library aimed at use in scalable distributed or concurrent applications. This library was chosen because it Open Source, it works under the most of the major operating systems, it supports various types of messages and network topologies and supports many programming languages. The capability of developing messaging clients in various programming languages enables the developers and system integrators to write the interfaces in the language the legacy system is build. The other basic system components such as the process point, notification, service discovery, message loggers are written in Java with the Spring [15] framework. The mobile application used in the testing was written for Android [16]. The component diagram can be seen on figure 3. This kind of system organization enables scalability, reliability and easy system component replacement in case of problems.

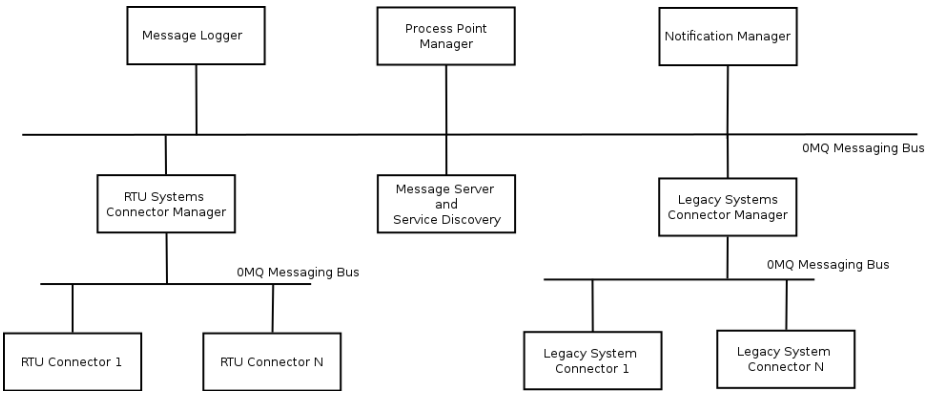


Fig. 3. System component diagram

On the system deployment diagram shown on figure 4 it is shown where and how the system is connected in the company's control systems IT infrastructure. It can easily connect to various systems on various levels of the IT network such as plant floor, operations network and business network.

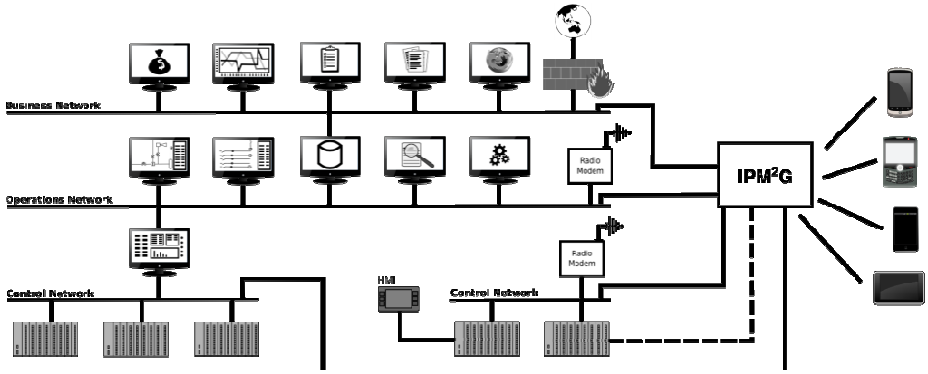


Fig. 4. System deployment diagram

5 Testing and discussion

Because the main goal of the research was system design and not its implementation, instead of building a complete system and testing its capabilities only a few components of the whole system were made. The main focus of the tests was to ensure that the components are communicating between themselves as it was designed and to check the processing of the process parameters and their delivery to the mobile devices.

The communication tests were made by examination of the messages passed in the system between its components. Basically the messages were monitored in a normal operation mode, their content was analyzed and was checked against the predefined use cases and communication rules of the defined communication protocol.

The processing of the process parameters values was tested in order to see if the messages are correctly interpreted on high level (to check the predefined rules by which few parameters together define some process state and the message is abstracted) and to see if the process parameter values are properly filtered and only the relevant information is sent to a specific employee.

The last conducted tests were system load tests which were performed by simulating few legacy system connections and few RTU connections sending data and the observation was made if the data is properly and timely processed.

All of the tests passed without any significant problems.

6 Conclusion

By the analysis of the system design and conducted test of the prototype system it can be concluded that the primary goals have been totally accomplished. If this kind of system needs to be implemented for a real industrial process, the whole system must be adjusted to that specific process. The build system can work out of the box for some small process, but for bigger processes there might have to be done some adjustments in order the system to handle all that big amount of data. Maybe just adding some extra system components to share the processing time will be enough. Maybe some adjustments of the data processing algorithms should be done. It all depends on the industrial process for which the system will be used. Besides the data processing adjustments, some changes in the content delivery components and the content adaptation components should be made in order to meet the requirements of the used mobile devices.

Generally speaking this system design is a good base for building real world back-end servers for mobile applications for industrial process.

7 References

1. Enterprise Resource Planning https://en.wikipedia.org/wiki/Enterprise_resource_planning
2. SCADA Systems https://en.wikipedia.org/wiki/SCADA
3. SAP PM http://www.erptips.com/Learn-SAP/SAP-Module-Overviews/Plant-Maintenance-PM.html
4. SAP homepage - http://www.sap.com
5. OpenERP homepage http://www.openerp.com
6. Hendrik Witt, „User Interfaces for Wearable Computers“, 2007
7. Barfield Woodrow, Caudell Thomas, „Fundamentals of Wearable Computers and Augumented Reality“, Lawrence Erlbaum Associates, Inc. , 2000
8. Asim Smailagic and Daniel Siewiorek, „Application Design for Wearable and Context-Aware Computers“, IEEE PERVASIVE computing, 2002
9. Google Maps Coordinate http://www.google.com/enterprise/mapsearth/products/coordinate.html
10. Hidekazu Yoshikawa, Distributed HMI System for Managing All Span of Plant Control and Maintenance, Nuclear Engineering and Technology Vol.41 No.3 April 2009
11. Gudrun Klinker, Oliver Creighton, Allen H. Dutoit, Rafael Kobylinski, Christoph Vilsmeier, Bernd Brügge, „Augmented maintenance of powerplants: A prototyping case study of a mobile AR system“, 2001
12. Hiroshi Shimoda, Hirotake Ishii, Yuichiro Yamazaki, Wei Wu, Hidekazu Yoshikawa, „A Support System for Water System Isolation Task in NPP by Using Augmented Reality and RFID“, 2004
13. Pekka Siltanen, Tommi Karhela, Charles Woodward, Paula Savioja, Augmented Reality for Plant Lifecycle Management
14. 0MQ homepage - http://zeromq.org/
15. Spring Framework homepage - http://www.springsource.org/
16. Android homepage - http://www.android.com/

Using hidden space in optimization of space utilization

Riste Marevski[1], Ivan Chorbev[1], Viktor Todorovski[1]
[1] Faculty of computer science and engineering, University of Ss Cyril and Methodius,
"Rugjer Boshkovikj" 16, P.O. Box 393, 1000 Skopje,R. of Macedonia
riste.marevski@yahoo.com, ivan.chorbev@finki.ukim.mk, viktor.todorovski@yahoo.com

Abstract. The topic of this paper is the use of advanced algorithms in order to solve the problem of optimal use of available space. There are a lot of algorithms that try to solve this problem but most of them are not taking into consideration the available space into the concave elements. In this paper we describe how to use this space in order to find the optimal solution. Most of the algorithms that solve this problem use genetic algorithms as a base for the optimization. Some of them also use heuristics in order to implement expert knowledge. Our approach is based on an algorithm that groups the elements utilizing the available space from concave elements and then continues with the optimization phase. The optimization phase is implemented as a genetic algorithm that uses specific problem heuristics.

Keywords: space utilization, transportation optimization, combinatorial optimization, packaging problem, cargo loading optimization, genetic algorithms, heuristics

S. Markovski, M. Gusev (Editors): ICT Innovations 2012, Web Proceedings, ISSN 1857-7288

1. Introduction

Optimal use of space is a generic problem area that includes various problem subtypes that are being solved with varying efficiency and quality. A lot of researches try to solve problems of this type and there are a lot of proposed algorithms that ought to solve them. But the proposed algorithms are very often limited to simple geometric forms, usually rectangular or in some cases cylindrical forms [8]. Since these algorithms offer an optimal placement of elements with the mentioned limitations, our aim was to build an algorithm that can be used to place elements with complex shapes. Also, we aimed at utilizing the space inside the elements e.g. the holes in the elements is a space that is not taken into consideration. We refer to this space as "hidden" space since is not "visible" for the algorithms (not taken into consideration). Our approach is mainly focused on using this "hidden" space in order to produce an optimal solution.

We propose a two phase algorithm where in the first phase the elements are grouped, while the second phase is a somewhat standard loading optimization phase. The grouping phase of the solution that we propose can be incorporated in any optimization algorithm as an enhancement. When this phase is over, the algorithm can continue with an optimization phase. The second phase can implement any optimization algorithm, not limited to genetic algorithms, b-tree search etc. In this paper we present an application of an adapted genetic algorithm.

Researchers have looked for a solution to this problem for a long time in the past. Apparently, back in the sixties of the last century [4] the need for optimal utilization of space emerged as a topic of more serious research. Since then this problem is quite researched and thus a number of techniques for its solution are proposed. However, optimizing the utilization of space has remained a popular research topic even today. Proofs of the continued interest are the papers that are published in the last few years focusing on this subject [1,4,6,24]. The fact that researchers are still working on finding more appropriate and more complete solution to these problems suggests that this is an area where there is room for further improvement.

Optimization of space utilization finds application in areas such as transport of material goods, warehouse storage, packaging, etc. [5].

The rest of the paper is organized as follows: section 2 describes the previous research on this topic, the limitations of the proposed solutions and the areas that can be improved. Section 3 describes our algorithm in detail. Section 4 describes the experimental results, while in section 5 the conclusion of this research is stated. In section 6 we present the future work aims and the planned improvements.

2. Previous research

Several ways of solving the problem of optimal utilization of space can be found in the literature. These solutions have reached the maximum utilization and offer an optimal solution for specific cases only. However, most of them are characterized by

certain limitations that make these algorithms applicable to only a small number of real world situations. Most common assumption for the algorithms that solve the problem of optimal placement of elements in a given finite space is that all elements must be in a form of a box [2,3,5,14,16,17,18]. For this specific case there are a lot of algorithms that can find the optimal solution. But in the real world elements often are in a form different from the form of a box. In this case the optimum utilization of space is more complicated and should be done in a different way. The actual research in this area offers some examples of solving the problem of optimal utilization of space when items should be placed in a form different than the shape of a box. In most of the cases the elements are in the shape of a cylinder [8].

The main characteristic of most of the applications that solve the problem of optimal placement of the elements is to satisfy two main conditions. The first condition to be satisfied is the positioning of the elements in a way that does not exceed the space available for loading. The second condition is that the elements must not overlap each other. The applications that try to find the optimal solution are mainly based on these two conditions. However, a much larger number of factors affect the optimal placement of material goods such as restrictions on the rotating elements, stability of the packed elements, the complexity of the arrangement of the elements, limiting the maximum weight and fragility of the elements [1,4,21]. Some specific research and commercial applications for this purpose offer solutions that partially implement a fraction of these factors during decision making. Another feature that most studies do not take into consideration is the use of extra space that comes from elements that contain holes [6.9], which certainly leaves room for further research in this area. The main purpose of this paper is to make an improvement in this area. The purpose of our research is to implement a solution that will use this unused space and thus make a further step forward and offer a more complete solution to this real word problem.

3. Loading Algorithm

The loading algorithm consists of two phases. The first phase is referred to as a Grouping phase. In this phase we group the elements aiming to utilize the available hidden space (space that arises from holes in the elements) in an optimal way. In the second phase we place the already grouped elements in the available space of the storage using genetic algorithm in order to find the optimal arrangement of elements.

3.1 Grouping phase

In the initial phase the elements are grouped placing one or more elements in the empty space within another element. This way the set of elements for storage is reduced enabling more elements to be placed in the available space. The implementation of the grouping algorithm works as follows:

The algorithm maintains a list of elements and a list of available spaces that arise from the elements. Starting from the element with maximum volume we loop through the elements and try to find an available space within one of the elements using the

Best Fit approach. This approach makes sure that elements will be compressed as much as possible.

Within the list of elements that are maintained, each element is described with a set of values that enable simple implementation of the Best Fit approach. The set of values includes the volume of the element, the shape of the element, the size of the element, list of void spaces inside the element as well as values that are used to describe the real position of the element in the space (orientation of the element and actual position). Shape representation is implemented using regular geometric forms. Each element can be described with a set of simple geometric forms. Simple elements are described as a box, a cylinder or a sphere. More complex elements are described by a composition of simple geometric forms. The available space which needs to be filled with elements as well as void space in the elements is also represented using the same notation. The use of simple geometric forms enables finding the space that best fits with simple checks. This reduces the complexity of the algorithm itself as well as its execution time.

For each element the algorithm tries to find the space that fits best. For example, if the element has a shape of a box and also the space is in a shape of a box with the same dimensions the algorithm will place the element in that space utilizing 100% of the space. If this is not possible, the algorithm searches for the element-space combination that has the smallest difference between the volume of the space and the volume of the element. The space that is left unfilled is represented as a composition of simple geometric forms. In many cases, when one element is placed inside another, the space that is left empty can be represented as different compositions of simple geometric forms (Figure 1). For example, if we place a smaller box into another box the empty space can be represented with composition of three other boxes in six different ways. Three of them are shown on Figure 1. When this is the case, our algorithm adds all eighteen boxes to the list of available space. This is done in order to support all different elements that can be placed in this space. These spaces are referenced to each other and when one of them is used for placing of another element, the spaces from five other combinations are removed from the list. The available space that arises from the selected space is divided using the same logic.

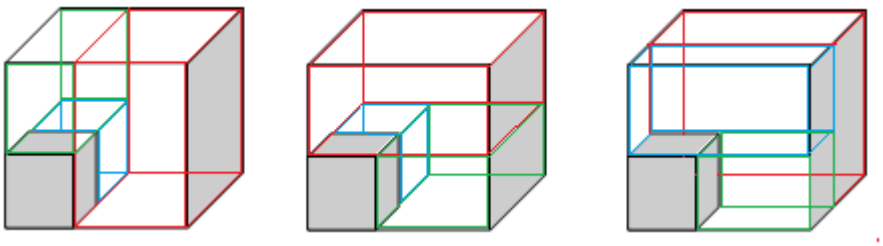


Figure 1. Different division of space

3.2 Optimization phase

The optimization phase is implemented using a genetic algorithm. This phase can also be implemented with other algorithms, but for the needs of this research we have used an already established approach. The chromosome is represented with three lists of integers (Figure 2). The first list represents the index of the elements in the element set. The second list is the orientation of the elements. The third list represents the side of the previous element to which the current element is placed to. With this representation we represent the whole loading plan. The loading starts from the left bottom corner and each element are placed next to the previous one using a bottom-up approach.

The crossover and mutation operations are done separately on each sub-list and are adapted to be in accordance with this representation [9].

The fitness function is implemented as a penalty function. If an element exceeds the available space the algorithm assigns 100 negative points to this solution. The second measure is the volume of the void space and can be from 0 to 100. The best solution is the solution with minimum negative points.

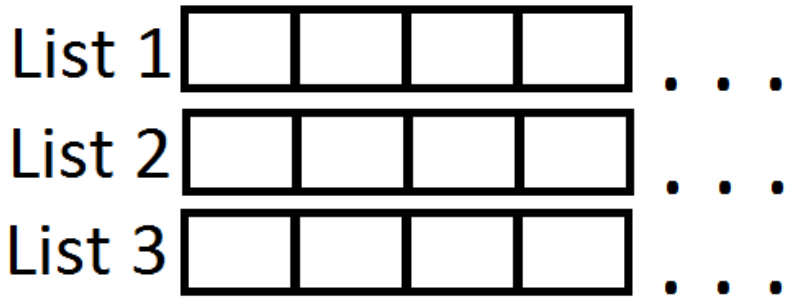


Figure 2. Chromosome representation

3.3 Heuristics

Both of the phases implement heuristics in order to use expert knowledge during the placement of the elements. Using the expert knowledge we take in consideration a lot of important factors like fragility of the elements, the max weight that can be placed on top of them, the allowed orientation of the elements. The expert knowledge is implemented as a set of rules. When an element needs to be placed in a space, the algorithm checks the rules and decides if the element can be placed that way. For example, when an element needs to be placed inside another element by the grouping algorithms, or when the element needs to be placed next to another element by the genetic algorithm operators, the algorithm checks if the element can be rotated. If not, the algorithm does not consider this case as a possible solution which results in a reduced time complexity of the algorithm.

4. Experimental results

The utilization of hidden spaces is highly dependent of the elements structure and properties. If the element set is consisted of elements that have a lot of free space (pipes for instance) and also the element set contains small elements, this algorithm is much better than the ones that work only with regular geometric forms and without considering the holes in the elements.

We have made different simulations with different variations of the algorithm. The experiment was done with a small data set (with 10 elements) in order to compare the different variations. As a next step we are going to test the solution with larger data sets. At the beginning we tried the algorithm omitting the first phase. In this case we used a genetic algorithm that can arrange elements that contain holes. Then we added the grouping phase. Use of the grouping phase reduced the time needed to achieve the maximum result. With element grouping the element set was simplified which enables the genetic algorithm to reach the maximum producing less generations. We also tried to use a genetic algorithm that works only with simple geometric forms. This time we did not take into consideration the holes that stayed empty at the end of first phase because this space was already used in an optimal way. This change reduced the time complexity because the genetic algorithm fitness function was much simpler. There was no change in the efficiency of the solution despite this simplification. Figure 3 shows the comparison between different variations of the algorithm. The x axis represents the number of generations while the y axis represents the fill level of the container in percentage. The chart shows that simplification of the element set with inclusion of the grouping phase reduces the generations needed to reach the optimal solution.

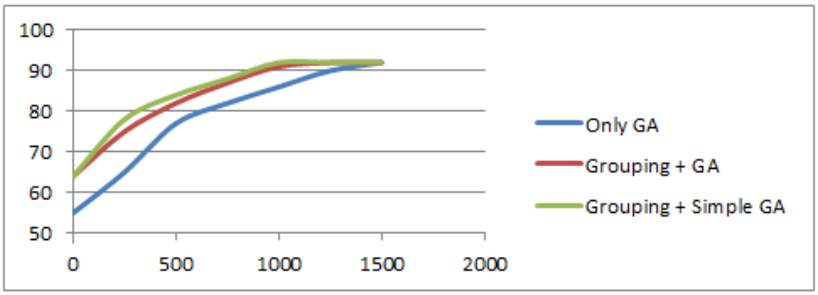


Figure 3. Comparison of the different variations

5. Conclusion

Current loading optimization algorithms have not reached the maximum efficiency and can still be improved. Using the hidden space offers a big improvement when there are a lot of elements with such characteristics. Such algorithms are applicable in many real world areas, for example for packing furniture, pipes etc.

Although this optimization problem can be solved with only an optimization algorithm, adding a simple phase at the beginning reduces the element set (depending on the structure of the element set). This enhancement reduces the time complexity of the algorithm.

Since this kind of algorithms are applicable in many areas [24] it is worth to research further in this area.

6. Future work

In the real world, the set of elements that need to be loaded is often the same with the set of some of the previous loading operations. With continual usage of the algorithm, the solutions can be stored and the next time the algorithm is initiated, it can start from the best previous solution for the given set of elements. Also, the stored solution can be applied if the solution is not improved during a limited number of algorithm iterations.

Also, implementing packing patterns as an expert knowledge will improve the time complexity and the efficiency of the algorithm. The pre-saved patterns can be a starting point for the algorithm.

We also want to improve this algorithm by implementing more factors as expert knowledge. By implementing new factors the solution can be applicable for optimizing the loading for transport purposes.

7. References

1. Bai-Sheng Chen, Yu-Fu Huang, Intelligent Cargo Loading System for Two-stages Truck Loading Problem, Takming University of Science and Technology, 2011
2. Li Pan, Joshua Z. Huang, Sydney C.K. Chu, A Tabu Search Based Algorithm for Cargo Loading Problem, University of Hong Kong, 2008
3. Shigeyuki Takahara, A Multi-start Local Search Approach to the Multiple Container Loading Problem, Kagawa Prefectural Industrial Technology Center Japan, November 2008
4. Rafael García-Cáceres, Carlos Vega-Mejía and Juan Caballero-Villalobos, Integral Optimization of the Container Loading Problem, Escuela Colombiana de Ingeniería & Pontificia Universidad JaverianaColombia, 2011
5. Oana Muntean, An Evolutionary Approach For The 3d Packing Problem, Proceedings of the International Conference on Knowledge Engineering, Principles and Techniques, 2007
6. Santosh Tiwari, Development and Integration of Geometric and Optimization Algorithms for Packing and Layout Design, Clemson University, 2009

7. Kelly FOK, Ming Ka & Andy CHUN, Hon Wai, Optimizing Air Cargo Load Planning and Analysis, Department of Computer Science City University of Hong Kong, 2004

8. H.T. Dean, J. N. Baggaley and R.J.W. James, Three Dimensional Container Packing of Drums and Pallets, University of Canterbury, New Zealand, 1999

9. Ilkka Ikonen, William E. Biles, Anup Kumar, Rammohan K. Ragade, John C. Wissel, A genetic algorithm for Packing Tree-Dimensional Non-Convex Objects Having Cavities and Holes, University of Louisville, 1997

10. G¨unther R. Raidl, Gabriele Kodydek, Genetic Algorithms for the Multiple Container Packing Problem, Department of Computer Graphics Vienna University of Technology, 1998

11. Shyi-Ching Liang and Chi-Yu Lee, Hybrid Meta-heuristic for the Container Loading Problem, Department of Information Management, Chaoyang University of Technology, 2007

12. Eva Hopper, Two-dimensional Packing utilising Evolutionary Algorithms and other Meta-Heuristic Methods, University of Wales, Cardiff, May 2000

13. Sadaaki Miyamoto, Yasunori Endo, Koki Hanzawa, Yukihiro Hamasuna, An optimization system for container loading based on metaheuristic algorithms, University of Tsukuba, Ibaraki, Japan,

14. Guntram Scheithauer, Algorithms for the container loading problem, Dresden University of Technology, 1992

15. Nikhil Bansal, Alberto Caprara and Maxim Sviridenko, Improved approximation algorithms for multidimensional bin packing problems, IBM T.J. Watson Research Center & DEIS, University of Bologna, 2006

16. Tobias Fanslau, Andreas Bortfeldt, A Tree Sarch Algorithm for Solving the Container Loading Problem, University of Hagen, 2008

17. Mykolas Juraitis, Tomas Stonys, Arūnas Starinskas, Darius Jankauskas, Dalius Rubliauskas, A Randomized Heuristic For The Container Loading Problem: Further Investigations, Department of Multimedia Engineering, Kaunas University of Technology, 2006

18. Robert H. Storer, Joseph C. Hartman, The Container Loading Problem with Tipping Considerations, Lehigh University

19. Reinaldo Morabito, Marcos Arenales, An AND/OR-graph Approach to the Container Loading Problem, Universidade Federal de Sao Carlos & Universidade de Sao Paolo, 1994

20. A. Bortfeldt, H. Gehring, D. Mack, A parallel tabu search algorithm for solving the container loading problem, A. Bortfeldt et al. / Parallel Computing 29, 2003

21. Søren Gram Christensen, David Magid Rousøe, Container loading with multi-drop constraints, Technical University of Denmark, 2007

22. Felix T. S. Chan†, Niraj Kumar and Tse Chiu Wong, Three-Dimensional Air-Cargo Loading Problem: An Evolutionary Algorithm Based Approach, Department of Industrial and Manufacturing Systems Engineering, The University of HongKong, 2006

23. Sadaaki Miyamoto, Yasunori Endo, Koki Hanzawa, and Yukihiro Hamasuna, Metaheuristic Algorithms for Container Loading Problems: Framework and Knowledge Utilization, Journal of Advanced Computational Intelligence and Intelligent Informatics, 2007

24. Riste Marevski, Kostadin Solakov, Enhancing Company's Every Day logistics Using Cargo Loading Optimization, International Conference for Entrepreneurship, Innovation and Regional Development, Sofia 2012

S. Markovski, M. Gusev (Editors): ICT Innovations 2012, Web Proceedings, ISSN 1857-7288

MPI Parallel Implementation of Jacobi

Edmond Jajaga[1] and Jolanda Kllobocishta[2]

[1]South East European University, eLearning Center
e.jajaga@seeu.edu.mk
[2]State University of Tetova, Department of Mathematics
jolanda.kllobocishta@hotmail.com

Abstract. Parallel computing has become a key technology to efficiently tackle complex scientific and engineering problems. The ability of parallelism of an algorithm provides a useful rationale to recourse to it. Forgotten in years, the Jacobi algorithm has been identified with a high parallelism affinity, which is used today as a preconditioner for multigrid methods. This paper describes the message passing implementation of Karniadakis and Kirby as tested in a 9-node cluster with distributed memory. We outline details about tasks distribution and their mapping to the processes accompanied with a thorough description of communications running in a parallel environment.

Keywords. iterative methods, systems of linear equations, C++ MPI, parallel programming

1 Introduction

There exists a lot of research regarding methods for solving systems of linear equations of the form Ax=b. Many algorithms from different math and computer science researchers are currently in place. Basically they are separated into two groups: direct and iterative methods. From the non practical application of Gaussian elimination as a direct method for solving systems Ax=b, iterative methods were widely studied from the scientists. There are lot of advantages and disadvantages between Gaussian elimination as an example of direct method and an iterative one like Jacobian [10].

The general framework of an iterative process is as simple as this: first, an initial assumption-solution is generated intuitively for the vector-solution $x^{(0)}$. Then, using this assumption the algorithm provides us with a possible solution $x^{(1)}$. Now the role of the solution $x^{(1)}$ becomes the input for the next possible solution. This process goes repeatedly, providing an array of vector-solutions $x^{(0)}$, $x^{(1)}$, $x^{(2)}$…, until we get into a satisfactory solution. Apparently there are a lot of questions which need to be answered: how does the initial assumption will be like? What kind of algorithm should be used? Do my iterative results converge to "the real" solution, and if yes, will they converge as fast as they would be better then Gaussian elimination? Much work has been done in this direction and it has been revealed that for certain algorithms and

S. Markovski, M. Gusev (Editors): ICT Innovations 2012, Web Proceedings, ISSN 1857-7288

certain types of matrices of A i.e. if A is symmetric and positively determined or at least not singular, the solution actually will converge.

Some of classical iterative methods include: Jacobian, Gauss-Seidel and Richardson method. In this paper we will focus on the Jacobian method as one of the oldest iterative methods. It is worth mentioning that the most modern method for applications of numerical calculations seems to be method of Crylov subspaces [6].

Because of the high level of parallelization we have addressed the parallel implementation of Jacobi iterations. Currently, there exist some parallel implementations of Jacobi iterations like in C, C++, Fortran77 and Fortran90 [7, 9], CUDA and OpenGL [8]. We have followed the implemented Jacobi iterations given in [1] in a parallel environment through library routines of Message Passing Interface (MPI) of C++ language. MPI is designated for high performance on massive parallel machines and in cluster workstations. The application is implemented and tested in SEEUCluster of the South East European University. This cluster consists of 9 workstations PC computers 1.5Ghz/128MB/20GB operating in Linux Red Hat Enterprise 4.0 operating system, configured with OSCAR 4.2 (Open Source Cluster Application Resources) software. LAM 7.0.6/MPI 2 C++/ROMIO - Indiana University is the parallel environment for programming.

The paper is organized as follows: Section 1 gives the mathematical background of the Jacobi iterative method; the decomposition method and process mapping are treated in Section 2; a thorough description of the implemented parallel algorithm of Jacobi with MPI routines is given in Section 3.

2 Jacobi iterative method

When solving the systems of linear equations of the form:

$$\begin{cases} a_{11}x_1 + \cdots a_{1n}x_n = b_1 \\ \vdots \qquad\quad \ddots \qquad \vdots \\ a_{n1}x_1 + \cdots a_{nn}x_n = b_n \end{cases} \quad \text{or} \quad Ax = b \qquad (*)$$

there exist two kinds of methods [1]:

Direct methods, through which we obtain the solution of the system (*) after a finite and known number of steps and the error of these kind of methods is 0, therefore are also called exact methods.

Iterative methods, which solving the system (*) produce an array of approximate values (vector-solution), which converges in some defined circumstances to the exact solution of the system.

The Jacobi method is one of the oldest iterative methods which are engaged in solving the system (*). In order to describe the method procedure firstly we set the system in more proper form like the following:

$$\begin{cases} x_1 = -\dfrac{\sum_{j=2}^{n} a_{1j}x_j}{a_{11}} + \dfrac{b_1}{a_{11}} \\ \quad\vdots \qquad\ddots\qquad \vdots \\ x_i = -\dfrac{\sum_{\substack{j=1 \\ j\neq i}}^{n} a_{ij}x_j}{a_{ii}} + \dfrac{b_i}{a_{ii}} \\ \quad\vdots \qquad\ddots\qquad \vdots \\ x_n = -\dfrac{\sum_{\substack{j=1 \\ j\neq i}}^{n-1} a_{nj}x_j}{a_{nn}} + \dfrac{b_n}{a_{nn}} \end{cases} \quad \text{or} \quad x^{(k+1)} = Tx^{(k)} + c \qquad (1)$$

Hence, the formula in terms of its elements would look like the following:

$$x_i^{(k+1)} = \frac{1}{a_{ii}}\left(b_i - \sum_{j\neq i} a_{ij}\, x_j^{(k)}\right), \quad i = 1, 2, \ldots, n. \qquad (**)$$

Formula (**) meanwhile represents the Jacobian iterations. This method assumes we have all the input values of x in the previous iteration (k). But, usually there are not given all the x values. What we can do is to make an initial prepossession for x and to generate another group of solutions for x from the equation (**), which in fact will represent the input values for the next iteration (k+1). After we have found the group of x values for the previous iteration we continue generating new groups again and again until we arrive at an acceptable solution. These iterations produce an array of approximate values for the "real" solution of the system (*).

With the word "acceptable solution" we will intend that group of solutions, respectively to the approximations, of x with the required accuracy.

If the vector-solution values are getting closer to the expected solution with the growth of iterations, then it is said that it converges to the solution of the system (*). In the majority cases the approximation array results with a good estimation for the values of x i.e. converges. Note that this algorithm is nonfunctional for all matrixes. One of the wide exceptions is any matrix with any 0 in diagonal.

We stress out the following requirements of the Jacobi method, in order for it to be functional:

1. All the members of the main diagonal of the matrix A must be nonzero ($a_{ii}\neq 0$, $\forall i = 1,2,..n$).
2. This method requires a duplicate storage for the vector-solution x. This, because none of the members (elements) of x can be overwritten while all the x-elements are calculated for that iteration i.e. it is needed an array of current solution to manipulate with and another array, we will note it with *xold*, holding values of the previous iteration.
3. Components (elements) of the new iteration vector solution are not dependent of each other, hence can be computed at the same time. This identifies the high potential of the parallelization of Jacobi method.

4. The solution provided by the Jacobi method not always converges. It is ensured only within some specified circumstances, which will be discussed in the next subsection.

3 Parallelism of Jacobi iterations

Referring to the equation (**) it can be observed the high potential of the parallelism of Jacobi. In contrast to Gaus-Seidel method, in which are used x-values of the previous and current iteration when finding the x-values of the next iteration, Jacobi puts borders between iterations; values of the vector-solution x are calculated only from the vector-solution of the previous iteration (noted with xold). Because of this, when it comes to parallelization Gauss-Seidel has certain disadvantages, even that the convergence rate of the Jacobi-type iteration is no better than the convergence rate of the corresponding Gauss-Seidel iteration for any nonnegative matrix A [2].

3.1 Data decomposition

Based on equation (**) we partition the problem into the following sub problems:

D1: sum1 = 0
D2: sum2 = $\sum_{j=2}^{n} a_{ij}x_j$
D3: x1 = (-sum1 - sum2 + b1)/A11
D4: sum1 = a11x1
D5: sum2 = $\sum_{j=3}^{n} a_{ij}x_j$
D6: x1 = (-sum1 - sum2 + b2)/A22
D7: sum1 = a11x1+ a12x2
D8: sum2 = $\sum_{j=4}^{n} a_{ij}x_j$
D9: x1 = (-sum1 - sum2 + b3)/A33
D10: sum1 = $\sum_{j=1}^{3} a_{ij}x_j$
D11: sum2 = $\sum_{j=5}^{n} a_{ij}x_j$
D12: x1 = (-sum1 - sum2 + b4)/A44
⋮ ⋮

If we observe the problem decomposition we will find out that there are independent sub problems like: 1, 2, 4, 5, 7, 8, 10, 11, etc. and those dependent: 3 dependent from 1 and 2, 6 from 4 and 5 etc. This way the process of parallelization looks more feasible and every process can do many of these tasks depending of the mapping topology. The dependency graph of the problem decomposition would look like in Fig. 1.

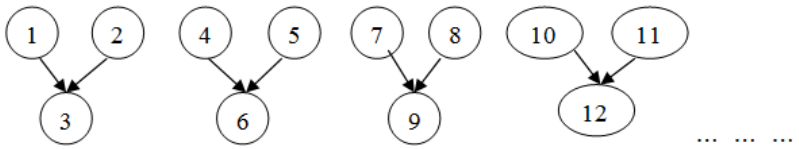


Fig. 1. Dependency graph of the decomposed sub problems

For a more balanced workload, the matrix dimension is divided with the total number of processes which will be initialized in the MPI environment; the obtained result will be the number of rows to be computed by a process, except the last process which will in turn get also the remaining last rows of the residue.

4 MPI implementation

MPI as a message passing library accommodates a natural and easy partitioning of the problem, it provides portability and efficiency, and it has received wide acceptance by academia and industry [1].

Function ReadMatrix serves as input for capturing coefficients of matrix A of system (*), while ReadVector serves for initialization of values for right side vector of (*), namely b. These functions are called from the main function main() and will be executed only by process 0, while at this stage pointers of matrix A, and vectors b and x from all other processes are null.

Within the main function of the application of Jacobi parallel implementation i.e. main() function, the following variables are declared:

- `nbproc` will hold the number of nodes which will do calculation in the parallel environment
- `myrank` indicates the corresponding number of the node (process) of the base communicator COMM_WORLD
- `n` holds the dimension of the square matrix A_{nxn} and also the number indicating the amount of variables of vector solution `x[n]`
- `abstol` provides the permitted error of the required result. We have instantiated it with $\varepsilon = 10^{-2}$ as a random value.

In order to give the user free choice of matrices values there are used dynamic allocation arrays. This rationale does not create close blocks, which means the operating system is more comfortable when deciding where to place things in memory. All the arrays used in our application use this methodology. After the instantiation of the input values the MPI routines for initialization and setting the parallel environment are executed.

The value of n received from the user will be distributed through the MPI routine Bcast to all the initialized processes by the communicator. In order to ensure all the processes received what was dedicated to them it is used Barrier() function.

After the input instantiation the function Jacobi takes the control of running the application. This function implements the Jacobi iterations based on system (**). It receives the number of processes initialized (mynode), the total number of nodes (numnodes), matrices dimension (N), matrix A as double** array, a vector x where will be placed the solution, a vector b holding the right vector of the system (*) and the permitted error (abstol). The main node is considered process 0 which delegates the tasks to other nodes, returns the vector solution and the number of iterations maxit.

As mentioned earlier on Section II. 2., the rows per node mapping are done through `rows_local` variable. It records the number of rows each node is responsible for calculation. For example, if `N=10` and number of nodes is `numnodes=3` than the first two nodes will make calculation within an iteration for 3 rows (since `floor(10/3)=3)`, whereas the last node will do calculations for `rows_local = 10-3*(3-1) = 4` rows.

After specifying the number of rows per node, next comes the distribution of matrices A and b from the equation $Ax = b$.

Distribution of matrices rows are done from node 0, because it holds the whole matrices. It sends the tasks to other processes based on their corresponding rank (node number) through the MPI routine for point-to-point communication, Send, specifically for our example:

- for i=1 a message will be send to *process 1* for j = 0, 1, 2 → 3 rows i.e. `A[1*3+(0,1,2)]` or `A[3], A[4]` and `A[5]`
- for i=2 a message will be send to *process 2* for j = 0, 1, 2, 3 → 4 rows i.e. `A[2*3+(0,1,2)]` or `A[6], A[7], A[8]` and `A[9]`

Rows `A[0], A[1]` and `A[2]` will be processed from process 0 because they are already on process 0 and there is no need of distribution. Note that the rows distribution of process 2 is done through the last rows distribution code. Likewise to matrix A distribution, the vector b distribution is done.

Normally, before task reception of processes (except process 0), because we are dealing with dynamic allocation arrays, it is required to create their structures in these nodes and then through Recv routine to be received messages dedicated for the corresponding nodes.

Vectors x and b will be of length `rows_local` i.e. every node will process the rows taken over. Specifically for our example, process 1 will process rows 3, 4 and 5; therefore it will return the values for variables x3, x4 and x5, in each iteration.

After rows distribution has been performed, the Jacobi iterations are ready to begin. But yet to do it, locations are reserved in corresponding processes for values that will be hold for vectors `xold`, `count` and `displacements`. Vector xold will serve for saving the vector solution from previous iteration of the current one. Understandably that this vector is of length N i.e. the number of variables of system (*) and as initial values for its components we will take 1. The other two vectors length is of the same as the communicator size, namely the value of variable numnodes. In our

example these vectors would be of length 3. Before entering the final stage of the Jacobi iterations the following variables needs to be described:

- `i_global` holds the indices of rows processed by the corresponding nodes i.e. on process 2 `i_global` takes values 6, 7, 8 and 9, while on process 0 takes values 0, 1 and 2 and on process 1 takes 3, 4 and 5.
- `displacements` is an integer type vector of size of the communicator. Input *i* specifies the dislocation (depending on the recvbuf) where will be placed data incoming from process *i.*
- `local_offset` the local swerve or the index of the first row of matrices A and b of each process. For example, process 1 ka `local_offset=3` i.e. the first row that will be processed in this node is the third row.
- `count` saves the number of rows which are received from the previous vector solution (`xold`).

As an example, for testing the Jacobi implementation so far explained a system of 10 linear equations is used running on 3 nodes. The Jacobi iterations are implemented with the following lines of code of the Jacobi function:

```
for(k=0; k<maxit; k++){
  error_sum_local = 0.0; sum1 = 0.0; sum2 = 0.0;
  for(j=0; j < i_global; j++)
    sum1 = sum1 + A[i][j]*xold[j];
  for(j=i_global+1; j < N; j++)
    sum2 = sum2 + A[i][j]*xold[j];
  x[i] = (-sum1 - sum2 + b[i])/A[i][i_global];
  error_sum_local += (x[i]-xold[i_global])*(x[i]-
xold[i_global]);
}
COMM_WORLD.Allreduce(&error_sum_local,&error_sum_global,1
,DOUBLE,SUM);
COMM_WORLD.Allgatherv(x,rows_local,DOUBLE,xold,count,disp
lacements,DOUBLE);
  ...
```

These lines of code do the following:

- Find the values of vector solution `x[]` based on formula (**)
- Calculate the sums of the local errors referring to the new values of vector solution `x[]` and those of the previous iteration `xold[]` based on the Euclidean norm (***)

The function *Allgatherv* enables us to specify how much information we are expecting from every node. Recall that standard *Allgather* assumes that each node sends the same amount of data [3]. Specifically, nodes will send their corresponding results of rows, respectively of the corresponding components of vector solution `x`. Based on Table 1 we can indicate the following:

Process 0: sends the first three components of the vector solution `x(0,1,2)`

Process 1: sends the next coming three components of the vector solution `x(3,4,5)`

Process 2: sends the last four components of the vector solution `x(6,7,8,9)`

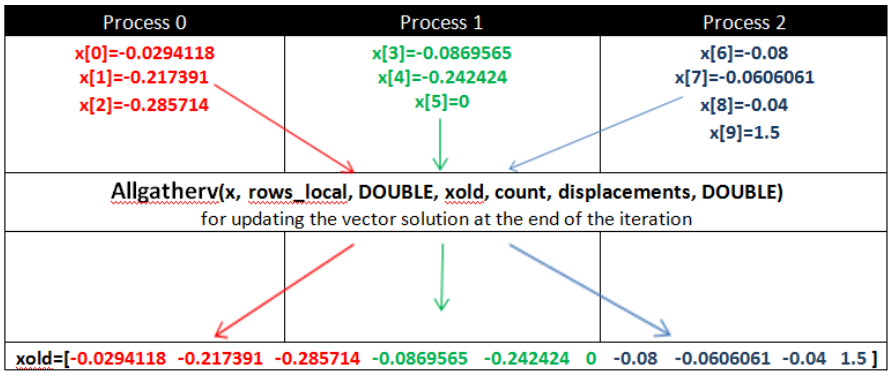


Table 1. Updating the vector solution `xold` through *Allgatherv*

For each iteration, MPI routine *Allreduce* will help us to sum up the whole local errors in one variable `error_sum_global`, which will be compared with the permitted error `abstol` (Table 2). If the result holds the required accuracy then the iterations will stop, the memory will be cleared out of arrays `A`, `b`, `x`, `xold`, `count` and `displacements` and the function will return the value of the last iteration `k`. Otherwise, the iterations will continue until an accepted solution is obtained. In order iterations to not last forever, except the Euclidian norm, there is placed another stopping criteria – the maximum number of iterations `maxit` which is instantiated in our example with the value 1000. Finally, the execution returns to the main() function, which gives the final iteration and closes the parallel environment.

During the testing of our example on SEEUCluster, an acceptable result is obtained at the seventh iteration when error_sum_global takes a value 0.00867182 with the vector solution x=[0.0267993, -0.0943336, -0.060582, -0.000628417, 0.000524287, 0.00245644, -0.00043372, -0.00262186, 0.0878125, 0.0629833].

5 Conclusion

As we outlined on this paper, the initial distribution of the main matrix A and right one b, the Jacobi method is very parallelizable. Only two MPI routines are needed in every iteration – Allreduce for calculating the error and Allgatherv for distributing the vector solution x updated at the end of each iteration. The algorithm gathers the calculated components of vector solution x from every node to formulate the whole vector solution. This method implies that matrix vector multiplications in different nodes are done independently.

As opposed to the high level of parallelism, the Jacobi algorithm suffers from numerous communications issues and do not scale easy for a big number of nodes. However, this paper demonstrates the ability of Jacobi parallelism based on message passing technology through MPI routines. It can serve as a "floor" for further studies, measurements of parallel programming performance, and comparison with other iterative methods.

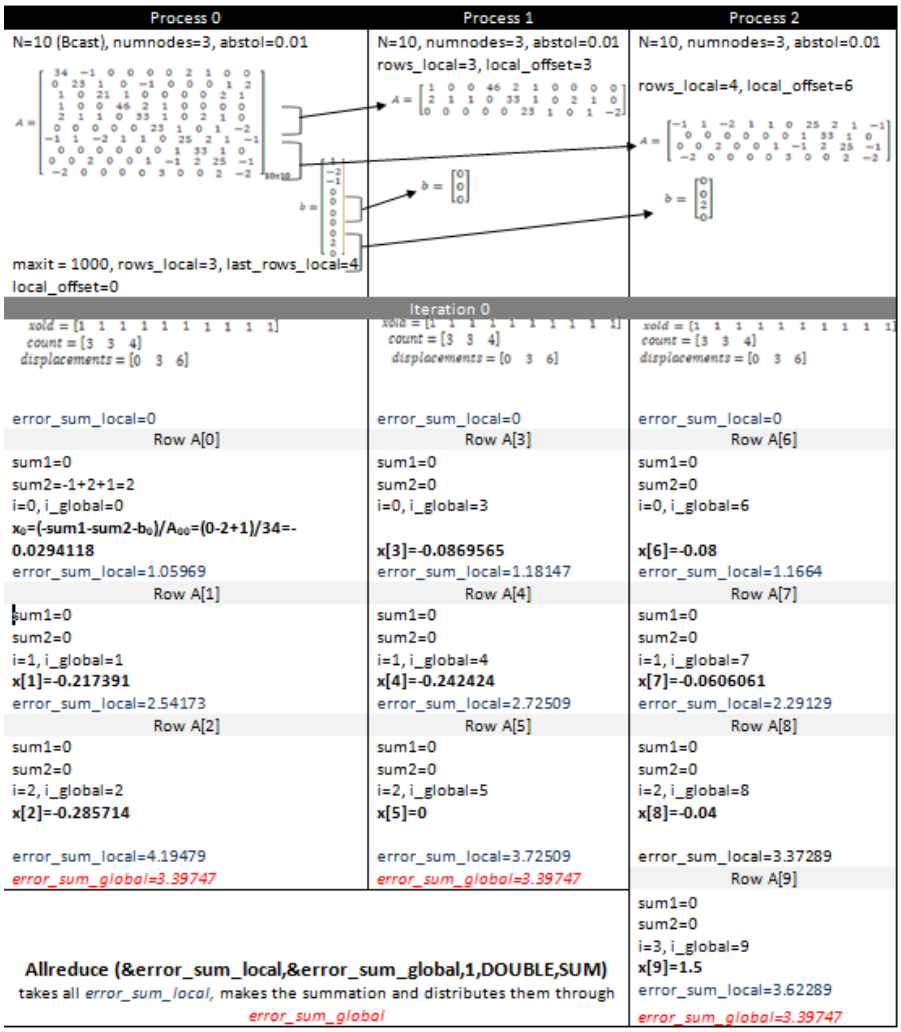


Table 2. The first iteration of our running example illustrated in Fig. 4 calculated through Jacobi method in a parallel environment with 3 nodes

References

1. G. E. Karniadakis dhe R. M. Kirby II, Parallel Scientific Computing in C++ and MPI, Cambridge University Press, 2003
2. J. N. Tsitsiklis, *A Comparison of Jacobi and Gauss-Seidel Parallel Iterations*, Massachusetts Institute of Technology, Appl. Math. Left. Vol. 2, No. 2, pp. 167-170, 1989
3. M. Snir, S. Otto, S. Huss-Lederman, D. Walker, dhe J. Dongarra, MPI: The Complete Reference, Massachusetts Institute of Technology, 1996
4. F. Hoxha, "Ushtrime të analizës numerike", SH. P. "Libri universitar", 1997
5. A. Grama, A. Gupta, G. Karypis, V. Kumar, Introduction to parallel computing, 2nd Edition, January 16, 2003
6. H. A. van der Vorst, *Krylov subspace iteration*, Computing in Science & Engineering, January/February 2000
7. JACOBI A Program or the Jacobi Iteration, Available at http://people.sc.fsu.edu/~jburkardt/vt2/fsu_open_mp_2008/jacobi/jacobi.html
8. R. Amorim, G. Haase, M. Liebmann, dhe R. W. dos Santos, *Comparing CUDA and OpenGL implementations for a Jacobi iteration*, 19 December 2008
9. Parallel Jacobi Iterative Scheme, Available at http://scv.bu.edu/~kadin/alliance/apply/solvers/jacobi_parallel.html
10. W. R. Fraser, *Gaussian Elimination vs. Jacobi Iteration*, Project report, 21 November 2008
11. S. Youssef, Iterative methods for sparse linear systems, second edition, January 2000

On the Convergence of Distance Vector Routing Protocols

Dejan Spasov, Marjan Gushev

Faculty of Computer Science and Engineering
Skopje, Macedonia
{dejan.spasov, marjan.gushev}@finki.ukim.mk

Abstract. In this paper we give an overview of distance vector routing protocols. We focus on the convergence mechanisms in two widely known distance vector routing protocols: EIGRP and RIP. With the aim to provide open source protocol, we propose a solution that inherits the simplicity of the RIP protocol and the fast convergence of the EIGRP protocol. We believe that our proposal will provide faster convergence and better scalability in large networks.

1 Introduction

In a computer network, it is vital to know the shortest paths between each pair of nodes (routers), because shortest pats are preferred choice for directing the flow of end-user traffic. In the early networking days, network administrators were manually configuring routes that were under their administrative domain. However it became obvious that this approach did not scale well and it was prone to errors. As the number of nodes in computer networks grew linearly, the number of links among the nodes grew with quadratic speed. Hence it became impossible for administrators to catch up on such a growth, i.e. to maintain best routes, to keep second the best routes as back-ups, and so on.

In the early ‘80s, routing protocols started to emerge on the commercial routers. A routing protocol is a network protocol that implements graph-based algorithm for finding shortest paths to distant networks. In addition, routing protocols specify message format and communication procedures that will allow them to share information about the remote networks. Routing protocols determine the best path to each network which is then added to the routing table. Most often, it is considered that routing protocols operate at layer three of the OSI model, with the exception of the IS-IS protocol which operates at layer two.

Internet can be seen as interconnection of separate routing domains or autonomous systems. This formulation divides routing protocols in two categories:

- *Interior gateway routing protocols* (IGPs) – protocols used for intra-domain routing

adfa, p. 1, 2012.

S. Markovski, M. Gusev (Editors): ICT Innovations 2012, Web Proceedings, ISSN 1857-7288

- *Exterior gateway routing protocols* (EGPs) –routing protocols used for routing between autonomous systems ([1]-[3]).

Interior Gateway Protocols exchange routing information within a single routing domain. Prominent members of IGP family are: OSPF, EIGRP, and RIP routing protocols. Considering the type of the shortest-path algorithm they use, these protocols are further subdivided in two categories:

- *Distance vector* routing protocols (EIGRP and RIP)
- *Link-state* routing protocols (OSPF)

In distance vector routing protocols routes to distant networks are advertised as vectors (objects with distance and direction). A *metric* must be defined within these protocols and the distance is measured according to this metric. The direction represents the neighbor router along the path to the advertised distant network. Well-known example of the algorithm for finding best routes in distance vector protocols is the Bellman-Ford algorithm. Early distance vector routing protocols were designed to periodically send their complete routing table to all neighbors. This approach guaranteed consistent routing information among all routers in a network, but did not scale well for large networks [1]-[3].

Link-state routing protocols need to have a complete view of the topology before applying the Dijkstra's shortest path algorithm. Thus the first step for link-state routers is to exchange information about the topology. In contrast to early distance-vector protocols, link-state protocols offered faster convergence with almost zero control traffic taxing the network. However, with the advent of the EIGRP it has been shown that distance vector protocols can maintain fast convergence with the same amount of control traffic as link-state protocols [1]-[3].

A disadvantage of link-state routing protocols is that if a link goes down then entire network will be down for the time the re-computation of shortest paths takes place. This can be alleviated with dividing the entire routing domain in sub-domains – but this step requires a knowledgeable administrator and more configuration commands on routers. In distance vector networks if a link goes down, only the routes that were going through that link will be unavailable for the time the re-computation takes place. Another disadvantage of link-state routing protocols is that they require more processor time than distance vector protocols.

In this paper we analyze the metric and convergence mechanisms of distance vector routing protocols. We say that a network has *converged* if all routers have complete and accurate knowledge about the network. Our goal is to propose a new routing protocol that is based on the RIP protocol.

2 Distance vector routing protocols

First we will illustrate the differences between distance vector and link-state routing protocols. Imagine a road infrastructure of a country, but without accompanying guide lines or information signs. How would a driver know to drive from city A to city B? An obvious solution is the government to install guide and information signs. This solution represents distance-vector routing protocols. Another solution is in each car the government to install GPS navigating device. This solution represents link-state protocols. The question that arises is which approach is better? It is obvious that GPS solution is more expensive, but doable; though twenty years ago this would have been impossible task.

The following example demonstrates disadvantages of link-state protocols: Assume that a distant road, (not on the route from A to B) is under construction. Then imagine that all GPS devices will be updating the topology change for 24 hours. In addition, imagine that each 60 days all GPS devices will not be working for 24 hours due to maintenance reasons.

The following example demonstrates disadvantages of distance vector routing protocols: let A' and B' are two neighboring cities on the road between A and B. Let assume that the road between A' and B' is closed for repair. Then a driver will be driving a car in a loop around city A and its neighboring cities.

When we speak about computer networks, we use the term *autonomous system* to refer to a collection of routers interconnected with links and operated by single administrative authority. End-users, or hosts, usually are not considered in the network model. *Routers* are special-built computers with the ability to find the shortest path to each network in the entire domain. Shortest paths to all networks in the autonomous system are kept in the fast memory of the router as a special data structure known as *routing table*. For each network in the autonomous system there is only one entry in the routing table usually composed of: the IP network address, metric, next-hop router, exit interface and expiration timer.

The existence of shortest paths implies that there must be a metric by which routes will be measured and compared. Simple metrics are based on hop count, or the number of transiting routers, while more complex metrics include bandwidth and delay in their calculations, even the waiting times in router's ques. Usually, a concrete metric value is referred as *cost* – which is a term from weighted graphs. Let $d(i,j)$ represents the cost between edges i and j. We will assume

$$d(i,j)=\begin{cases}\infty & i \text{ and } j \text{ are not adjacent}\\ \in \mathrm{N} & i \text{ and } j \quad \text{are} \quad \text{adjacent}\end{cases} \tag{1}$$

Assuming that costs are additive, the best metric between any two nodes $D(i,j)$ can be found by finding the minimum

$$D(i,j)=\min_{k\in N}\{d(i,k)+D(k,j)\} \tag{2}$$

where N represents all routers and $D(i,i)=\infty$. It has been proved that procedure (2) will lead to shortest paths and several algorithms have been designed according to this procedure [4].

The theory of shortest paths on graphs, though useful in finding the shortest routes, does not solve all problems that may show up in reality. For example, networks have frequent changes in topology due to router failure or network maintenance. The mechanism that distance vector protocols use against router crashes is route timing out. For example, in RIP routing protocol the time out mechanism is set to 180 seconds. If a router does not get an update message that a certain route is alive for 180 seconds, it declares that route unreachable. If a network becomes unreachable, the nearest router upon noticing this will advertise that network as unreachable. Each distance vector protocol has reserved a special *infinity-metric* value for unreachable destinations [4].

The above procedure equipped with route time out and infinity metric will always converge to appropriate shortest paths for each router. However, we did not mention the time needed for routers in a network to converge to shortest paths list. For example, consider a simple network of four routers (Fig. 1) and assume that all routers are in a state of consistency, i.e. routers A and B know that to get to the server S their packets must pass through C [4].

Now assume that connection between C and D fails. With the help of timeout timers, C will notice that D is unreachable, but meanwhile A and B will falsely advertise a route to D through themselves. C will accept this false advertised route with bigger metric and it will advertise back to A and B a slower and unreachable route. This process of mutual deception, known as "counting to infinity", will continue until infinity metric has been reached. In RIP, for example, hop count is used as metric and the infinity metric is represented by the number 16. Route time-out timers are set to 180 seconds. This means that the convergence process of the network on Fig. 1 will last unacceptable 48 minutes.

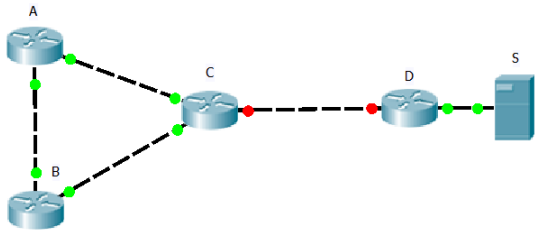


Figure 1. Example of routing instability.

Several mechanisms have been proposed to speed up the convergence of the distance vector protocols. *Split horizon* rule forbids sending back routes to the neighbor from which these routes have been learned. *Split horizon with poisoned reverse* mechanism will advertise back these routes, but with infinite metric, thus improving the convergence. *Triggered updates* is a rule that requires routers to send update messages immediately when they notice a change of metric in their routing tables. The receiving routers will change their metric if their routes were through the sending router and will trigger update message to their neighbors [4].

3 Routing Information Protocol

Routing Information Protocol (RIP) is one of the oldest and still alive routing protocols. Its development began in the late '70s from the Xerox's XNS protocol. The first document that describes RIP was published in 1988 [5], however recent RFC extensions that were proposed to support IPv6 [6] and cryptographic authentications [7] secured its future existence.

RIP metric is an integer between 1 and 15, with 16 being reserved for infinity. The way the costs for traversing networks are associated is not specified in the standard, but due to the limit of 15, the cost is usually 1. This is the well known *hop-count* metric used by RIP.

RIP packets are encapsulated in UDP segments before being sent over IP network. RIP configured routers send and receive RIP packets on port 520. RIP packet format is given on figure 2 [4]:

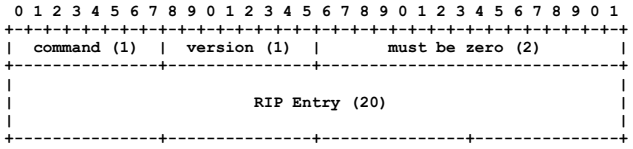


Figure 2. RIP packet format.

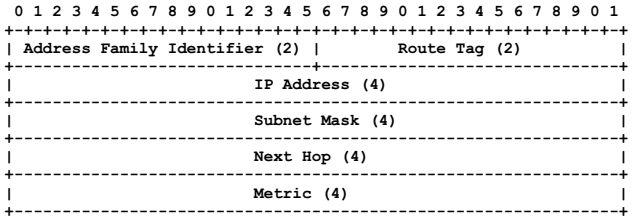


Figure 3. RIP entry.

We can notice that RIP packets are aligned on 32 bit boundaries. Version field (fig. 2) helps to distinguish between RIP version 1 and RIP version 2 packets. The command field defines two types of messages:

1. *Request* – from a neighbor router to send all or part of the routing table
2. *Response* – from the neighbor router with all or part of the routing table.

Each RIP packet (fig. 2) can carry information for up to 25 routes. Parameters requested or sent back for one route are carried with one RIP entry (fig.3). Response packets can be generated for three reasons: response to request packet, regular update or triggered update. Every 30 seconds each router will send its routing table to every neighbor with response packets. In order to avoid synchronization and unnecessary collisions over broadcast networks, each 30 second interval is jittered with a small random time less than 5 seconds. Triggered updates can over flood the network. Thus

after a triggered update is sent, a timer is set for a random time less than 5 seconds. If other trigger events occur before the timer expiration, a single update is sent after the timer expires. The timer is then reset to another random value between 1 and 5 seconds [4].

With each route, RIP process on a router associates two timers: the *time out* timer (as described in previous section) and *garbage collector* timer. Once a route enters the routing table, the timeout timer is set to 180 seconds and reset each time if an update for the route is received. If the timer expires, the garbage collector timer is set to 120 seconds and the route is considered unreachable. The route will remain in the routing table for the duration of the garbage collector, but it will be advertised as unreachable. After the garbage collector expires, the route is removed from routing table [4].

It is obvious that if we increase infinity in the RIP protocol, we will create more space for manipulating route costs. However, this will create backward compatibility problem and will confuse older versions of RIP. The best thing we can hope is that older versions will ignore routes with costs greater than infinity. Thus, the committee responsible for maintaining the RIP standard remained adamant to demands for increasing the infinity value.

4 Enhanced Interior Gateway Protocol

Enhanced Interior Gateway Routing Protocol (EIGRP) is a CISCO proprietary distance vector routing protocol that was developed to address shortcomings of the RIP protocol, like the hop-count as a metric, maximum network diameter of 15, and the periodic broadcasts of the entire routing table [1]-[3].

The proprietary part of EIGRP is protected with *Protocol Dependent Modules* (PDM) and *Reliable Transport Protocol* (RTP). Protocol Dependent Modules gave to EIGRP capability to operate over various Layer 3 network protocols: IPv4, IPX, and AppleTalk, while RTP provided connectionless and connection oriented services over these networks. In other words, RTP offers TCP-like and UDP-like services to EIGRP that do not depend on the protocol stack.

EIGRP defines four packet types needed for its communication: Hello, Update, Acknowledgement, Query, and Reply [1]-[3].

The first thing an EIGRP router must do is to establish adjacency with its neighbors. This is done with the help of *Hello* packets and this is lifelong adjacency. Hello packets are usually exchanged over 5 second intervals.

The next step for neighbor routers is to exchange routing information. This is done with *Update* and *Acknowledgement* messages. This communication is connection oriented thus eliminating the need for periodic route refreshment and route timeout timers. A new Update message for a particular route is sent only if the metric for that route changes. EIGRP uses the term partial and bounded to describe Update messages. Partial refers to the fact that only routes with changed metric are included in the update, and the term bounded refers to the fact that updates are sent only to those routers affected by the change.

If a route becomes unavailable, *Query*, and *Reply* messages are used in the search for alternative routes. Again, these two types of messages are sent in connection-oriented manner accompanied with Acknowledgement message.

EIGRP uses the most complex metric of all routing protocols. It can be made of four parameters: bandwidth, delay, reliability, and load. Reliability and load are dynamic parameters measured at each interface, but they are seldom used in calculations. Thus most often the well-known *bandwidth + delay* formula is used in metric calculations. Let $L(R,D)$ be a route from the router R to a destination D. Let $L(R,D)$ be made of links l_i with bandwidth w_i measured in bps and delay d_i. Then the EIGRP's bandwidth+delay metric for $L(R,D)$ is computed according to the formula

$$metric\{L(R,D)\} = \frac{256 \cdot 10^7}{\min\{w_i\}} + \frac{256}{10} \cdot \sum d_i \quad (3)$$

The bandwidth is usually specified on the interface by the producer. Cisco's defaults are 100 Mbps for LAN interfaces and 1.544 Mbps for WAN interfaces. Default delays on Cisco's routers are given on the following table:

Media	Delay
100 M ATM, Fast Ethernet, FDDI,	100 μs
T1, 512K, DSO, 56K, 1HSSI	20 000 μs

Table 1. Default delays on Cisco routers.

EIGRP uses *Diffusing Update Algorithm* (DUAL) to perform the shortest path computation. Although, it is still a distance vector protocol, it is advanced version that is supposed to be better than the Bellman-Ford algorithm that is used by RIP.

In order to explain how it works, we have to explain the terms used by DUAL (all terms refer to one destination):

1. *Successor* – this is the next-hop neighbor on the route to a destination network;
2. *Feasible Distance* (FD) – is the best (lowest) metric to the destination network;
3. *Feasible Successor* (FS) – is a neighbor who has a loop-free backup route, should any router on the best route fails;
4. *Reported Distance* (RD) – is the feasible distance to the destination network of the neighboring routers

Feasibility Condition (FC) – is a criterion based on which backup loop-free routes to destination network are found. EIGRP's DUAL algorithm maintains a *topology table* separate from the routing table. The topology table includes the best path to a destination network and backup path (via the Feasible Successor) that DUAL has found to be loop-free. In order a neighbour to qualify for Feasible Successor, it has to pass the Feasibility Condition:

$$FD > RD \tag{4}$$

In [8] it has been proved that if (4) holds true, that neighbor has loop-free path to the destination network (a path that does not pass through the router that performs the feasibility test.

The EIGRP protocol explained so far, though better than rip, will not scale well on large networks. Thus several patches have been proposed that improve the convergence time in large networks. Stuck-in-active, stub router, graceful shutdown, graceful restart, multiple AS support, and so on.

5 Analysis of the EIGRP Protocol

EIGRP's superiority has been attributed to the use of the DUAL algorithm, while other distance vector protocols use inferior Bellman-Ford or Ford-Fulkerson algorithms. We believe that main advantage of EIGRP over RIP is EIGRP's metric. Using (3) EIGRP is capable of finding faster routes than RIP. However, we believe that this metric does not always find shortest paths. Consider the network on figure 4.

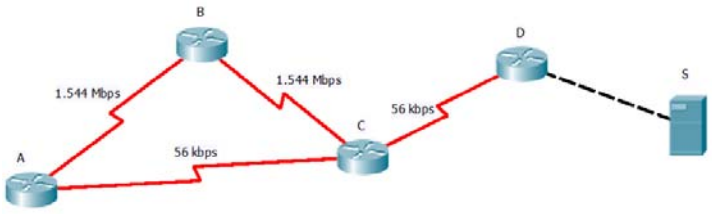


Figure 4. Shortest path demonstration on EIGRP network.

The first thing an administrator should do to ensure proper operation of the EIGRP protocol is to set appropriate bandwidth values on each router's interface. However, if default values for the delay are used, then the shortest path from the router A to the server S would be

`A->C->D->S`

On this example intuitively it is clear that the shortest path is `A->B->C->D->S`. The reason for error in EIGRP's computations is the default value for delay over serial links. The default delay value is same for T1 and 56k links (table 1). Some scholarly papers [10] suggest that we use propagation delay in (3). We believe that this approach is too expensive and again will lead the protocol to wrong conclusions. The best option is to use the serialization delay in (3), thus (3) will become

$$metric\{L(R,D)\} \sim \sum \frac{1}{w_i}. \tag{5}$$

This result is proportional to the metric of the OSPF link-state routing protocol ([1] ch. 11).

It is frequently advertised that one of the main advantages of EIGRP over RIP is that EIGRP maintains a topology table in which it stores backup routes. These backup routes are used if the best route fails, thus reducing the convergence time. We believe that this mechanism of EIGRP is not very useful. Consider the example on figure 5.

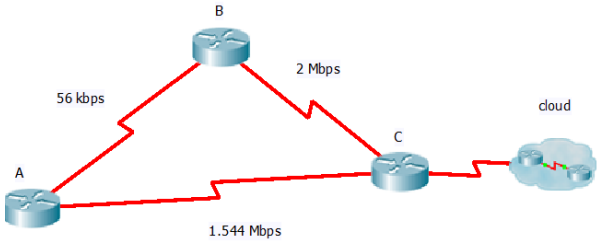


Figure 5. Backup route demonstration on EIGRP network.

In this example, only router A will have a backup route through router B to cloud networks. If the link A-C fails, router A will use its backup route through the link A-B. If the link B-C fails, then B engages in regular search for new route and eventually will find that it can use the link A-B. If any link in the cloud fails, then backup link may contribute to longer convergence, by installing wrong paths. Eventually, this problem will be solved with split horizon and triggered updates mechanisms.

To summarize, with Feasibility Condition (4) only a fraction of the routers in a network will have backup routes in their topology tables. Failure of even smaller fraction of links in a network will trigger proper usage of the EIGRP's backup route mechanism.

Finally, we would like to emphasize that EIGRP does not use periodic updates for disseminating routing updates, but a sophisticated system of queries and replies. This mechanism provides faster convergence in case a route becomes unavailable.

6 New RIP-like Routing Protocol

The purpose of this paper was to analyze the convergence mechanisms of the EIGRP protocol and to propose new RIP-like routing protocol. With this proposal we are proposing a new protocol that is not backward compatible with previous RIP protocols. Our work is based on the implementation of the RIP protocol in the open-source Qugga routing software [11] .

The first upgrade that was done was to improve RIP's metric. Our solution is to introduce additional field in RIP entries (fig. 6). From previous chapter we concluded that it would be simpler and more efficient if we use OSPF's metric instead of EIGRP's.

The second step in building a better routing protocol is to eliminate the dependence on periodic broadcast updates and route aging timers. These mechanisms provide reliable but slow convergence. In order to achieve faster and reliable convergence, we have to use the TCP protocol for the RIP packets (fig.2). If a route becomes unreachable neighbor routers will be queried for alternative routes. This implies that stuck in active timers must be implemented. In TCP usage requires periodic UDP broadcast to

router's neighbors to establish associativity. We propose this messages to include the number of entries in the routing table. If two neighbors disagree on this number will prompt routers to exchange their routing tables.

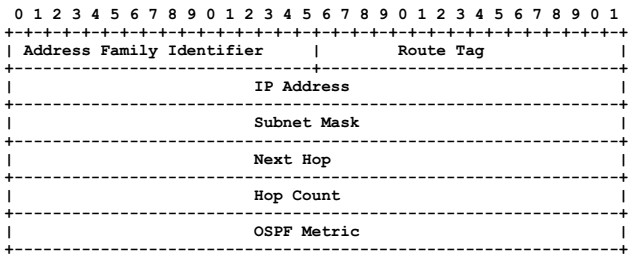


Figure 6. Modified RIP entry.

The final step will be to implement a system for finding and keeping backup routes. Here we propose more powerful feasibility condition

$$FD \geq RD \tag{6}$$

This condition is better because we require the feasible distance (FD) to be greater than or equal to the reported distance (RD). In addition we apply (6) over two metrics: hop count and OSPF metric. With our proposal more backup routes will be found. For example, with (6) router B (fig. 5) will have a backup route through router A, and vice versa. Thus we increase the number of backup routes in a computer network.

References

1. Cisco Networking Academy: CCNA Exploration 4.0 - Routing Protocols and Concepts.
2. Cisco Networking Academy: CCNP Advanced Routing Protocols.
3. Graziani, R.: Lecture notes on CCNA exploration and CCNP courses. Available at: http://www.cabrillo.edu/~rgraziani.
4. Malkin G.: RIP version 2. RFC 2453, IETF standard, November 1998.
5. Hedrick C.: Routing Information Protocol. RFC 1088, IETF standard, June 1988.
6. Malkin, G., Minnear, R.: RIPng for IPv6. RFC 2080, IETF standard, January, 1997.
7. Atkinson, R., Fanto, M.: RIPv2 Cryptographic Authentication. RFC 4822, IETF standard, February, 2007.
8. Garcia-Lunes-Aceves, J. J.: Loop-Free Routing Using Diffusing Computations. *IEEE/ACM Transactions on Networking*, vol. 1, no. 1, February 1993.
9. Albrightson, B., Garcia-Lunes-Aceves, J. J., Boyle, J.: EIGRP – A Fast Routing Protocol Based on Distance Vectors.
10. Yee, J. R.: On the Internet Routing Protocol EIGRP: Is it Optimal? In: IFORS, 2005.
11. http://www.nongnu.org/quagga

Accessibility and Inclusion in e-Learning

Mimoza Anastoska-Jankulovska, Jove Jankulovski, and Pece Mitrevski

St. Clement Ohridski University, Faculty of Technical Sciences, Ivo Lola Ribar bb
7000 Bitola, Republic of Macedonia
{jankmj2,jove.jankulovski}@yahoo.com, pece.mitrevski@uklo.edu.mk

Abstract. Rapid development of technology is opening new perspectives for business, entertainment, and education. Learning with the help of technology is providing new opportunities for learners. E-learning is grasping with big steps into diverse areas engaging more and more different methods and activities. Lot of developments and lot of experiences are present in the field of e-learning everywhere in the world. One of the important features of technology is allowing access to unprivileged, either with permanent or temporary disability. Using technology to make everything more inclusive has become important issue in developed countries. The focus of any new development is to be accessible to as wider audience as possible. How accessibility affects e-learning? This paper intends to review current experiences in the field of e-learning with the focus on accessibility. It will present basic accessibility guidelines and will reflect on how learners with special needs can access e-learning environment.

Keywords: e-learning, accessibility, inclusivity

1 Introduction

New technologies and especially Internet have become very important part of our everyday work and life. Being so, the question has arisen how accessible it is to all groups of population. Internet, or web, accessibility for all has grown to be a burning issue in European Union. Some of the countries have adopted laws on accessibility, the others are implementing different projects' activities in order to improve the accessibility of own web sites. US have adopted legislation on accessibility followed by rules and standards for accessible electronic and information technology purchased by federal agencies. There are a lot of initiatives on the international level that are helping in designing and evaluating if and how much the web site and its content are accessible. Imperative for a successful accessible web site is to start from the very beginning - in the planning phase. The web site should be planned inclusive and accessible.

Web accessibility means that it will be possible to meaningfully use the web for all people regardless of their abilities, preferences or available technologies [9]. Web accessibility means two ways communication: every person can perceive, understand, navigate, and interact with the web, and everybody can contribute to the web. From web accessibility benefits all, people with disabilities but also older people with

S. Markovski, M. Gusev (Editors): ICT Innovations 2012, Web Proceedings, ISSN 1857-7288

changing abilities due to aging, or people with temporary disabilities or just changing preferences of the web use. Web accessibility includes all abilities, including visual, auditory, physical, speech, cognitive, and neurological different level abilities.

E-learning is a new way of learning where learner is using available technology to learn in own way, on its own pace, with the tools and methods that are most adequate for him/her. E-learning is already present in our everyday life. It has influenced and changed a lot the curricula and academic teaching and learning [5]. The Web and e-learning offers the possibility of unmatched access to information and interaction for different groups of people, including disadvantaged and people with disabilities. That is, the accessibility barriers to audio and visual media can be much more easily overcome through web technologies. The web is becoming important source in many areas of our life: education, employment, government, commerce, health care, recreation. Vital is that the web is accessible to all in order to provide equal opportunity to everybody regardless their abilities or preferences [1]. An accessible web to all can also help different groups of people to participate in society more actively.

New technologies such as mobile telephony are adding to the complexity of the issue. Now the web can be accessible through most of the today's mobile devices. That is helping final users and is increasing chances for accessibility. On the other hand, it is increasing the requirements placed in front of the web developers, since everything should be working properly and should be accessible, regardless which browser is used or from which mobile device the access is required.

Another important consideration is that web accessibility is required by laws and policies in some countries. Government legislation and organizational policies can encourage inclusive design by fostering awareness of the need for inclusion and by setting broad expectations for society [4].

Web is becoming more accessible in different parts of the world and it is creating new opportunities for all users. Accessibility of web based materials is introducing a shift in education also. Accessible web is opening new doors for learners and is increasing their chances for success. So, how is this improvement of web accessibility influencing education and e-learning is important research question.

2 Making the Web Accessible

There are two aspects when talking about web accessibility. Most of the stress on web accessibility has been on the web developers' responsibilities. But also web software has an important role in web accessibility. Software needs to help developers produce and evaluate accessible web sites, and be usable by people with different types of abilities.

World Wide Web Consortium (W3C) has started an initiative Web Accessibility Initiative (WAI) [11]. One of the roles of the WAI is to develop guidelines and techniques that describe accessibility solutions for Web software and Web developers. These WAI guidelines are considered the international standard for web accessibility and are used as basis for concrete national accessibility guidelines, legislations or standards. There can be found a lot of documents describing the different web acces-

sibilities, and how concrete improvements could increase Web accessibility. It can be very useful tool while developing new web sites or wanting to improve accessibility of already existing sites, to have at least some check list with all basic requirements for web accessibility.

US have reviewed WAI guidelines and have developed own Section 508 [10], an amendment that requires electronic and information technology developed by or purchased by the Federal Agencies to be accessible by people with disabilities. Federal Agency's purchase power is tremendous. IT developers and market in general, faced with accessibility requirements from Federal Agencies will not be able to develop two parallel lines, so, will continue developing accessibility technology and software for all its clients. Most of web accessibility guidelines are present in Section 508, but are appended with technical standards, functional performance criteria and part about information, documentation and support. Representatives from industry, academics, government, and disability advocacy organizations have proposed standards for accessible electronic and information technology.

The most effective way to ensure that some web site is accessible, is to plan accessibility while starting with the sketch of the site. Making a web site accessible can be simple or complex, depending on many factors such as the type of content, the size and complexity of the site, and the development tools and environment. If planned from the beginning, many accessibility features are easily implemented. It is regardless of if it is about starting completely new Web site development or redesigning the existing one. Fixing inaccessible Web sites can require significant effort, especially sites that were not originally "coded" properly, or sites with specific types of content such as multimedia.

Web accessibility guidelines are requiring web content to be perceivable, operable, understandable and robust.

Perceivable Content.

Text alternatives for non-text content convey the purpose of an image or function to provide an equivalent user experience. Text alternatives should be provided for non-text content such as: short equivalents for images, including icons, buttons, and graphics; description of data represented on charts, diagrams, and illustrations; brief descriptions of non-text content such as audio and video files; text description for labels for form controls, input, and other user interface components.

People who cannot hear audio or see video need alternatives for multimedia. Well written text records containing the correct sequence of any auditory or visual information provide a basic level of accessibility and facilitate the production of captions and audio descriptions. Examples can be: text record and captions of audio content, such as recordings of people speaking; audio descriptions, which are narrations to describe important visual details in a video; sign language interpretation of audio content, including relevant auditory experiences.

Content should be developed in that way that it will be given the possibility to present it in different ways for different users. In order for users to be able to change the presentation of content: headings, lists, tables, and other structures in the content are marked-up properly; sequences of information or instructions are independent of any

presentation way; browsers and assistive technologies provide settings to customize the presentation.

Content should be made in such a manner that it will be easier to see and hear. Distinguishable content is easier to see and hear. Meeting this requirement helps separate foreground from background, to make important information more noticeable.

Operable Content.

Functionality should be available also from a keyboard. Many people do not use the mouse and rely on the keyboard to interact with the Web. This requires keyboard access to all functionality, including form controls, input, and other user interface components. Meeting this requirement helps keyboard users, including people using alternative keyboards such as keyboards with ergonomic layouts, on-screen keyboards, or switch devices. It also helps people using voice recognition (speech input) to operate websites and to dictate text through the keyboard interface.

Users should have enough time to read and use the content. Some people need more time than others to read and use the content. That means that users should be able to stop, extend or adjust time limits.

Users can easily navigate, find content, and determine where they are. Content that is well organized helps users to orient themselves and to navigate effectively. Meeting the above requirement helps people to navigate through web pages in different ways, meeting their particular needs and preferences. Some people may be using the content with only a mouse or a keyboard, while others may be using both. While some people rely on hierarchical navigation structures to find specific web pages, others rely on search functions on websites instead. Some people may be seeing the content while others may be hearing it, or seeing and hearing it at the same time.

Understandable Content.

Material presented by text is readable and understandable for majority of users. Content authors need to ensure that text is readable and understandable to the broadest audience possible, including when it is read aloud by text-to-speech software. Software, including assistive technology, will be able to process text content correctly having all above requirements met. These requirements help software to generate page summaries, and to provide definitions for unusual words such as technical jargon. It also helps people who have difficulty understanding more complex sentences, phrases, and vocabulary. In particular, it helps people with different types of cognitive disabilities.

Many people rely on predictable user interfaces and consistent appearance. Users can be disoriented or distracted by inconsistent appearance or behavior. That is why web developers should ensure that content appears and operates in predictable ways throughout the all pages on the web. People can easy and quickly learn the functionality and navigation mechanisms on a website. So, people can use them according to own specific needs and preferences. Some people assign personalized shortcut keys to functions they frequently use to enhance keyboard navigation. Others memorize the steps to reach certain pages or to complete processes on a website. That can be done

with predictable content because they both rely on predictable and consistent functionality of the web sites.

Forms and other interaction can be confusing or difficult to use for many people. As a result, they may be more likely to make mistakes. Web site should have explanatory interactions. If web site itself helps to avoid mistakes, or explains what mistakes were, it will be much easier for users to interact on the web. Helping users to avoid and correct mistakes can be done by: descriptive instructions, error messages, and suggestions for correction; context-sensitive help for more complex functionality and interaction; opportunity to review, correct, or reverse submissions if necessary.

Robust Content.

Robust content means it can be used with all possible browsers, assistive technologies or other agents, existing in the moment of development, without losing any information or feature. This will maximize compatibility with current and future user agents, including assistive technologies and all kind of browsers. It enables assistive technologies to reliably process the content, and to present or to operate it in different ways. This includes non-standard (scripted) buttons, input fields, and other controls.

2.1 Accessible Content Management Systems

Education, especially e-learning, is based on databases with necessary data and information used for learning processes. These databases are important part of a learning structure. Content Management Systems (CMS) are systems that are allowing storing, publishing, editing or modifying content, while big number of user can access, share and contribute to the stored data. CMS can be commercial or open source and all are with varying features. Content published on the CMS can be documents, videos, audio, text, pictures, diverse types of data, articles, etc. Therefore, evaluating CMS is necessary to determine at what level content is accessible and which one to use.

CMS are systems that, no matter how fancy and good-looking are, still are remaining just tools for managing data. If it does not comply with necessary accessibility requirements, it should be changed with something more accessible. First step in selecting CMS is thinking about its functional requirements. CMS should be able to deliver what is needed for the activities. After the functional requirements are determined, decide whether you want to follow any kind of accessibility standards or guidelines. It is advisable to follow the W3C Web Content Accessibility Guidelines (WCAG). It gives a list of recommendations, which will help to make the system more accessible. It is important to check accessibility of all the CMS: theme, module of the CMS, but also administrator part.

CMS systems are penetrating in the area of learning, hence their accessibility should be observed from the perspective of user/learner too [2]. They are increasingly drawing and maintaining the attention of a learner due to use of new technologies. Learning is taking place only in cases when learners are actively involved in it. In order to have active learners, existing content should be attractive to them, which is

taking us to the basic precondition that learners need to be able to access it in a fist place. Furthermore, content generated or uploaded by learners while using CMS should be stored and could be retrieved and used by other learners/users. That is why it is crucial to conduct capacity building activities with those users that will be filling the system with information in terms of accessibility and its requirements. All users of the CMS must have some knowledge of accessibility. Without this knowledge, no CMS can ensure that only accessible content is published to the site. Personnel using a CMS must have the necessary training and support to enable them to produce and publish accessible content.

2.2 Web Interactivity

Interactive web sites are web sites that are not just transmitting information to the users, but are allowing communication with users. They are not just simply delivering information but are adapting it according to the users' actions. There are different ways to provoke interactivity on the web sites.

The most effective way of learning is by providing interactivity. Interactivity should be secured by content that is allowing learners to communicate virtually, to exchange, to share, to collaborate. This means that both content and software are to have interactive nature that is attractive for learners. In such environment, the role of teacher/educator is changed from person that possesses the knowledge into person that is facilitating the learning processes. Interactivity is narrowing down the boundaries between learner and educator; they are starting to disappear. Both categories of users of these systems should be aware of the accessibility rules and to apply them. Especially, the content that is developed by users need to be developed according accessibility rules. Namely, learners with disabilities access and navigate the web in different ways, depending on their individual needs and preferences. Sometimes even learners without disabilities configure standard software and hardware according to their needs, and sometimes people use specialized software or hardware that helps them perform certain tasks. Acting this way, equal opportunities for learning, and inclusiveness of all learners in secured.

Some common approaches for interacting with the web include:

- Assistive Technologies - software or hardware that people with disabilities use to improve interaction with the web. These include screen readers that read aloud web pages for people who cannot read text, screen magnifiers for people with some types of low vision, and voice recognition software and selection switches for people who cannot use a keyboard or mouse.
- Adaptive Strategies - techniques that people with disabilities use to improve interaction with the web, such as increasing text size, reducing mouse speed, or turning on captions. Adaptive strategies include techniques with standard software, mainstream browsers, or with assistive technologies.

3 Evaluating the Web Accessibility

Evaluating accessibility early and throughout the development process is a must when developing a new one or redesigning an existing site. It can identify accessibility problems when it is easier to address them. Simple techniques such as changing settings in a browser can determine if a Web page meets some accessibility guidelines. Or, trying to access the web site through different mobile devices can point some of the possible issues needed improvement. A comprehensive evaluation to determine if a site meets all accessibility guidelines is much more complex.

There are different evaluation tools that can be found online and can help with evaluation. However, no tool alone can determine if a site meets accessibility guidelines. Knowledgeable human evaluation is necessary to determine if a site is accessible and to what extent [6].

Most commonly available browsers provide bookmark functionality, and screen readers provide functions to list headings, links, and other structures on a web page. Nevertheless, the design of the content is an essential factor to support different styles of navigation [3]. Examples include:

- Person with visual disabilities or from any other reason not being able to see the screen, needs to get an overview and orient themselves by scanning the headings on a web page; the headings need to be designed to also support such purposes.
- After screen magnification, person can only seeing small portions of the screen at a time, and need to orient him/herself using visual cues; the visual design needs to also support such purposes.
- The structure of web pages need to be designed to support and effective use of the keyboard because there can be a person using only the keyboard (or keyboard alternatives) to navigate through the web content.
- Web browsers need to provide supporting functionality that is easy to use and remember, since some person can have difficulty remembering the addresses, names, or particular functionality of websites.
- Websites need to provide alternative mechanisms for navigation because user can be a person who does not think and organize concepts hierarchically, as how most navigation menus are designed to be.

There are a few pitfalls to avoid when creating an accessible document (printed or electronic).

Floating graphics are a particular issue and these are generally not accessible. Word art is one example that is interpreted as a graphic and is supposed to be text. This also includes images set to float behind or in front of text.

Headers and footers are accessible by screen reader software, but the user needs to know that they are there. It can't be assumed that an access technology user will automatically look for headers and footers. Depending on the document it may be appropriate to include a line to say that information is contained in the header or footer or simply to ensure that important information is also repeated in the main body of the text.

Preparing for other needs. No matter how well designed and produced your document is, any printed material will never meet the needs of all people. Some people will not be able to read print, and may prefer the same information in another format like braille or audio, or simply by accessing the electronic file. It is always a good idea to keep a text based electronic original of your document.

Web based materials should be inclusive for all [7]. E-learning can be online and offline, but should be supportive to widest possible audience in order to be able to secure successful results. There are a lot of materials for e-learning with big variety of topics covered. Not all of them are inclusive and accessible to wide audience. Some, developed for concrete known target group, can stay that way. But others, developed for wide population should be adapted to be accessible for all. Assessment is a starting point in improving the accessibility of already existing e-learning materials.

4 Conclusion

People navigate through the web and find content using different strategies and approaches depending on their preferences, skills, and abilities. Someone using a website for the first time may need more thorough guidance than someone who has more experience with that particular website. Many functions to support different styles of navigation are built directly into web browsers and assistive technologies.

Using the Web on a mobile device with a small screen may need more orientation cues than someone using a desktop computer. While these are generally considered to be usability aspects that affect people with and without disabilities, some situations affect people with disabilities more directly. Accessibility solutions are beneficial for people with and without disabilities and are becoming increasingly available in standard computer hardware, mobile devices, operating systems, web browsers, and other tools.

E-learning materials are designed to support learning. A research of e-learning interactive materials' usage shows that different users use different approaches in order to secure cognitive learning [8]. It is clear that it depends on learning styles and also on available technology and its functionality. E-learning materials should have built-in accessibility and inclusiveness in them in order to be successful in delivering learning tool for all learners.

According to the UN's millennium development goals, access to learning must be secured for all. The audience that is using technologies for various purposes is rapidly increasing on a daily basis, as well as, possibilities for learning via web content. Various learners might have various disabilities (temporary or long-term) that might be limiting their access to digital learning content. Or, some learners might need to be able to customize the learning content. In such case, it is inevitable, that all learning content should be developed according to accessibility guidelines.

References

1. Baguma R. et al.: A Web Design Framework for Improved Accessibility for People with Disabilities, Beijing, China (2008)
2. Fletcher K.: Diary of a Trainer: Learning to Create Online Learning Experiences, West Virginia, USA (2010)
3. Guild of Accessible Web Designers; Feb 2012; <http://www.gawds.org>
4. Inclusive Design and Research Center; Feb 2012; <http://idrc.ocad.ca/>
5. Kirkpatrick D.: Who owns the curriculum? Victoria University, Melbourne, Australia, (2001)
6. Kobayashi, A.M.R et al.: Relationship between accessibility and software evolutibon, Brasil (2011)
7. Lopes, R.: The Semantics of Personalized Web Accessibility Assessment (2010)
8. Masemola S.S. et al.: Towards a Framework for Usability Testing of Interactive e-LearningApplications in Cognitive Domains, Illustrated by a Case Study (2006)
9. Paciello, M.: Web Accessibility for People with Disabilities(2001)
10. US Government: Electronic and Information Accessibility Standards (Section 508); Feb 2012; <http://www.access-board.gov/sec508/standards.htm>; <www.section508.gov>
11. World Wide Web Consortium (W3C); Feb 2012; <www.w3.org>

S. Markovski, M. Gusev (Editors): ICT Innovations 2012, Web Proceedings, ISSN 1857-7288

Germany's Local Bureaucracy in the Era of Social Media
- From Advanced eGovernment to Affordable Open Government

Dr. Norbert Jesse

TU Dortmund University, Dortmund,
Germany (Tel: +49 231 755 6221; e-mail: norbert.jesse@udo.edu

Abstract: Germany's local governments face considerable pressures: Budget restrictions, increasing citizens' expectations concerning service quality, socio-demographic distractions, and environmental issues – to name just a few. From the beginning of the Internet era, governments have tried to exploit modern ICT. For some time now, the focus has been on a new eGovernment dimension: Open Government. The basic idea is to get citizens more involved in local issues. In line with the Social media paradigm, Open Government targets openness, closeness to people and participation. From a technological perspective, it is evident that this political agenda requires robust ICT, enabling fast, flexible and affordable solutions, tailored to specific local needs. This paper outlines the path from eGovernment to Open Government and presents an affordable Open Government platform, the OpenGovernment Suite, which fits into the general IT roadmap of the German Government.

Keywords: eGovernment, Open Government, SAGA, OpenSAGA

1 Introduction

From the citizens' point of view, local communities are the most delivering administrative units in Germany. Responsible for schools, traffic issues, cultural activities and the implementation of thousands of rules and regulations, they have a direct impact on the quality of life. It is estimated that more than 90% of all contacts address local governments. But times have become increasingly difficult for governments: an explosive blend of financial constraints, socio-demographic rejections and challenges concerning education, infrastructure and the environment demand careful but effective decisions.

More than in previous years, local politicians and administrators need sensitive "antennae" for the troubles, concerns and needs of their citizens. The label "Open Government" circumscribes these efforts which target a closer interaction between local administrations and residents. The basic idea is to endow all stakeholders with flexible and barrier-free access to the many facets of government. It is expected that an intense interaction between administrations and citizens will strengthen regional identity, encourage new business models (i.e. by publishing local data) and support a local agenda setting (making use of the knowledge of locals). Moreover, while joining

S. Markovski, M. Gusev (Editors): ICT Innovations 2012, Web Proceedings, ISSN 1857-7288

the global Social media paradigm, public institutions have even more reason to discuss the challenges and opportunities of an Open Government agenda.

From the early days, ICT technology has been a pivotal driver in local institutions in terms of productivity, service quality, communication speed, effortlessness of access to information etc. Early in the 1990s, when the Internet began to unfold its potential, the German Government opted for a comprehensive IT agenda, i.e. for a systematic, coordinated approach to stimulate and implement innovative, Internet-based eGovernment solutions. Obviously, the time has come for trials with social media-like solutions within the realm of eGovernment.

However, the simplicity of the Open Government concept contrasts with its political, organizational and technological complexity. This document focuses on some basic technological aspects of Open Government and presents an open-source platform for Open Government solutions. This platform relies on software standards advocated by the German Government and will enable local governments to bridge the gap between local government and its citizens with an affordable, flexible and easy-to-customize tool.

2 Germany's Path from eGovernment to Open Government

When the Internet began to unfold its potential in the 1990s, the German Government decided to develop a comprehensive ICT agenda for Internet-based eGovernment. Initially, eGovernment was a fuzzy concept aimed at a more agile administration. The Federal Government soon initiated the Initiative Bund-Online2005, a project for a comprehensive modernization of federal services. Result: 350 different services in more than 100 units became "Internet-ready" with a financial investment of 1.65 billion euros [1].

Despite of the success of BundOnline 2005, Germany's eGovernment is not a European role model [2]. Consequently, in December 2010, representatives of politics, administration and industry opted for a fast expansion of eGovernment and Open Government, i.e. to ease interaction between governments and citizens, to strengthen local identity and to encourage new business models with local data [3]. Also in 2010, the IT Planning Council outlined the objectives for a national eGovernment strategy by addressing eight key areas and a number of targets [4]. The ambitious objective is to set international standards for an effective and efficient administration in a federal structure. Among others, one focus is on transparency and social participation. The Federal Government transformed this agenda for the modernization of government into precisely defined projects [5]. Three projects can be regarded as critical:

- An "Act on E-Government" to set the legal framework for secure electronic administration (electronic files, payment and access) and to encourage efficiency [6];
- The "Process Data Accelerator", aimed at methods, open standards and architectures for an electronic sharing of data between public administration and the private sector (10,000 reports/year, implying 50 billion euros in total/year) [7];

- An explicit Open Government strategy leading to more transparency, participation and innovation [3].

3 Social Media and Open Government

Today, we are witnessing the penetration of life by social media applications. Facebook, Twitter, LinkedIn, YouTube, Yammer, Socialcast – to name just a few – have changed the expectations of people in terms of a fast, less formal exchange of data and information, unrestricted by space and time. Companies like Nike or VW have reacted to this process. Marketing in the print media is reduced in favor of faster, more dynamic communication with potential customers on social media platforms. These companies are re-defining communication and carefully learning from experience where customer opinions have initiated a kind of information tsunami.

Citizens' expectations towards government are likely to move in the same direction. People who are used to evaluation and feedback on e-commerce platforms like amazon.com, booking.com and the like and are used to communicating with their friends on facebook expect similar options regarding access to government. Recent polls indicate clearly that citizens prefer a much greater openness: 71% of citizens expect greater satisfaction with their administration if there were additional options to contact local offices. Only 29% are currently satisfied with the present communication channels on the Internet [8]. For some time now, German governments have looked for concepts and technologies to respond to this new mindset. Local projects target open budget, open data and consultation management. However, the challenge is significant considering the persistence of public administration: it is about the acceleration of communication without a loss in quality, a more flexible responsibility of public employees and a new kind of liability in communication.

The social media paradigm opens up extraordinary opportunities, and it is helpful to reflect business experiences. It was Friedrich Krupp who strategically introduced a systematic approach to leverage the knowledge of employees. Now, 130 years later, an increasing number of companies experiment with social media – referred to as Social Enterprise or Enterprise 2.0 – to stimulate innovative ideas from their workforce. While companies use the knowledge of employees, local governments want to leverage the know-how of their citizens. Citizens are constitutive and the first and final authority when it comes to initiating new projects and the evaluation of past decisions. With social media technology at hand, citizens can get involved as "experts for the daily life" and become drivers for local innovations. Three different dimensions are essential:

- Transparency concerning political decisions-making;
- Direct participation in relevant issues by means of comments, suggestions and voting mechanisms;
- A fast way to communicate and collaborate.

What may initially sound abstract is in fact of tangible value. Only one example: In the city of Bonn, the local government used brochures, meetings and questionnaires to

identify critical topics related to budget decisions. On average, only 40 people per year contributed directly to a budget. In January 2010, the city of Bonn started an Open Budget project. Focused on budget-relevant activities in the sectors of sports and nature/environment protection, the population was encouraged to suggest and comment on investments. Using a web-based social media-like platform, more than 12,000 citizens registered on the website suggesting almost 1,500 approaches which stimulated 14,000 comments [9].

4 Open Government Requirements

Open Government has a number of pivotal objectives:

- To enable access to local data (socio-demographic, shopping patterns, traffic data etc.) and background information about community-related issues (information from the local council, etc.);
- To encourage suggestions and ideas concerning all kinds of local issues;
- To foster identification with the city, yielding a higher level of consensus;
- To support easy and flexible communication.

Without adequate software, the transfer of Open Government into real-world applications remains critical. A number of requirements are paramount:

- Interoperability with available systems, ERP and GIS software in particular;
- Open data that encourage new business models;
- Cross-linking of data (linked data; budget, traffic, pollution, demographic, economic etc.) and an automated, machine-assisted enrichment enhancement? process;
- Communication features to comment on statements, contribute ideas and vote for or against proposals;
- An architecture that is flexible enough to fit into new trends and experiences.

The overall technological concept should address communities of all sizes and follow the concept of “start small then grow”. In addition, the system should be easily customizable to suit local requirements, and it must be affordable [10].

5 Technical Enabler: SAGA and OpenSAGA

5.1 SAGA Standard

SAGA (Standards and Architectures for eGovernment Applications) is an ambitious standardization approach for Germany's public administrations, published formally by the German Government, in order to:

- Define technical standards and architectures for eGovernment applications;
- Standardize processes and data in administrations to achieve interoperability and compatibility [11 and 12].

In technical terms, SAGA concentrates on five fundamental objectives: interoperability, reusability, openness [13], reduction of costs and risks, and scalability.

The Federal Ministry of the Interior released SAGA version 5.0 at the end of 2001 with a refined collection of methods, specifications and software tools. Whenever reasonable, ICT solutions for governments should be based on these modules.

5.2 OpenSAGA Platform

The basic concept of OpenSAGA is outlined under [14]. The mission of the OpenSAGA platform – officially released in May 2010 under the GPL V2 license – is to deliver an open-source framework for developing SAGA-compliant, Java-based web applications (for instance for a food emergency warning, Fig. 1). The innovative idea is to generate SAGA-compliant applications for eGovernment from domain descriptions aiming at an 80% automation level while programming an additional 20% for more complex business logic.

Four characteristics are constitutive for OpenSAGA:

- It describes domain interrelationships (“what?"), while the appropriate technical implementation ("how?") is generated (if possible) automatically;
- It makes reasonable assumptions, especially in terms of SAGA compliance; i.e. software developers do not have to take care of SAGA technologies;
- It concentrates on semantic relations, i.e. programmers can concentrate their work on business logic instead of a large number of technical details;
- It follows a model-driven approach.

Fig. 1. Food emergency warning: Generated with OpenSAGA [15].

In architectural terms, OpenSAGA consists of four elements:

- The domain models describe the domain;
- The process model describes the behavior of the application and its reaction to user input;

- The view model describes the general structure and content of a view state within a process;
- The generator creates a runtime model from the models' XML files.

With its strong reference to SAGA, OpenSAGA is obviously the first-choice platform for developing Open Government applications, and an Open Government suite.

6 The Open Government Suite

The Open Government suite (OGS) is based on OpenSAGA and exploits this platform's innovative features [16]. Following the architectural capabilities of OpenSAGA, the OGS is a modular, integrating, and configurable solution (Fig. 2) characterized by

- An integrated, state-of-the-art user interface;
- A single sign-on for all modules;
- A homogeneous concept for the administration;
- Functionalities for elective elements and the evaluation of ideas.

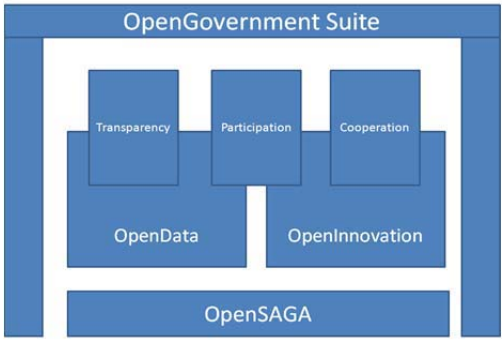


Fig. 2. The Open Government Suite (basic scheme)

Currently, the OGS features three core modules: the data catalog (open data), budget participation, and collaboration (Fig. 3). Additional modules may be integrated during the evolution of the OGS (council information module, business process modeling, reporting etc.). The overall advantages of the OGS are controllability, openness to extensions, easy implementation and a "start small, and grow according to experiences and needs" approach.

6.1 Open Data

Open Data is currently the most evident European Open Government topic [17and 18]. The OGS's Open Data module enables a well-structured compilation of freely available local data. Users get a straightforward overview of available data, may sub-

mit suggestions to the administration to add specific data to the pool or add relevant information (like links, assessments, Meta data etc.).

Specific characteristics of the OGS:

- Simple uploading of data following the REST approach;
- A variety of search filters.

Fig. 3. The Open Government Suite: Home Page

6.2 Budget Participation

A municipality's budget is a political program defined in terms of financial figures. An increasing number of German cities have started to experiment in this field by explicitly encouraging citizens to comment on carefully selected topics [19 and 20].

Specific characteristics of the OGS (Fig. 4):

- Freely configurable categories (i.e. traffic, kinder garden, schools);
- Filter concept to follow monitor? selected topics;
- Easy to contribute to discussions;
- Configurable voting procedures;
- Statistical functionalities;
- Transfer of master data to the following year.

Fig. 4. Citizens may contribute to selected budget issues (screenshot of the OGS, in German)

6.3 Collaboration

The OGS comprises a collaboration tool introduced as an event stream. Users may enroll and contribute to a topic. Whenever something "interesting" occurs on the platform (new information or data sets, new evaluations, votes completed, etc.), participants will be informed. Analogous to modern social media systems, users may feed messages into that event stream at any time.

Specific characteristics of the OGS:

- Clustering of information into “streams”;
- Usability follows well-known standards for social media;
- Automatic information about new entries;
- News-based information between participants;
- Tagging.

Summary

Governments face a considerable challenge, and the pressure to deal with the many problems grows every year. With the advent of social media, an increasing number of German local governments consider the inclusion of their citizens in a large number of local affairs. Referred to as Open Government, the focus is on transparency, participation, and collaboration.

For some time, the German Government has propagated SAGA as an open, interoperable and sustainable standard for government applications. OpenSAGA is a SAGA-compliant platform to develop Java-based web solutions. Almost naturally, the OpenGovernment Suite is based on OpenSAGA and provides a comprehensive choice for communities to move forward on the path to much closer cooperation between citizens and local governments.

However realistic this perspective may be: This is only the first of many steps forward. New ideas like cloud computing and mobile solutions supplement the picture,

posing a constant need for additional ICT concepts. Creative initiatives in terms of open data, for instance in the German capital Berlin, are beginning to take shape, and the German Government has initiated lead projects with respect to cloud computing (also in the sector of eGovernment), the outcome of which will have to be evaluated carefully in the years to come.

References

1. Zypries, B. (2006): BundOnline 2005 - die nächsten Schritte der eGovernment-Initiative des Bundes, in: Schubert, S., Reusch, B., Jesse, N., „Informatik bewegt“, 32. Jahrestagung der Gesellschaft für Informatik e.V. (GI), Lecture Notes in Informatics, Procedings, Part II p. 29 ff.
2. Altmeier, P. (2012): Nicht einmal im Ansatz erkannt. Veränderungen administrativer und legislativer Willensbildung durch Web 2.0, in: Behörden Spiegel, Mai 2012.
3. 5th National IT Summit, “The Dresden Agreement – Utilising the opportunities for ICT in Germany”, Dresden, 7 December 2010
4. IT-Planungsrat (2010): National E-Government Strategy, IT Planning Council decision. Sept. 24, 2010.
5. The Federal Government, Government Programme „Network-Based and Transparent Administration“, Federal Ministry of the Interior, September 2010
6. Referentenentwurf der Bundesregierung, zum Gesetz zur Förderung der elektronischen Verwaltung sowie zur Änderung weiterer Vorschriften (2012), Stand 05.03.2012
7. http://www.p23r.de/
8. Stemper, J.; Schulz-Dieterich, A. (2012): Erfolgsfaktoren für die Verbreitung von E-Partizipation. Ergebnisse einer bundesweiten Studie zur digitalen Beteiligung. in: innovative Verwaltung, 1-2/2012, S. 14 ff.
9. Stadt Bonn. Rechenschaftsbericht „Bürgerbeteiligung am Haushalt 2011/2012
10. Jesse, N. (2012-2): Smarter Cities with the OpenGovernment Suite – A German IT-Platform for a Dialog between Citizens and Local Governments, in: International Conference on Management and Service Science (MASS 2012), 10.-12.08, 2012, Shanghai, China. (accepted)
11. Der Beauftragte der Bundesregierung für Informationstechnik (2011): Konzept für SAGA 5.0, Version de.bund 5.0.0, 3.11.2011
12. Federal Minister of the Interior (2008): SAGA. Standards and Architectures for eGovernment Applications. Version 4.0, Bonn 2008
13. Jesse, N.: Increasing Germany’s Government Performance with Open Source Software? – From Strategy to Implementation, in: International Conference on e-Commerce, e-Administration, e-Society, e-Education, and e-Technology, Hong Kong 2012
14. http://www.opensaga.org
15. http:/www.lebensmittelwarnung.de
16. http://www.opengovernmentsuite.de
17. Schellong, A., Stepnanets, E. (2011): Unbekannte Gewässer. Zum Stand von Open Data in Europa. Studie der Computer Science Corp. (CSC), 2011
18. Shadbolt, N. (2010): Towards a pan EU data portal - data.gov.eu, http://ec.europa.eu/information_society/policy/psi/docs/pdfs/towards_an_eu_psi_portals_v4_final.pdf, 15th December 2010, Version 4
19. http://www.buergerhaushalt.org/

20. Von Lucke, J., Geiger, Ch. P., Hoose, A., Schreiner, M. (2011): Open Government Data. Budget 2.0 & Open Budget Data. Öffnung von Haushaltswesen und Haushaltsdaten. Gutachten für die Deutsche Telekom AG zur T-City Friedrichshafen. Dated 24 October 2011

A Comparative Review of Contention-Aware Scheduling Algorithms to Avoid Contention in Multicore Systems

Genti Daci, Megi Tartari

Abstract. Contention for shared resources on multicore processors is an emerging issue of great concern, as it affects directly performance of multicore CPU systems. In this regard, Contention-Aware scheduling algorithms provide a convenient and promising solution, aiming to reduce contention, by applying different thread migration policies to the CPU cores. The main problem faced by latest research when applying these schedulers in different multicore systems, was a significant variation of performance achieved on different system architectures. We aim to review and discuss the main reasons of such variance arguing that most of the scheduling solutions were designed based on the assumption that the underlying system was UMA (Uniform Memory Access latency, single memory controller), but modern multicore systems are NUMA (Non Uniform Memory Access latencies, multiple memory controllers). This paper focuses on reviewing the challenges on solving the contention problem for both types of system architectures. In this paper, we also provide a comparative evaluation of the solutions applicable to UMA systems which are the most extensively studied today, discussing their features, strengths and weaknesses. For addressing performance variations, we will review Vector Balancing, OBS-X and DIO scheduling for UMA systems. While for NUMA systems, we will compare and discuss DINO and AMPS Schedulers which supports NUMA architectures aiming to resolve performance issues and also introduce the problems they have. This paper aims to propose further improvements to these algorithms aiming to solve more efficiently the contention problem, considering that performance-asymmetric architectures may provide a cost-effective solution.

Keywords: Uniform Memory Access(UMA), Multicore CPU systems, Contention-Aware Scheduling, Non Uniform Memory Access(NUMA), Vector Balancing Scheduling, OBS-X Scheduler, DIO Scheduler, DINO Scheduler, AMPS Scheduler.

1 Introduction

Contention for shared resources in multicore processors is a well-known problem. The importance of handling this problem is related with the fact that multicore processors are becoming so prevalent in desktops and also servers, that may be considered a standard for modern computer systems and also with the fact that this problem causes performance degradation. Let's consider a typical multicore system

S. Markovski, M. Gusev (Editors): ICT Innovations 2012, Web Proceedings, ISSN 1857-7288

described schematically in Figure 1, where cores share parts of memory hierarchy, that we call "memory domains", and compete for resources like last level cache (LLC), memory controllers, memory bus and prefetching hardware.

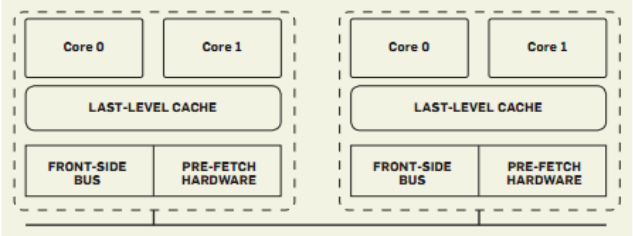


Fig. 1. A schematic view of a multicore system with 2 memory domains

Preliminary studies considered cache contention [5] as the most crucial factor responsible for performance degradation. Driven by this assumption, they focused on finding mechanisms to reduce cache contention like Utility Page Partitioning [8] and Page Coloring [7]. Successive studies [1] calculated the contribution that each of the shared resources in multicore processors have in degrading performance of such systems, concluding that contention for last level cache (LLC) was not the dominant factor in degrading performance. Based on this new conclusion, recent studies selected scheduling as an attractive tool, as it does not require extra hardware and it is relatively easy to integrate into the system. Contention-Aware scheduling [1][2][3][4] is proposed as a promising solution to this problem, because it reduces contention, by applying different thread migration policies. The major part of these studies, found solutions that could be applied only in UMA (Uniform Memory Access) systems, that are not suitable for NUMA (Non Uniform Memory Access). So for UMA systems we will discuss DIO Scheduler, that uses thread classification schemes like SDC [9], LLC miss rate [2], Pain metric [2], Animal Classes [10] to take the best scheduling decision; OBS-X scheduling policy based on the data provided by the OS dynamic observation of tasks behavior; Vector Balancing scheduling policy, that reduces contention for shared resources by migrating tasks based on the task activity vector information, that characterizes tasks regarding resource usage. For NUMA architecture, that still requires further research, is proposed DINO Scheduler. We will also discuss AMPS scheduler design for asymmetric-architecture multicore systems, that supports NUMA.

The rest of the paper is organized as follows: In Section 2 we argument why contention-aware algorithm is considered a promising solution to mitigating contention. In Section 3 we review and discuss the scheduling algorithms valid for UMA systems. In Section 4 we review and discuss the scheduling solutions proposed for NUMA architectures and we conclude in Section 5.

2 Contention-Aware Scheduling a Promising Solution

Preliminary studies on improving thread performance in multicore systems were mainly focused on the problem of contention for the shared cache. Cache partitioning has a significant influence on performance closely relating with execution time. J. Lin, Q. Lu, X. Ding, Z. Zhang, X. Zhang, and P. Sadayappan [5], implemented an efficient layer for cache partitioning and sharing in the operating system through virtual-physical address mapping. Their experiments showed a considerable increase of performance up to 47 %, in the major part of selected workloads. A number of cache partitioning methods have been proposed with performance objectives [7] [8] [25]. A. Fedorova, M. I. Seltzer, M. D. Smith [17] designed a cache-aware scheduler that compensates threads that were hurt by cache contention by giving them extra CPU time.

The difficulty faced from S. Zhuravlev, S. Blagodurov and A. Fedorova [2] in evaluating the contribution that each factor has on performance was that all the degradation factors work in conjunction with each other in complicated and practically inseparable ways.

To take into consideration the result of their work, it is proposed Contention-Aware Scheduling, that separates competing threads onto separate memory hierarchy domains to eliminate resource sharing and, as a consequence to mitigate contention. To design a contention-aware scheduler, initially we must choose a thread classification scheme, that predicts how they will affect each other when they will compete for shared resources and a scheduling policy, which assigns threads to cores given their classification. So the classification scheme serves to identify applications that must be co-scheduled or not. S. Zhuravlev, S. Blagodurov and A. Fedorova [2] help us with their contribution in analyzing the effectiveness of different classification schemes like:

- SDC (Stack Distance Competition), a well known method [9] for predicting the effects of cache contention among threads, based on the data provided from stack distance profiles, that inform us on the rate of memory reuse of the applications.
- Animal Classes is based on the animalistic classification of application introduced by Y. Xie and G. Loh [10]. It allows classifying applications in terms of their influence on each other when co-scheduled in the same shared cache.
- Miss Rate is considered as the heuristic for contention, because it gives information for all the shared resources contention.
- Pain Metric is based on *cache sensitivity* and *cache intensity*, where sensitivity is a measure of how much an application will suffer when cache space is taken away from it due to contention; intensity is a measure of how much an application will hurt others by taking away their space in a shared cache.

The results of evaluation of effectiveness of these classification schemes, show that the best contention predictor is Miss Rate [2][11]. A high miss rate exacerbates the contention for all of these resources, since a high-miss-rate application will issue a large number of requests to a memory controller and the memory bus, and will also be typically characterized by a large number of prefetch requests, while SDC performed

worse because it does not take into account miss rates in its stack distance competition model and it models the performance effects of cache contention, which is not the only cause of degradation.

As a perfect scheduling policy, it is used an algorithm proposed by Y. Jiang, X. Shen, J. Chen, R. Tripathi [16].This algorithm is guaranteed to find an optimal scheduling assignment, i.e., the mapping of threads to cores, on a machine with several clusters of cores sharing a cache as long as the co-run degradations for applications are known. Jiang's methodology uses the co-run degradations to construct a graph theoretic representation of the problem. The optimal scheduling assignment can be found by solving a min-weight perfect matching-problem.

3 Proposed Schedulers for UMA Systems

Several studies investigated ways of reducing resource contention and as mentioned above in Section2, one of the promising approaches that emerged recently is contention-aware scheduling [2][3][4]. This represents a promising solution, as several research co-scheduled tasks based on memory bandwidth or other shared resources. We mention here the co-scheduling tasks proposed for SMP [19,22] and for SMT [20]. These studies of contention-aware algorithms were focused primarily on UMA (Uniform Access Memory) systems, where there are multiple shared LLCs, but only a single memory node equipped with a single memory controller, and memory can be accessed with the same latency for any core. In this section we will review and evaluate OBS-X, Vector Balancing scheduling policy, and DIO scheduler by discussing their features, merits, but also their gaps.

3.1 OBS-X Scheduling Policy based on OS Dynamic Observations

According to R. Knauerhase, P. Brett, B. Hohlt, and S. Hahn [3], in a multicore environment the Operating System (OS) can and should make observations of the behavior of threads running in the system. These observations, combined with knowledge of the processor architecture, allow the implementation of different scheduling policies in the Operating System. Good policies can improve the overall performance of the system or performance of the application.

The performed experiments on this study have included various software and hardware environments. The lack of intelligent thread migration and also the fact that OS handles cores as independent, without taking into account that they share resources represent the challenges faced by R. Knauerhase, P. Brett, B. Hohlt, and S. Hahn during this study, where they found a policy to address these challenges, as the traditional operating system scheduler does not take into account the fact that amount of contention is quite dynamic because it depends on each task's behavior at a given time. After analyzing this study, we faced a problem with the authors.

They developed an observation subsystem that collects historical and hysteretic data by inspecting performance-monitoring counters and kernel data structures, gathering so information on a per-thread basis. They introduced OBS-X scheduling policy, that uses observations of each task's cache usage. OBS-X's goal is to distribute cache-heavy threads throughout the system, helping so to spread out cache load. When a new task is created, OBS-X looks for the LLC group with the smallest cache load, and places the new task in this group. OBS-X strength relates with the fact that this policy include the notion of overloaded tasks.

They ran two sets of experiments across four cores in two LLC groups. The first set of experiments consisted of four instances of cachebuster, an application that consumes as much cache as possible and four instances of spinloop, that consumes CPU with a minimum of memory access. They used [cb,sl][cb,sl] pairing, which represents the worst performance because both cachebuster applications contend for cache resources at the same time. With the addition of OBS-X, cachebuster performance increased between 12 % and 62 %, comparing with the default Linux default load balancing. The reason for the increase is that OBS-X distributed the cache-heavy tasks across LLC groups, thus minimizing the scheduling of heavy tasks together. To approximate real-world workloads, they ran OBS-X with a set of applications from the SPEC CPU 2000 suite run. The overall speedup increases to 4.6 %.

3.2 Vector Balancing Scheduling Policy

This policy reduces contention by migrating tasks, led by the information of *task activity vector* [18], that represents the utilization of chip resources caused by tasks. Based on the information provided form these vectors, it has been proposed from A. Merkel, J. Stoess and F. Bellosa [4] the scheduling policy that avoids contention for resources by co-scheduling tasks with different characteristics. The definition of activity vectors requires the read of a small number of the performance-monitoring counters (PMC) and asymmetric observations. This policy can be easily integrated in the OS balancing policy, so we can exploit the existing strategies. The weakness of this proposed solution by A. Merkel, J. Stoess and F. Bellosa [4] lies in the fact that these authors to avoid complexity in their research, assumed that tasks do little I/O, do not communicate with each other, they are independent. They used compute-intensive tasks. This assumption is a weakness because it limits the space where the Vector Balancing can be applied successfully. If there is communication, co-scheduling based on resource utilization can have conflicting goals. This is a topic of future work.

3.3 DIO (Distributed Intensity Online) Scheduler

S. Zhuravlev, S. Blagodurov and A. Fedorova [2] proposed DIO contention-aware scheduler. DIO scheduler continuously monitors the miss rates of applications, as we

argued in Sector 2 that it was the best contention predictor, then finds the best performance case and separates threads. It obtains the miss rates of applications dynamically online via performance counters. This makes DIO more attractive since the stack distance profiles, which require extra work to obtain online, are not required. Furthermore, the dynamic nature of the obtained miss rates makes DIO more flexible to application that have a change in the miss rate due to LLC contention. DIO was experimented in AMD Opteron with 8 cores, 4 cores for each domain. DIO improved performance by up to 13 % relative to default. Another use of DIO is to ensure QoS (Quality of Service) for critical applications, since it ensures to never select the worst performance case of the scheduler.

4 Adaptation of Contention-Aware Schedulers for NUMA Systems

Research studies on contention-aware algorithms, were primarily focused on UMA (Uniform Memory Access) systems, where there are multiple shared last level caches (LLC), but they have only one memory node associated with a memory controller, and the memory can be accessed with the same latency from every core. Modern multicore systems are using massively the NUMA (Non Uniform Memory Access) architecture, because of its decentralized and scalable nature. In these systems there is one memory node for each memory domain. Local nodes can be accessed for a shorter time than the distant ones, and each node has its own controller. According to S. Blagodurov, S. Zhuravlev, M. Dashti and A. Fedorova [1], when existing contention-aware schedulers designed for UMA architectures, were applied on a NUMA system (illustrated on Figure 3 [1]), they did not effectively manage contention, but they also degraded performance compared with the default contention-unaware scheduler (30% performance degradation).

4.1 Why existing Contention Management Algorithms degrade Performance on NUMA Systems?

S. Zhuravlev, S. Blagodurov and A. Fedorova [2] proposed DIO contention-aware scheduler. DIO scheduler continuously monitors the miss rates of applications, as we argued in Sector 2 that it was the best contention predictor, then finds the best performance case and separates threads. It obtains the miss rates of applications dynamically online via performance counters. This makes DIO more attractive since the stack distance profiles, which require extra work to obtain online, are not required. Furthermore, the dynamic nature of the obtained miss rates makes DIO more flexible to application that have a change in the miss rate due to LLC contention. DIO was experimented in AMD Opteron with 8 cores, 4 cores for each domain. DIO improved performance by up to 13 % relative to default. Another use of DIO is to ensure QoS

(Quality of Service) for critical applications, since it ensures to never select the worst performance case of the scheduler.

4.2 DINO Contention-Management Algorithm for NUMA Systems

As argued above, previous contention-aware algorithms were valid only on UMA architectures, but when applied to NUMA architectures, used in today's modern multicore processors hurt their performance. To address this problem, a contention-aware algorithm on a NUMA system must migrate the memory of the thread to the same domain where it migrates the thread itself. However, the need to move memory along with the thread makes thread migrations costly. So the algorithm must minimize thread migrations, performing them only when they are likely to significantly increase performance, and when migrating memory it must carefully decide which pages are most profitable to migrate. These are the challenges of designing a new contention-aware scheduling algorithm, which is appropriate with NUMA architecture. These challenges are handled in the study of S. Blagodurov, S. Zhuravlev, M. Dashti and A. Fedorova [1]. They have designed and implemented Distributed Intensity NUMA Online (DINO).

DINO scheduler uses the same heuristic model for contention as the DIO (Distributed Intensity Online) scheduler discussed in Section 3.3, that uses the *LLC miss rate* criteria for predicting contention. First of all, DINO tries to co-schedule threads of the same application on the same memory domain, provided that this does not conflict with DINO's contention-aware assignment. This is true for many applications [14]. DINO organizes threads in broad classes according to their miss rates as shown in the research study of Y. Xie and G. Loh [10]. The classes in which threads get divided are:

- Turtles: less than 2 LLC miss rates for 1000 instructions
- Devils: 2-100 LLC misses for 1000 instructions
- Super_Devils: more than 100 LLC misses for 1000 instructions

So the migrations will be performed only when threads change their classes, while they preserve their thread-core affinity relation as much as possible. For multithreaded applications DINO tries to co-schedule threads of the same application, in the same memory domain, but always avoiding to create conflicts in DINO's definitions regarding contention management. It also uses techniques to evaluate if it is convenient to co-schedule threads in the same domain or it would be better to separate them? DINO in this situation should at least avoid memory migration back and forth, preventing so performance degradation. DINO achieves this by separating threads in classes as explained above.

Results of DINO implementation showed that DINO achieved up to 30 % performance improvements for jobs in the MPI workload.

4.3 AMPS the Scheduling Algorithm for Performance-Asymmetric Multicore System NUMA & SMP

Since industry is going towards multicore technology, and traditional operating systems are based on homogenous hardware, and performance-asymmetric architectures (or heterogeneous) [21][23], present a very convenient solution regarding the cost they have, it appears the necessity to setup the relation between two different technologies. As a first step towards this, T. Li, D. Baumberger, D. A. Koufaty and S. Hahn [6] designed the operating system scheduler AMPS, that manages efficiently both SMP and NUMA-style performance-asymmetric-architectures. AMPS contains three components:
- Asymmetry-aware-load-balancing, that balances threads to cores in proportion with their computing power
- Faster-core-first scheduling, that controls thread migrations based on predictions of their overhead.

Our evaluation demonstrated that AMPS improved stock Linux for asymmetric systems in the aspect of performance and fairness.

AMPS uses thread-independent policies, which schedule threads independently regardless of application types and dependencies. This is considered a weakness that should be eliminated in the future. Thread-dependent policies mostly exists in research. H. Zheng, J. Nieh [24] dynamically detect process dependencies to guide scheduling.

5 Related Works

Research on solutions for the problem of resource contention on multicore systems is wide and dates back many years. Initial research in this field, were based on the idea that the primary factor on degradation performance in such systems was contention for shared cache. We mention the study of J. Lin, Q. Lu, X. Ding, Z. Zhang, X. Zhang, P. Sadayappan [5] who evaluated the impact of existing cache partitioning methods on multicore system performance. They observed with most workloads, a significant performance improvement (up to 47 %). The only limitation of this study was that their experiments were limited by the hardware platform they used.

S. Zhuravlev, S. Blagodurov, A. Fedorova [2] through extensive experimentation on real systems, determined that along with it, other factors like memory controller contention, memory bus contention and prefetching hardware contention all combine in complex ways to create the performance degradation. They proposed DIO which improved performance by up to 13 % relative to the default operating system scheduler. Prior to DIO, were proposed also other scheduling policies like OBS-X [3], which uses the operating systems observations of behavior of threads running in the systems and then makes a decision on how to migrate threads for a better performance; All these studies were primarily focused on UMA systems, while

DINO contention-aware scheduler, remains the most appropriate until today for NUMA and uses miss rate as a contention predictor, like DIO does.

Prior research demonstrated that compared to homogeneous ones, asymmetric architectures deliver higher performance at lower costs in terms of die area and power consumption. T. Li, D. Baumberger, D. A. Koufaty, S. Hahn [6] proposed AMPS scheduler that manages efficiently both SMP and NUMA-style performance-asymmetric architectures. The problem of contention of heterogeneous architectures is almost uncovered, that is why it is a field of future research.

6 Conclusions and Discussions

Based on the wide dissemination of multicore processors, we chose to handle the topic of contention for shared resources in such systems, as it affects directly their performance. One of the major difficulty encountered during design of such schedulers, was selecting the most effective thread classification scheme, used to choose the best performance case respective to a specific pairing of co-scheduled threads. To mitigate contention for shared resources, we discussed and reviewed the best scheduling algorithms and policies, that do not perform equally when applied to different multicore architectures. So for UMA systems, we reviewed OBS-X scheduling policy, that uses the operating system dynamic observations on tasks behavior to migrate threads; Vector Balancing scheduling that takes migration decisions based on the task activity vector information and DIO contention-aware scheduling which is the best solution for UMA, because it mitigates contention for all shared resources, not only for cache contention, as OBS-X does. Moreover, Vector Balancing provides a limited solution, as it is based on compute-intensive and independent tasks, that do little I/O. These previously proposed contention-aware scheduling policies applied to NUMA modern multicore systems proved to hurt these systems' performance, because they fail to eliminate memory controller contention and create additional interconnect contention, that is why they needed adaptation to this new architecture. The most appropriate solution for NUMA systems is the DINO contention-aware scheduler, as it solves the performance degradation problem associated with the previous contention-aware solutions by migrating the thread along with its memory and also eliminates superfluous migrations. AMPS is the first scheduler proposed for the performance-asymmetric architectures, that supports both NUMA and SMP-style performance-asymmetric architectures, but it does not completely address contention, requiring further research in the future.

References

1. Blagodurov, S., Zhuravlev, S., Dashti, M., Fedorova, A.: A Case for NUMA-aware Contention Management on Multicore Systems. In: The 2011 USENIX Annual Technical Conference, pp. 1-9 (2011).

2. Zhuravlev, S., Blagodurov, S., Fedorova, A.: Addressing Contention on Multicore Processors via Scheduling. In: Proceedings of ASPLOS, pp.1-6 (2010).
3. Knauerhase, R., Brett, P., Hohlt, B., Hahn, S.: Using OS Observations to Improve Performance in Multicore Systems. In: IEEE Micro 28, 3 , pp. 54-58 (2008).
4. Merkel, A., Stoess, J., Bellosa, F. : Resource-Conscious Scheduling for Energy Efficiency on Multicore Processors. In: Proceedings of EuroSys, pp.6-8, pp.11-13 (2010).
5. Lin, J., Lu, Q., Ding, X., Zhang, Z., Zhang, X., Sadayappan, P.: Gaining Insights into Multicore Cache Partitioning: Bridging the Gap between Simulation and Real Systems. In: Proceedings of International Symposium on High Performance Computer Architecture, pp. 1-5 (2008).
6. Li, T., Baumberger, D., Koufaty, D.A., Hahn, S.: Efficient Operating System Scheduling for Performance- Asymmetric Multi-core Architectures. In: Proceedings of Supercomputing, pp.1-4, pp.8-10 (2007).
7. Zhang, X., Dwarkadas, S., Shen, K.: Towards practical page coloring-based multicore cache management. In: Proceedings of the 4th ACM European Conference on Computer Systems 2009.
8. Qureshi, M. K., Patt, Y. N.: Utility-based cache partitioning: A low overhead, high-performance, runtime mechanism to partition shared caches. In MICRO 39: Proceedings of the 39th Annual IEEE/ACM International Symposium on Microarchitecture , pp. 1-3 (2006).
9. Chandra, D., Guo, F., Kim, S., Solihin, Y. : Predicting InterThread Cache Contention on a Chip Multi-Processor Architecture. In HPCA '05: Proceedings of the 11th International Symposium on High Performance Computer Architecture (2005).
10. Xie, Y., Loh, G.: Dynamic Classification of Program Memory Behaviors in CMPs. In: Proceeding of CMP-MSI, pp. 2-4 (2008).
11. Blagodurov, S., Zhuravlev, S., Fedorova, A.: Contention-aware Scheduling on Multicore Systems. ACM Trans. Comput. Syst. 28 (December 2010).
14. Zhang, E. Z., Jiang, Y., Shen, X.: Does Cache Sharing on Modern CMP Matter to the Performance of Contemporary Multithreaded Programs? In: Proceedings of PPOPP (2010).
16. Jiang, Y., Shen, X., Chen, J., Tripathi, R.: Analysis and Approximation of Optimal Co-Scheduling on Chip Multiprocessors. In: Proceedings of the 17th International Conference on Parallel Architectures and Compilation Techniques (PACT '08), pp. 220-229 (2008).
17. Fedorova, A., Seltzer, M.I, Smith, M.D.: Improving Performance Isolation on Chip Multiprocessors via an Operating System Scheduler. In: Proceedings of the Sixteenth International Conference on Parallel Architectures and Compilation Techniques (PACT'07), pp.25-38 (2007).
19. Zhang, X., Dwarkadas, S., Folkmanis, G., Shen, K.: Processor Hardware Counter Statistics as a First-Class System Resource. In: Proceedings of the 11th USENIX workshop on Hot topics in operating systems (HOTOS'07).
20. McGregor, R. L., Antonopoulos, C. D., Nikolopoulos D. S.: Scheduling Algorithms for Effective Thread Pairing on Irbid Mutiprocessors. In: Proceedings of the 19th IEEE International Parallel and Distributed Processing Symposium (IPDPS'05).
21. Shelepov, D., Saez Alcaide, J.C., Jefferym S., Fedorova, A., Perez, N., Huang, Z. F., Blagodurov, S., Kumar, V.: A Scheduler for Heterogeneous Multicore Systems. In: SIGOPS Operating Review,43(2) (2009).
22. Antonopoulos, C., Nikolopoulos, D., Papatheodorou, T.: Scheduling Algorithms with Bus Bandwidth Considerations for SMPs. In: International Conference on Parallel Processing (October 2003).

23. Balakrishnan, S., Rajwar, R., Upton, M., Lai, K.: The Impact of Performance Asymmetry in Emerging Multicore Architectures. In: Proceedings of the 32th Annual International Symposium on Computer Architecture, pp. 506-517 (June 2005).
24. Zheng, H., Nieh, J.: A Scheduler with Automatic Process Dependency Detection. In: Proceedings of the First Symposium on Networked Systems Design and Implementation, pp. 183-196 (March 2004).
25. Suh, G. E., Devadas, S., Rudolph, L.: A New Memory Monitoring Scheme for Memory-Aware Scheduling and Partitioning. In: Proceedings HPCA'02, pp.117-128 (2002).

S. Markovski, M. Gusev (Editors): ICT Innovations 2012, Web Proceedings, ISSN 1857-7288

Cloud e-University services

Kiril Kiroski, Marjan Gusev, Magdalena Kostoska, and Sasko Ristov

University Ss Cyril and Methodius, Faculty of Computer Sciences and Engineering, Rugjer Boshkovikj 16, Skopje, Macedonia,
{kiril.kjiroski, marjan.gushev, magdalena.kostoska, sashko.ristov}@finki.ukim.mk

Abstract. This paper examines and elaborates various electronic services used at universities on a world-wide level. Cloud e-University services provide learning help as well as practical knowledge for students, and enable teaching and administrative staff to fulfill their tasks through provided services and integrate various solutions into greater, encompassing platform for e-driven education. Electronic services include virtualization and Cloud services, computer clusters, GRIDs and storage, science and engineering software primarily used for development, video conferencing and distance learning services, social computing and support services, as well as Student Information Systems (SIS) and Student Lifecycle Management Systems (SLM). We give brief overview of how iKnow [1] functionalities can be mapped into ELF framework [2].

Keywords: e-University, cloud computing, SaaS, electronic services, E-Learning

1 Introduction

Higher education institutions are amongst the few organizations facilitating development and deployment of new technologies. Cloud computing today is becoming "old news", but this status is reached through no small effort and support of Universities around the world. Their role in Cloud computing development varies, from initial designers and developers, through consulting and participation in cloud services testing, and consumers of end-user solutions on the other end of the spectrum. Therefore, we can safely say that Universities today are more and more becoming e-Universities, given the amount of electronic services developed, deployed, and consumed.

Universities, as leading researchers and innovators of new technologies have eagerly adopted benefits provided through Cloud computing, and developed various means to exploit these virtual resources.

S. Markovski, M. Gusev (Editors): ICT Innovations 2012, Web Proceedings, ISSN 1857-7288

For implementing a complete e-University system there are plenty e-Learning Framework variations developed throughout the years. On Figure 1, we present working e-Learning Framework developed through coordinated efforts of U.K Joint Information Services Committee (JISC) and Australia Department of Education, Science and Training (DEST) [3]. This Framework recognizes 59 distinct functionalities divided into three main groups: Sample User Agents, Learning Domain Services and Common Services.

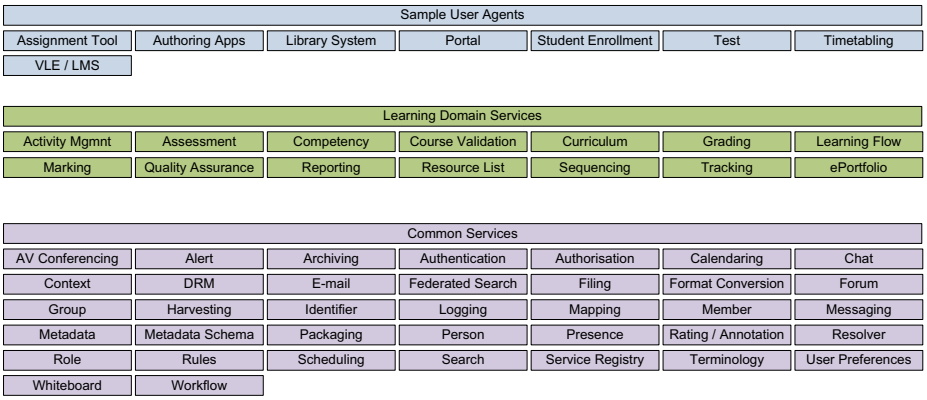
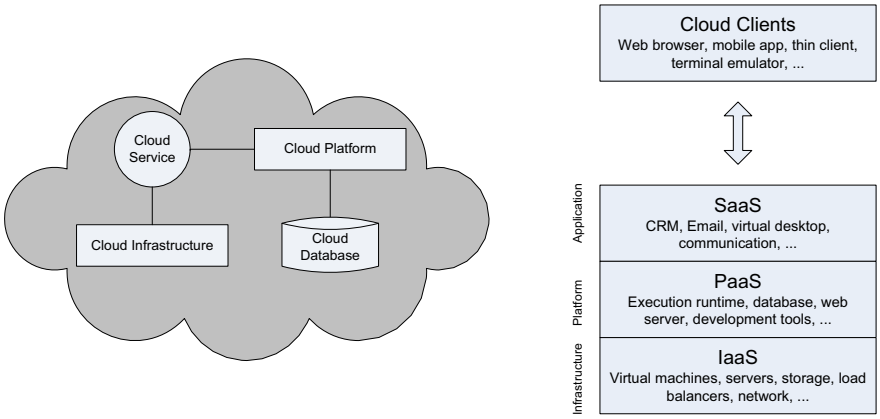


Fig. 1 The E-Learning Framework.

2 Cloud implementation of University student services

According to [4] there are "…[besides Service Oriented Architecture] two other common ways of integrating systems, which are to integrate at the user interface level using portals, or at the data level by creating large combined datasets or data warehouses." Out of these words, one can clearly see why we propose Service Oriented Architecture as building and Web Services as delivering model for University-wide integrated and maintained electronic services solution. Both of the alternative approaches require highly trained IT professionals, data warehouses, servers and infrastructure than any faculty possesses on its own.

A sample Cloud architecture and cloud computing service levels are shown in Figure 2. Figure 2(a) shows simple cloud architecture, where data from database residing in the cloud is providing cloud service using cloud based platform. All of these components are using cloud infrastructure, typically a Virtual machine, and are connected using loose coupling. Figure 2(b) illustrates 3 service levels in cloud computing, namely, Infrastructure as a Service, Platform as a Service and Software as a Service. Cloud clients use provided service through appropriate application.



a) b)

Fig. 2 (a) Cloud architecture, and (b) cloud computing service levels.

According to objectives stated in [5], "…university information system should automate the overall business processes of the university. … organize student data, employees, study programs, various data about the educational process and the faculty's research and scientific work. … automate the work of the student service, human resources management, schedule of lectures, management of student payments on various grounds, distribution of resources (professors and classrooms), organization of online services for the students and various other activities…"

According to [6], "…Cross-institutional cooperation [in Macedonia] by sharing information is a need which arises, because many courses [at Universities] are beginning to be taught collaboratively realizing the concepts of student mobility and lifelong learning…". Therefore, realizing iKnow system as cloud service(s), should become one of the most important goals for the next period, since interoperability and standardization of data and services is one of the benefits when using cloud infrastructure.

For example University Sts Cyril and Methodius, started the iKnow project [1] to establish a central placeholder for ideas, requirements, functionality descriptions and activities toward designing and developing Student Information Services and Student Enrolling platform. This solution consists of five modules, and is intended to implement 76 identified basic functionalities. Most of functionalities are specific only to one module, but there is a number of functionalities shared between modules, where Core module is involved in all of those situations.

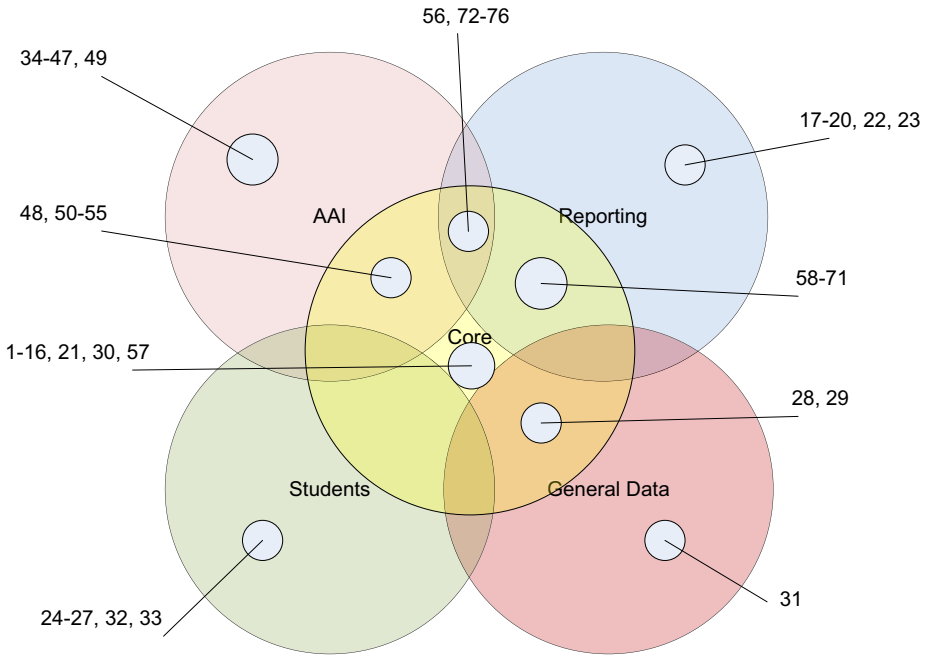


Fig. 3 iKnow system defines 5 modules and 76 basic functionalities [1].

Our goal is to provide insight into adjusting these services for cloud computing. In this paper we will not provide detailed description of needs and means to transform each and every service into cloud services. Rather, we will provide some suggestions how to divide responsibilities on service clusters, and general directions how to proceed to the desired goal – fully functional electronically driven University completely moved into the cloud.

The only requirement is that system must work, and must be reachable from any institution using its resources. What faculties need to do, is to provide their requirements regarding processing power, storage and bandwidth.

Regarding functionalities stated in provided project documents, we can make some additional observations, and give our ideas on improving service level and interaction. We recognize that full implementation of ELF framework through transformation of iKnow system into Cloud Computing solution will require substantial additional reworking.

Some of the cloud services beneficent for students and staff have been already implemented at the University, i.e. at the Faculty of Computer Sciences and Engineering. FCSE Computing Centre is using Microsoft's Live@Edu facilities in the form of Email service and SkyDrive as Cloud Storage [7]. Our experience is teaching us that these cloud services provide high-level functionalities without need to use technical resources and staff to implement, use and maintain these applications. Other educational institutions have positive experience with cloud services as well – for

example, various services at JISC CETIS (Centre for Educational Technology & Interoperability Standards) [8], such as Microsoft Live@Edu [9], Enrollment RX [10], Amazon Web Services [11], and so on.

3 Moving into cloud

Strategic Technologies Group at JISC CETIS in April 2011 organized Workshop [12], drawing a number of conclusions regarding cloud computing in higher education. In their view, there are a number of services suitable for moving into the Cloud, such as: Email, Travel expenses, Human Resources, Finance, Student network services, Telephone services, File storage and Infrastructure as a Service. They also postulated which services should stay "in house": Course and curriculum management, Admission process and Research processes. Another researcher at JISC CETIS, Yuan Li, formulates these conclusions with following observation [13]: "…Clearly, those are non core business or business critical to institutions are most likely to use cloud computing services. More complex, more customised data would be more difficult to move to cloud…". Clearly, in this context, "cloud" means public clouds, as opposed to private clouds, owned by service consumer.

Following these observations and [4], we provide our view on what iKnow functionalities can be moved to the cloud, and which should be institution governed. First, it should be mentioned that for some of the services, data should originate from faculties themselves and University, i.e. its Computer Center consumes that data; and with certainty, there are situations where Faculties use University provided services. Second, although we mention public and private clouds, it is recommendable for the initial phase, and services encompassed by iKnow project, only using private cloud (i.e. cloud owned by the University). Using functionality distribution among modules, illustrated on Figure 2, we briefly provide our view on functionalities' transfer into cloud.

After thorough analyzing, and mapping of ELF Services with appropriate iKnow functionalities, we divided all services into three tables, similar to division introduced by JISC CETIS E-Learning Framework. It is noticeable that a number of iKnow functionalities [1], such as functionalities regarding Administration and pricing, various reports intended for information exchange with Ministry and Statistical Office, do not map into ELF services. These functionalities can be implemented as SaaS, but require additional reworking.

iKnow Services supported by AAI module – involving services such as wireless infrastructure, authentication and authorization services, and so on - can be directly implemented as IaaS at the University level; at the same time, there is a number of iKnow functionalities directly mapping into ELF services, which require transformation from web application into web services available for different higher education institutions. In fact, most of iKnow functionalities can be mapped into Sample User Agent and Learning Domain Services, while requiring SOA approach.

4 Conclusions

In this paper, we have presented how a number of iKnow services can be mapped into ELF framework, thus becoming fully functional "as a Service" solution. From our brief introduction of offered services, we can see that most of the Universities have consolidated their computer centers and offer various cloud and non-cloud solutions. It is also evident that cloud solutions offered, present mix of various in-house and outsourcing solutions, such as Google Apps, Microsoft's Live@Edu, and others.

We have also offered insight into our past experience with various cloud solutions, which are positive, and encourage further investment in this direction. Computer Center at FCSE is also working on various cloud infrastructures and on establishing Authentication and Authorization Infrastructure with Shibboleth, which could be used at University-wide level.

iKnow system, intended to provide more than seventy five important functionalities which will enable concerned parties at Universities in Macedonia to fulfill several very important business processes, can only benefit from moving its functionalities into the cloud. We gave our arguments for this transfer, firmly believing that it will provide numerous benefits, such as increased interoperability, central bookkeeping and data warehousing, tightened security, campus-wide authentication and authorization, and so on

References

1. http://iknow.ii.edu.mk/
2. The E-Learning Framework, found at http://www.elframework.org/
3. The E-Learning Framework (http://www.elframework.org/framework.html)
4. Wilson, Scott, Blinco, Kerry, and Rehak, Daniel, Service-Oriented Frameworks: Modelling the infrastructure for the next generation of e-Learning Systems, Briefing paper, 2004
5. Marjan Gusev et al., E-Students Information System, Software Functional Requirements, UKIM Innovation Technologies Lab, 2010
6. Armenski, G., Gusev, M., e-Testing based on Service Oriented Architecture, 2005
7. http://students.finki.ukim.mk
8. Kraan, Wilbert, and Yuan, Li, Cloud Computing in Institutions, Briefing paper, 2010
9. Microsoft Live@Edu (http://www.microsoft.com/liveatedu/),
10. Enrollment RX (http://www.enrollmentrx.com/),
11. Amazon Web Services (http://aws.amazon.com/)
12. http://wiki.cetis.ac.uk/STG_workshop
13. http://blogs.cetis.ac.uk/cetisli/2010/04/07/cloud-computing-in-institutions-%E2%80%93-non-core-business-or-critical-data/

The Optimization of the Profit of a Parallel System with Independent Components and Linear Repairing Cost

Marija Mihova[1], Bojan Ilijoski[1], Natasha Stojkovich[2]

[1]Faculty of computer science and engineering, "Ss. Cyril and Methodius" University, Skopje, Macedonia

[2] Faculty of informatics, "Goce Delčev " University, Shtip, Macedonia

Abstract. The parallel system can be regarded as a multi-state system with graduate failure. When the system is not in its perfect state, it can be repaired to some higher level under some cost, in our case, to repair k components costs $C_0 + kC_1$. The objective of this research is to find the optimal repairing policy, so that the system makes the greatest possible profit. The main idea of the optimal solution is based on the analysis of the system performance during periods with a certain length, which allows us to use dynamic programming as optimizing technique. Additionally for the systems with unlimited working time we give a way for computation of the optimal repairing level.

Keywords: Optimal profit, multi-state system, parallel system.

1 Introduction

Consider a parallel system with n independent components, so that each can be in either working or failure state. Some examples of such system are equal machines in a factory that do the same work, computers in the laboratory, buses in a transportation company or n-triple transportation line. The cost to repair a failure component does not always depend only on the bill for individual repairing, but sometimes there are additional penalties that need to be paid like a transportation expenses or some influences on the system as a result of system reconstruction. For that reason, we assume that making a collection of recovers takes constant price C_0 and recovering of individual component takes come expected cost C_1. The whole system can be regarded as a multi-state system, so k effective operating components can be regarded as a system working in level k. During the operation some of the components may fail and we assume that the random variable "time to failure" of each component has exponential distribution with parameter λ. Since the our assumption is that components are independent, one level transition intensities can be regarded as independent Markovian transitions. If in the inspection there are failed components, we may decide to recover some of them. The objective here is to decide on which level it is best to get the system, thereby obtaining the maximal future operating profit.

S. Markovski, M. Gusev (Editors): ICT Innovations 2012, Web Proceedings, ISSN 1857-7288

A similar machine replacement problem is given in [1], where the problem is solved by using dynamics programming, so, here we are led by the same idea. The similar computations of the optimal policy for another type of multi-state systems are given in [3] and [4], where it is found that when the system works long enough there is a level on which the optimal profit is obtained, whenever the system is repaired when it is found under that level. On that line, we concentrate on analyzing the existence of such level.

2 The Optimal Policy for Constant Time Periods

Consider a system that operates m time periods with length T. During a period of operation some of the components may fail, i.e. the state of the system can become worst. We assume that at the start of each period, we know the state of the system and we must choose to let the system operate one more period in the state it currently is or repair k of the components for a cost $C_0 + kC_1$. Also we assume that the expected operating profit each component makes when it is in the working state for a unit of time is known and we will denote it by C. The problem we regard is to find the optimal policy for system repairing in order to obtain the benefit of bigger future operating profit.

To solve the problem using dynamics programming we need to identify it optimal substructure. Let $\tilde{C}_i(m)$ be the expected future optimal profit the system makes in the next k periods of length T, under assumption that it started in state i and at the beginning of the time interval mT the failure components are not repaired. By $\hat{C}_i(m)$ we will denote the expected future optimal profit the system makes in the next k periods of length T, under assumption that it started in state i. The problem has the following

Optimal substructure: For all $0 \le k \le n$

$$\tilde{C}_k(m+1) = \tilde{C}_k(1) + \sum_{i=0}^{k} \hat{C}_i(m) \binom{k}{i} e^{-i\lambda T} (1 - e^{-\lambda T})^{k-i} \tag{1}$$

$$\hat{C}_k(m) = \max(\{\tilde{C}_k(m)\} \cup \{\tilde{C}_{k+r}(m) - (C_0 + rC_1) \mid 0 < r \le n-k\}) . \tag{2}$$

Next in this chapter we will show that (2) can be simplified. It is clear that $\forall m, \hat{C}_n(m) = \tilde{C}_n(m)$, so we are concentrating on computation $\hat{C}_k(m)$ for $k < n$.

The expected profit one component makes for time T, if at the beginning of the period it is in failure state (and it is not repairing) is $\tilde{C}_0(1) = 0$, and if at the beginning of the period it is in working state is equal to

$$\tilde{C}_1(1) = \int_T^{\infty} CT\lambda e^{-\lambda t} dt + \int_0^T Ct\lambda e^{-\lambda t} dt = \frac{C(1 - e^{-\lambda T})}{\lambda} .$$

It is easy to conclude that the expected profit *k* component makes in time *T* if at the beginning of the period all of them are in working state is equal to

$$\widetilde{C}_k(1) = \frac{kC(1-e^{-\lambda T})}{\lambda}.$$

We will say that an *n*-component system is profitable if it is feasible to be repaired when all components are in the failure state. It means that there is a period m and number of components *k* such that

$$\widetilde{C}_k(m) - (C_0 + kC_1) \geq 0.$$

For an *n*-component profitable system we have that there is a integer *m* such that

$$\frac{C(1-e^{-\lambda mT})}{\lambda} = \frac{\widetilde{C}_k(mT)}{k} \geq \frac{C_0}{k} + C_1 > \frac{C_0}{n} + C_1. \tag{3}$$

By $\hat{m}$ we will denote the smallest integer such that (3) holds.

Proposition 2.1 $\forall m \in \boldsymbol{N}^+,\ m < \hat{m}$ and $\forall k = \overline{0,n}$, $\hat{C}_k(m) = \frac{kC(1-e^{-\lambda mT})}{\lambda}$.

Proof. The proposition is trivial for $\hat{m} = 1$. Let $\hat{m} > 1$ and *m* = 1. We need to proof that $\frac{kC(1-e^{-\lambda T})}{\lambda} > \hat{C}_n(1) - (C_0 + (n-k)C_1)$. Since $\hat{C}_n(1) = \frac{nC(1-e^{-\lambda T})}{\lambda}$ we have:

$$\frac{kC(1-e^{-\lambda T})}{\lambda} - \left(\frac{nC(1-e^{-\lambda T})}{\lambda} - (C_0 + (n-k)C_1)\right) = C_0 + (n-k)C_1 - \frac{(n-k)C(1-e^{-\lambda T})}{\lambda}$$

$$> C_0 + (n-k)C_1 - (n-k)\left(\frac{C_0}{n} + C_1\right) = \frac{kC_0}{n} > 0.$$

Suppose that the Proposition holds for all $m < m_1 < \hat{m}$. Using (1) it is easy to prove that $\hat{C}_n(m_1) = \widetilde{C}_n(m_1) = nC(1-e^{-\lambda m_1 T})/\lambda$ and $\widetilde{C}_k(m_1) = kC(1-e^{-\lambda m_1 T})/\lambda$. Again

$$\widetilde{C}_k(m_1) - \left(\hat{C}_n(m_1) - (C_0 + (n-k)C_1)\right) > C_0 + (n-k)C_1 - (n-k)\left(\frac{C_0}{n} + C_1\right) > 0.$$

Let $q = 1 - p$. Using identities $p^r(1-p)^l = p^{r+1}(1-p)^l + p^r(1-p)^{l+1}$ and $\binom{k+1}{i} = \binom{k}{i} + \binom{k}{i-1}$. we can easy prove the following identity

$$\sum_{i=0}^{k+1} A_i \binom{k+1}{i} p^i q^{k+1-i} - \sum_{i=0}^{k} A_i \binom{k}{i} p^i q^{k-i} = p\sum_{i=0}^{k} (A_{i+1} - A_i)\binom{k}{i} p^i q^{k-i} \tag{4}$$

Lemma 2.1 For all positive integers *k* and *m*, such that $m \geq \hat{m}$, $\hat{C}_{k+1}(m) - \hat{C}_k(m) \geq C_1$.

Proof: If $\hat{C}_k(m) = \hat{C}_{k'}(m) - (C_0 + (k'-k)C_1)$, $k' > k$, the Lemma is true since $\hat{C}_{k+1}(m) - \hat{C}_k(m) \geq \left(\hat{C}_{k'}(m) - (C_0 + (k'-(k+1))C_1)\right) - \left(\hat{C}_{k'}(m) - (C_0 + (k'-k)C_1)\right) = C_1$.

Now let $\hat{C}_k(m) = \widetilde{C}_k(m)$. The lemma is true for $m = \hat{m}$ since

$$\hat{C}_{k+1}(\hat{m}) - \hat{C}_k(\hat{m}) > \widetilde{C}_{k+1}(\hat{m}) - \widetilde{C}_k(\hat{m}) = \frac{C(1-e^{-\lambda \hat{m} T})}{\lambda} > \frac{C_0}{n} + C_1 > C_1.$$

Suppose that the lemma is true for all $i < m$. For $m + 1$, using (4) and $C(1-e^{-\lambda(\hat{m}-1)T}) / \lambda \leq C_0 / n + C$ we have

$$\begin{aligned}
\hat{C}_{k+1}(m+1) - \hat{C}_k(m+1) &\geq \widetilde{C}_{k+1}(m+1) - \widetilde{C}_k(m+1) \\
&= \frac{C(1-e^{-\lambda T})}{\lambda} + e^{-\lambda T} \sum_{i=0}^{k} \left(\hat{C}_{i+1}(m) - \hat{C}_i(m)\right) \binom{k}{i} e^{-i\lambda T} (1-e^{-\lambda T})^{k-i} \\
&\geq \frac{C(1-e^{-\lambda \hat{m} T})}{\lambda} - \frac{Ce^{-\lambda T}(1-e^{-\lambda(\hat{m}-1)T})}{\lambda} + e^{-\lambda T} C_1 \\
&\geq \frac{C_0}{n} + C_1 - \frac{Ce^{-\lambda T}(1-e^{-\lambda(\hat{m}-1)T})}{\lambda} + e^{-\lambda T} C_1 \\
&\geq \frac{C_0}{n} + C_1 - e^{-\lambda T}\left(\frac{C_0}{n} + C_1\right) + e^{-\lambda T} C_1 \geq C_1.
\end{aligned}$$

Theorem 2.1: Suppose that for some $m \in \boldsymbol{N}$, there is $k' > k$ such that

$$\hat{C}_k(m) = \hat{C}_{k'}(m) - (C_0 + (k'-k)C_1) \,. \tag{5}$$

and k' is the greatest integer that satisfied (5), then $k' = n$.

Proof: Let $k' < n$, then using Lemma 2.1

$$\begin{aligned}
\hat{C}_{k'+1}(m) - (C_0 + (k'+1-k)C_1) &= \hat{C}_{k'+1}(m) - \hat{C}_{k'}(m) + \hat{C}_{k'}(m) - (C_0 + (k'-k)C_1) - C_1 \\
&\geq C_1 + \hat{C}_k(m) - C_1 = \hat{C}_k(m),
\end{aligned}$$

So k' is not the greatest integer for which (5) holds, which is contradiction. So, $k'=n$.

The last Theorem simplifies the formula (2) to

$$\hat{C}_k(m) = \max\{\widetilde{C}_k(m), \widetilde{C}_n(m) - (C_0 + (n-k)C_1)\}. \tag{6}$$

Theorem 2.2 Let $k < n$ be an integer such that $\hat{C}_k(m) = \hat{C}_n(m) - (C_0 + (n-k)C_1)$. Then for all $k' < k$, $\hat{C}_{k'}(m) = \widetilde{C}_n(m) - (C_0 + (n-k')C_1)$.

Proof. $\hat{C}_k(m) = \hat{C}_n(m) - (C_0 + (n-k)C_1)$ implies $\widetilde{C}_k(m) < \widetilde{C}_n(m) - (C_0 + (n-k)C_1)$ i.e. $\widetilde{C}_n(m) - \widetilde{C}_k(m) > C_0 + (n-k)C_1$.

For $k' = k - 1$, using the proof of Lemma 1 we have

$$\widetilde{C}_n(m) - \widetilde{C}_{k-1}(m) = \widetilde{C}_n(m) - \widetilde{C}_k(m) + \widetilde{C}_k(m) - \widetilde{C}_{k-1}(m) > C_0 + (n-k)C_1 + C_1.$$

This implies $\widetilde{C}_{k-1}(m) < \widetilde{C}_n(m) - (C_0 + (n-(k-1))C_1)$, so the Theorem holds for k-1. By induction we have that the theorem holds for all $k' < k$.

The last Theorem tells us that at the beginning of each time interval mT, there is a level $k \le n$, so that for all levels smaller then k the optimal policy is obtained by re-pairing all the failure components and all the levels bigger and equal to k, the optimal policy is obtained when the failure components are not repaired. We will call this level boundary level for m-th step.

3 The Algorithm for Evaluation of the Optimal Repairing Policy

Using the earlier analysis we can construct an algorithm for evaluation of the optimal repairing policy, that takes O(mn). The psevdocode of the algorithm is

```
Input: C0, C1, C, m, λ, T.
Output: The boundary levels for k-th step, k = 1,m
for k=1 to n do
        Ĉ[k] = 0 ;
        C̃[k] = kC(1 − e^{−λT}) / λ ;
for 1 to m do
      Ĉ[n] = C̃[n] + Σ_{i=0}^{n} Ĉ[i] (k choose i) e^{−iλT} (1 − e^{−λT})^{k−i}
       k=0;
       while  Ĉ[k] + Σ_{i=0}^{k} Ĉ[i] (k choose i) e^{−iλT} (1 − e^{−λT})^{k−i} < Ĉ[n] − (C0 + (n − k)C1)  do
                    Ĉ[k] = Ĉ[n] − (C0 + (n − k)C1) ;
                     k=k+1;
       print k
      for j = k to n do  Ĉ[j] = Ĉ[j] + Σ_{i=0}^{j} Ĉ[i] (k choose i) e^{−iλT} (1 − e^{−λT})^{j−i} .
```

It is interesting that in the most of the experiments we made, for different boundary level for m-th step grows as m grows up, and there is some boundary level for $m \to \infty$. But there are some examples in which for some particular levels m and $m + 1$ the boundary level for m-th step is greater then the boundary level for $m + 1$-th step. Next we give such an example.

Example 3.1 Let $\lambda = 1$ $e^{-T} = 0.4545$, $C = 20.18$, $C_0 = 10$, $C_1 = 6$, $n = 2$. For $m = 1$ $\hat{C}_0(1) = \hat{C}_2(1) - (C_0 + 2C_1) = 0.016$ and $\hat{C}_1(1) = \widetilde{C}_1(1) = 11.008$ so the boundary level is 0. The boundary level for $m = 2$ is 1 since $\hat{C}_0(2) = \hat{C}_0(2) - (C_0 + 2C_1) = 10.028$ and $\hat{C}_1(2) = \hat{C}_2(2) - (C_0 + C_1) = 16.028$. But, the boundary level for $m = 3$ is 0 again because $\hat{C}_0(3) = \hat{C}_2(3) - (C_0 + 2C_1) = 17.5638$ and $\widetilde{C}_1(3) = \widetilde{C}_1(3) = 23.7621$. At the next steps, the boundary level remain at 0.

4 Boundary level for a system with unlimited working time

This boundary levels we get in our experiments, inspired us to make an additional analyzing in order to realize the existence of a boundary level when the working time of the system is unknown and we believe that that time is unlimited.

Again we regard a parallel system with n independent components so that the profit each component makes, if it is in the working state for a unit time, is C and the repairing cost for k components cost $C_0 + kC_1$.

Theorem 4.1 Suppose that whenever the system fails to level s, it is repaired to level k. Then the expected mean profit of such system is equal to

$$L_{k,s} = \frac{\lambda(k-s)\left(\frac{C}{\lambda} - \left(\frac{C_0}{k-s} + C_1\right)\right)}{\sum_{i=s+1}^{k} \frac{1}{i}}.$$

Moreover, the maximal expected mean profit is obtained in the case when $k = n$.

Proof. If the system starts with its work in state i, then the expected time to work in level i is equal to $(i\lambda)^{-1}$. The expected transition time from level k to level s is

$$\sum_{i=s+1}^{k} \frac{1}{i\lambda} = \frac{1}{\lambda} \sum_{i=s+1}^{k} \frac{1}{i}.$$

The profit it makes during that time is $\sum_{i=s+1}^{k} \frac{iC}{i\lambda} = \frac{k-s}{\lambda} C.$ To repair it to level k costs $C_0 + (k - s)C_1$. So, the expected mean cost is equal to

$$\frac{\frac{k-s}{\lambda}C-(C_0+(k-s)C_1)}{\frac{1}{\lambda}\sum_{i=s+1}^{k}\frac{1}{i}}=\frac{\lambda(k-s)\left(\frac{C}{\lambda}-\left(\frac{C_0}{k-s}+C_1\right)\right)}{\sum_{i=s+1}^{k}\frac{1}{i}}.$$

In order to proof the second stage of the theorem, we will show that the expected profit when the system is repaired to level $k + 1$ is greater then the expected profit when the system is repaired to level k, whenever it falls to level s, i.e. we will show that the following inequality is true

$$\frac{\lambda(k+1-s)\left(\frac{C}{\lambda}-\left(\frac{C_0}{k+1-s}+C_1\right)\right)}{\sum_{i=s+1}^{k+1}\frac{1}{i}}>\frac{\lambda(k-s)\left(\frac{C}{\lambda}-\left(\frac{C_0}{k-s}+C_1\right)\right)}{\sum_{i=s+1}^{k}\frac{1}{i}}.$$

The last inequality is equivalent to

$$\left(((k-s)+1)\sum_{i=s+1}^{k}\frac{1}{i}\right)\left(\frac{C}{\lambda}-\left(\frac{C_0}{k+1-s}+C_1\right)\right)>\left((k-s)\sum_{i=s+1}^{k+1}\frac{1}{i}\right)\left(\frac{C}{\lambda}-\left(\frac{C_0}{k-s}+C_1\right)\right).$$

It is clear that $\frac{C}{\lambda}-\left(\frac{C_0}{k+1-s}+C_1\right)>\frac{C}{\lambda}-\left(\frac{C_0}{k-s}+C_1\right)$. From the other hand

$$((k-s)+1)\sum_{i=s+1}^{k}\frac{1}{i}=(k-s)\sum_{i=s+1}^{k}\frac{1}{i}+\sum_{i=s+1}^{k}\frac{1}{i}=(k-s)\sum_{i=s+1}^{k+1}\frac{1}{i}+\sum_{i=s+1}^{k}\frac{1}{i}-\frac{k-s}{k+1}>(k-s)\sum_{i=s+1}^{k+1}\frac{1}{i}.$$

The last inequality follows from the fact that for all i, $i < k + 1$ which implies $\frac{1}{i}>\frac{1}{k+1}$, so

$$\sum_{i=s+1}^{k}\frac{1}{i}-\frac{k-s}{k+1}>\sum_{i=s+1}^{k}\frac{1}{k+1}-\frac{k-s}{k+1}=0.$$

If $\frac{C}{\lambda}\le\frac{C_0}{n}+C_1$, the system is unprofitable i.e. it is not profitable to repair it. So we will regard only the systems for which $\frac{C}{\lambda}>\frac{C_0}{n}+C_1$. Our goal is to find the level s for which the mean expected profit will be maximal. To do this we need to compare profits P_{n,s_1} and P_{n,s_2} for all $0 \le s_1, s_2 < n$. The next Theorem gives the boundary for C/λ that under which $P_{n,s} > P_{n,s+r}$, $0 < r < n - s$.

Theorem 4.2 For all $0 \le s < n$ and $0 < r < n - s$, $L_{n,s} > L_{n,s+r}$ if

$$A = \frac{1}{C_0}\left(\frac{C}{\lambda} - C_1\right) < \frac{\sum_{i=s+1}^{s+r} \frac{1}{i}}{(n-s)\sum_{i=s+1}^{s+r} \frac{1}{i} - r\sum_{i=s+1}^{n} \frac{1}{i}} = B_{s,s+r}. \quad (7)$$

Proof. By simple transformation the inequality $L_{n,s} > L_{n,s+r}$ becomes

$$\left(\frac{\sum_{i=s+1}^{n} \frac{1}{i} - \sum_{i=s+r+1}^{n} \frac{1}{i}}{(n-(s+r))\sum_{i=s+1}^{n} \frac{1}{i} - (n-s)\sum_{i=s+r+1}^{n} \frac{1}{i}}\right) C_0 + C_1 > \frac{C}{\lambda}.$$

Using $\sum_{i=s+1}^{n} \frac{1}{i} - \sum_{i=s+r+1}^{n} \frac{1}{i} = \sum_{i=s+1}^{s+r} \frac{1}{i}$ we get $B_{s,s+r}C_0 + C_1 > \frac{C}{\lambda}$, which is equivalent with (7).

Proposition 4.1 $\forall s, r, k$ such that $0 \le s, r, k$ and $s + r + k < n$, $B_{s,s+r} < B_{s,s+(r+k)}$.

Proof: By simple transformation $B_{s,s+r} < B_{s,s+(r+k)}$ becomes

$$\left(\sum_{i=s+1}^{s+r} \frac{1}{i}\right)\left((n-s)\sum_{i=s+1}^{s+r+k} \frac{1}{i} - (r+k)\sum_{i=s+1}^{n} \frac{1}{i}\right) < \left(\sum_{i=s+1}^{s+r+k} \frac{1}{i}\right)\left((n-s)\sum_{i=s+1}^{s+r} \frac{1}{i} - r\sum_{i=s+1}^{n} \frac{1}{i}\right).$$

It is easy to see that after multiplication of the both sides, the first terms will be equal. So this inequality is equivalent with

$$r\sum_{i=s+1}^{s+r+k} \frac{1}{i} < (r+k)\sum_{i=s+1}^{s+r} \frac{1}{i} \Leftrightarrow \sum_{j=s+1}^{s+r}\sum_{i=s+1}^{s+r+k} \frac{1}{i} < \sum_{j=s+1}^{s+r+k}\sum_{i=s+1}^{s+r} \frac{1}{i}$$

$$\Leftrightarrow \sum_{j=s+1}^{s+r}\left(\sum_{i=s+1}^{s+r} \frac{1}{i} + \sum_{i=s+r+1}^{s+r+k} \frac{1}{i}\right) < \sum_{j=s+1}^{s+r}\sum_{i=s+1}^{s+r} \frac{1}{i} + \sum_{j=s+r+1}^{s+r+k}\sum_{i=s+1}^{s+r} \frac{1}{i}.$$

Again we have equal first terms, so $B_{s,s+r} < B_{s,s+(r+k)}$ is equivalent to

$$\sum_{j=s+1}^{s+r}\sum_{i=s+r+1}^{s+r+k} \frac{1}{i} < \sum_{j=s+r+1}^{s+r+k}\sum_{i=s+1}^{s+r} \frac{1}{i} \Leftrightarrow \sum_{j=s+r+1}^{s+r+k}\sum_{i=s+1}^{s+r}\left(\frac{1}{i} - \frac{1}{j}\right) > 0.$$

The last inequality holds because for all i and j, $i \le s + r < j \Leftrightarrow 1/j < 1/i$.

Proposition 4.2 For all $0 < s < n$, $B_{s-1,\,s} < B_{s,\,s+1}$.

Proof: We need to proof that

$$\frac{\frac{1}{s}}{\frac{n-(s-1)}{s}-\sum_{i=s}^{n}\frac{1}{i}}<\frac{\frac{1}{s+1}}{\frac{n-s}{s+1}-\sum_{i=s+1}^{n}\frac{1}{i}},$$

which is equivalent to

$$\frac{n-s}{s(s+1)}-\frac{1}{s}\sum_{i=s+1}^{n}\frac{1}{i}<\frac{n-s+1}{s(s+1)}-\frac{1}{s+1}\sum_{i=s}^{n}\frac{1}{i}\Leftrightarrow-(s+1)\sum_{i=s+1}^{n}\frac{1}{i}<1-s\sum_{i=s}^{n}\frac{1}{i}\Leftrightarrow-\sum_{i=s+1}^{n}\frac{1}{i}<1-1=0.$$

Using last two propositions we can prove the following theorem that characterizes the optimal repairing level.

Theorem 4.1 Let $B_{-1,0}=\frac{1}{n}$, $B_{n,n+1}=\infty$ and $B_{s,s+1}$ are defined as in Theorem 4. Then if $B_{s-1,s}\le A<B_{s,s+1}$, the maximal profit is obtained when all failed components are repaired whenever the parallel system is found at level *s*.

Proof. Using Proposition 1 we have $A<B_{s,s+1}\le B_{s,s+k}$, for all *k*, $1\le k\le n-s$. From Theorem 3 we have that for all *k*, $1\le k\le n-s$ $L_{n,s}>L_{n,s+k}$.

From the other side, since $B_{s-1,s}\le A$, from the Proposition 2 we have that for all $1\le k\le s$, $B_{s-k,s-k+1}\le A$. Using Theorem ? we have that $L_{n,s-k}<L_{n,s-k+1}$, for all $1\le k\le s$. This imply $L_{n,s}>L_{n,s-k}$, for all $1\le k\le s$.

From the last two theorems, in order to find the optimal repairing level we need to calculate all B_{s-1}, *s* for $0\le s\le n$. These numbers grow as *s* grows up, so we need to find the interval in which A belongs. If it is found on the interval $(B_{s-1},s, B_{s,s+1})$, then we will conclude that the maximal mean profit will be obtained if the system is repaired in the moment when it is found at level *s*. These intervals are given in Fig. 1.

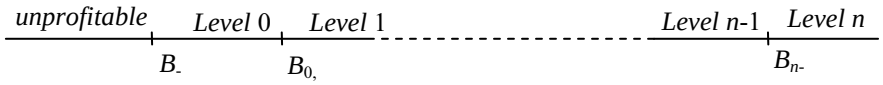


Fig. 1. Decision intervals.

Example 4.1 Consider a system with 4 components. The boundaries are $B_{-1,\,0}$ = 1/4, B_0, 1 = 12/23, B_1, 2 = 6/5 and B2, 3 = 4. Let C_0=3 and C_1 = 1. Then for C/λ = 1.5, A = 0.167 so unprofitable. For C/λ = 2, A = 0.33 so the optimal repairing level is 0. For C/λ = 6, A = 0.67 so the optimal repairing level is 2. For C/λ = 15, A = 4.67 so the optimal repairing level is 3.

The same result we obtained experimentally. The operation of such system during the time T1, much greater than T, was simulated. Whenever the system enters the specific level *k*, it is repaired to the level s, ∀s > k. The optimal profit was always received for *s* = *n* and the optimal level matches with the theoretical results.

We can design an O(*n*) algorithm for calculating the boundaries and finding the optimal repairing level.

```
Input: C0, C1, C, n, λ.
Output: if the system is profitable, the optimal repairing
level, else the message that the system is not profitable
A=(1/C0)(C/λ- C1)
if A < 1/n then print "the system is unprofitable" else
  S = 1/n
  s = n-1
  while A<=(1/(s+1))/( (n-s)/(s+1)-S) and s>=0 do
    S = S+1/(s+1);
    s=s-1;
if s ≠ -1 then print "the optimal repairing level is" s+1.
```

5 Conclusion

This paper deals with operating process of a parallel *n*-component system. The objective is to find the level to which the system with *k* efficiently operating components at the beginning of each considering time needs to be repaired, or to make decision to left it at the current state, in order to obtain the maximal future operating profit. We regard two types of systems. For the first type we assume that at the start of each period we know its state and only in that moment we are able to repair some of its components. For such systems, we showed that in order to obtain the optimal future operation profit, at the beginning of each period we need to make decision only between two choices, to repair all failure components or to left the system operate one more period in its current state. Moreover, it is shown that there is a boundary level under which the optimal policy is obtained if we leave the system in its current state. The current state for the second type of systems is known at any moment during their operation, which also allows us to make decision to repair its components in each moment. It is shown that there is a level *k* under which it is not profitable to fail, i.e. that the optimal mean profit is obtained if all failure components were repaired whenever the systems takes that level *k*.

Bibliography

1. Bertsekas, D.P., Dinamic Programming and Optimal Controll (2005)
2. Cormen, T. H., Leiserson, C. E., Rivest, R. L., Stein, C., Introduction to Algorithms, The MIT Press (2002)
3. Mihova, M., Ilijoski, B., The Optimal Profit Repairing Policy of One-component Multi-state System with Graduate Failure, CIIT (2012)
4. Mihova, M., Stojkovic, N., Simulation on the Profit of Work on Multi-state Two-terminal Transportation System, CIIT (2012)

E-business opportunities and challenges for SME's in Macedonia

Florim Idrizi[1], Fisnik Dalipi[2], Ilia Ninka[3]

[1,2]Faculty of Natural Sciences and Mathematics, State University of Tetovo
{florim.idrizi,fisnik.dalipi}@unite.edu.mk
[3]Department of IT, Faculty of Natural Sciences, University of Tirana
ilia.ninka@fshn.edu.al

Abstract. This paper explores the potential of adoption and use of ICT in small and medium sized enterprises (SMEs) in Macedonia. In the paper we present preliminary results of a survey of around 60 SMEs. The purpose of the study is to explore the factors enabling or impeding the successful adoption and use of ICT by SMEs. The study investigates the types of ICT adoption and applications, the overall motivation for ICT investments, the advantages gained from ICT, the motive of using Internet and the difficulties in implementing e-commerce applications. We find that SMEs are generally satisfied with their investment in ICT but they are concerned about the cost of such investments and are uncertain about the business benefits, failing to recognize ICT's strategic potential to increase business flexibility, to increase productivity and to support globalization. Besides the concern about the ICT related cost, other major obstacles in adopting ICT were lack of internal ICT capabilities and lack of information about selecting, implementing and evaluating suitable ICT and e-business solutions. Our findings have important implications for policy aimed at ICT and e-business adoption and use by SMEs and will provide a foundation for future research by helping policy makers to understand, assist and support the SME sector.

Keywords: SMEs, ICT, e-commerce, adoption, challenges, obstacles

1 Introduction

The adoption and use of ICT is critical for the competitiveness of Macedonian's SMEs in the emerging global market, while promoting significant positive consequences on the nation's economy. Through this research we would like to know more about the effects and usage of ICT by SMEs. We investigate the types of ICT adoption and applications, the overall motivation for ICT investments, the advantages gained from ICT, the motive of using Internet and the difficulties in implementing e-commerce applications.

S. Markovski, M. Gusev (Editors): ICT Innovations 2012, Web Proceedings, ISSN 1857-7288

2 ICT adoption in SMEs – theoretical framework

Many studies show that SMEs are the driving engine of growth, job creation, and competitiveness in domestic and global markets. They also play a pivotal role in innovation and productivity growth (Blackburn and Athayde, 2000). In the USA more than half of all employment comes from firms with fewer than 500 employees (Baldwin et al. 2001). In the UK, SMEs employ 67% of the workforce (Lange et al. 2000). In most EU member status SMEs make up over 99% of enterprises, 67% of jobs and 59% of GDP.

It remains a concern for many reasons including for example the scale of global ICT investment and the dissatisfaction expressed by Chief Executive Officers (CEOs) with ICT investment returns. ICT adoption in organisations has grown considerably throughout the past three decades. By 1998, in the developed world, ICT accounted for more than 50% of organisations annual capital investments and was expected to account for 5% of revenues by 2010 (Powell, 1999). The main driving force behind this large-scale ICT investment is the promise of increased competitive advantage (Hu and Plant, 2001; Piccoli and Ives, 2005), as ICT is regarded as a strategic weapon that can positively effect organisational change (Gregor et al, 2006). Most SMEs lag behind the large firms in their use of ICT both operationally and strategically. SMEs characteristically lack of managerial skills to conceive, plan and implement ICT and reluctantly update technology (Caldeira & Ward, 2002). Constrained by resources, hemmed in by competing demands, caution and suspicion often greet new technological opportunities. Large firms for example, have adopted e-commerce much faster than SMEs (Pool et al. 2006). There is certainly evidence that SMEs are reacting with caution to the possibilities of e-commerce, considering it a high-risk strategy (Al-Qirim, 2005), introducing e-commerce very slowly into their existing set of operations (Eriksson & Hultman, 2005).

3 Methodology

All the data for this survey was collected using a structured questionnaire. The SMEs are from different cities and regions of Macedonia, focusing upon three economically significant Macedonians sectors: financial services, touristic services and production. In total, we have surveyed 60 firms that have successfully established their business operations and are in market for more than fifteen years or more.

Through this study, our goal is to find answers to some of the following answers:

- What is the type of ICT used by SMEs?
- What kind of ICT application is being used by SMEs?
- What are the motivations for ICT investments?
- What are the barriers impeding ICT investments?
- Which are the benefits gained by ICT?
- What are the reasons of using the Internet?
- What are the challenges in E-commerce implementation?

- What are the reported sorces of ICT advice?
- How many times SMEs use the Internet to advertise their products or services?

The questionnaires were filled out by IT managers or other management people who did understand the nature of the issues investigated by this survey. Companies that do not use any form of ICT are not included in this study.

4 Key Findings

In order to identify and qualify the SMEs, for the purpose of this research, we use the definition given by (Kapurubandara et al, 2006) where businesses with less than ten employees are qualified as Micro Enterprise, between ten and fifty as Small Enterprises, and between fifty to two hundred and fifty employees as Medium sized Enterprises. Based upon these definitions, 32% of SMEs we have surveyed can be classified as Micro Enterprises, 48% as Small Enterprises and 20% are Medium sized Enterprises.

Table 1. Types of surveyed SMEs

Type of SMEs	Percentage
Micro Enterprises	32
Small Enterprises	48
Medium Sized Enterprises	20

4.1 Type of adopted ICT

In this section, we provide the results about the type of adopted and used ICT. Our goal was to investigate if these companies have: Internet connection, website, Extranet, e-mail, any computer network, including wireless technology and EDI (Electronic Data Interchange).

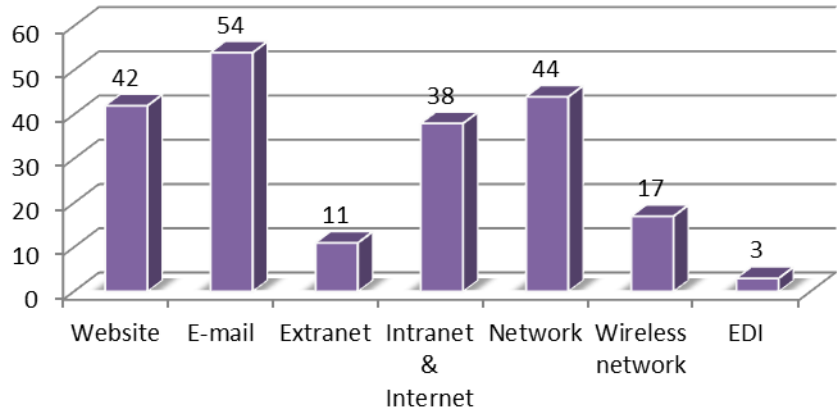


Fig. 1. Type of ICT adopted

Figure 1 indicates that across three sectors, e-mail is widely used by 54 firms. Through this survey we noticed that the majority of firms did not have official e-mail address, but they used standard mail such as hotmail, Yahoo Mail etc. About 44 firms have established network and 42 of them had their own website. Surprisingly, the survey reveals that 38 firms use Intranet whereas only 11 of them have Extranet to control access from the outside for their business purposes. Considering the more advanced and complex technology, only 3 firms from production sector use EDI to transfer electronic documents or business data.

4.2 Type of ICT application

Figure 2 shows the type of ICT application used by SMEs. We can distinguish that ICT applications are mostly used to automate the sales record, manage the documents and for processing orders. About 21 firms use system for design. Applications such as human resource management, market research, enterprise resource planning (ERP), and business intelligence are very modestly spread among investigated firms.

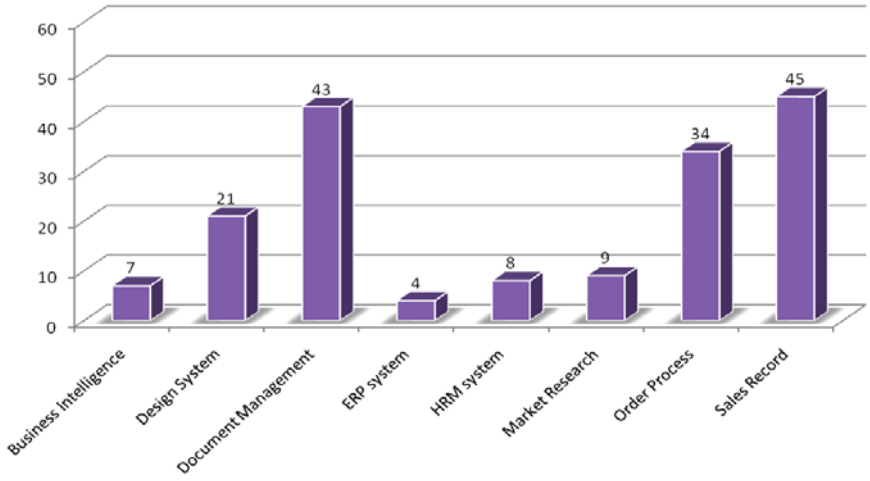


Fig. 2. Type of ICT application

4.3 Motivations for ICT investments

Here we study main driving force behind ICT investments. It is well known that SMEs are constrained by resources to invest in new technological opportunities. Nevertheless, our study reveals that almost all firms are open to invest in ICT in order to gain competitive advantage and to increase their business efficiency. As shown in Figure 3, about 90% of firms are motivated for ICT investments in order to improve relationships with clients and 85% of them are ready to invest in ICT to keep pace with rivals, followed by keeping up with new trend (41 firms) and then increase the operational efficiency. Very few firms (39 firms) invest in ICT to increase the satisfaction among employees.

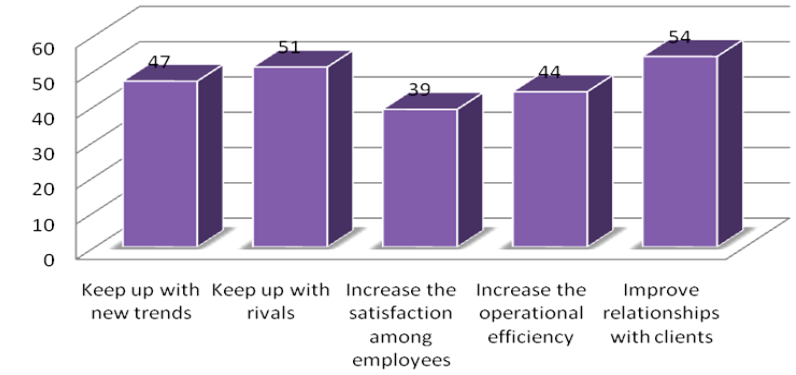


Fig. 3. Motivations for ICT investments

4.4 Barriers impeding ICT investments

Even though some SMEs in Macedonia are aware of ICT benefits, there exist some constraints and barriers to ICT investments. Figure 4 shows that cost and security are the largest barriers cited by firms (above 82% of firms). SMEs are also uncertain over the benefits to their business (28 firms). Just 13 firms (about 22%) answered that they did not have enough IT experience inside the firm. The only barrier which was insignificant and is less cited was the concern about reactions of the staff, cited by only 5% of the firms.

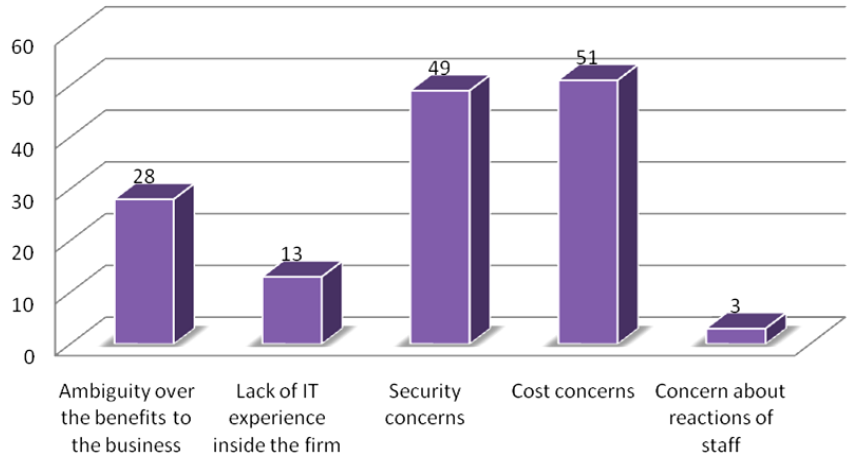


Fig. 4. Barriers impeding ICT investments

4.5 Benefits gained from ICT

In this section we report the benefits gained from ICT. As our next figure shows (Figure 5), the most cited benefit as a result of ICT use and adoption is improved quality of service (85% of firms). The second most answered benefit is keeping up with rivals, experienced by 75% of firms, followed by increasing sales which is cited by 65% of firms. Close to 57% of the firms cited increased productivity as a benefit. ICT is not seen as driving force to improve working on joint projects with other firms (cited by 7 firms), and reduce the staff number (cited by 4 firms). About 5% of firms highlighted that there are no experienced benefits from ICT.

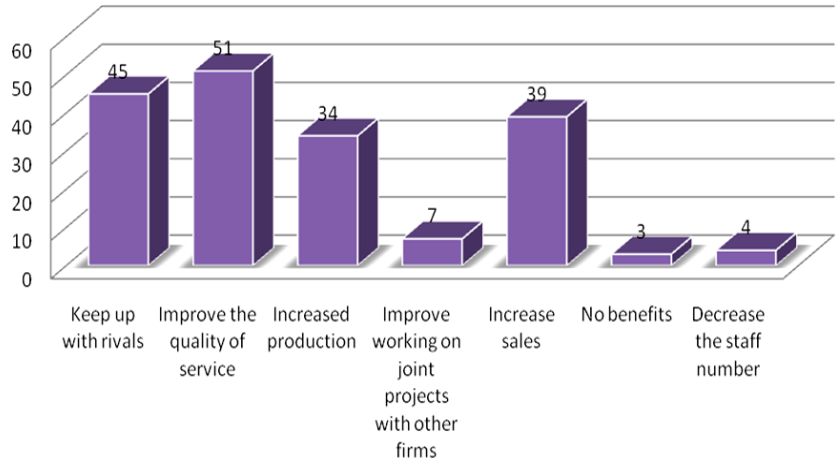


Fig. 5. Benefits gained from ICT

4.6 Reasons for using the Internet

We also questioned SMEs to list a few of the reasons why they're using the Internet. Figure 6 shows that most of the surveyed firms (95% of them) indicated that they use Internet to collect information. They are also using the internet to share information with suppliers (44 firms) and customers (38 firms), followed by making payments to suppliers (highlighted by 36 firms) and placing orders to suppliers (29 firms). Little interest is shown of using the internet to place orders from customers (19 firms) and to get involved in joint projects (only 4 firms).

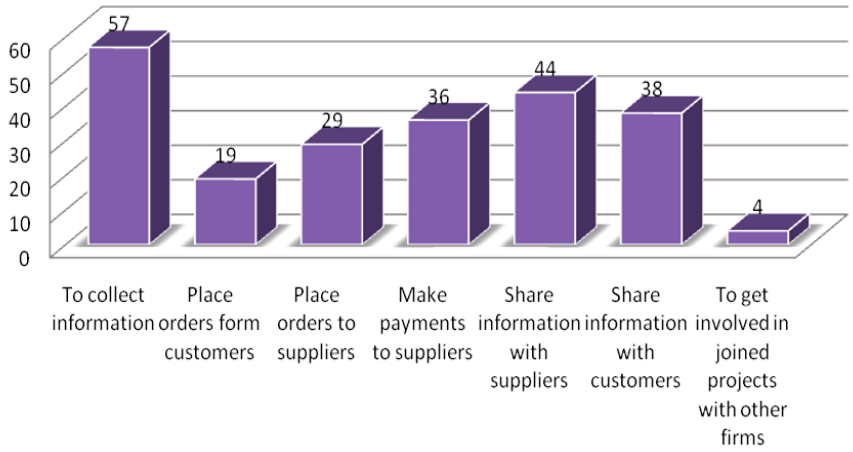


Fig. 6. Reasons for using the Internet

4.7 Challenges in e-commerce implementation

We have also investigated the challenges and problems that SMEs have in implementing e-commerce. As shown in Figure 7, majority of SMEs that participated in this research (51 firms) highlighted the security as a challenge that have found in e-commerce implementation. Another serious challenge to putting e-commerce in place appears to be the internet fraud, cited by 46 firms (about 77% of firms). Less importance is given to the following questions: difficulties in finding technical advice from outside, suppliers are not ready for e-business, the cost of developing the website and customers don't want to change.

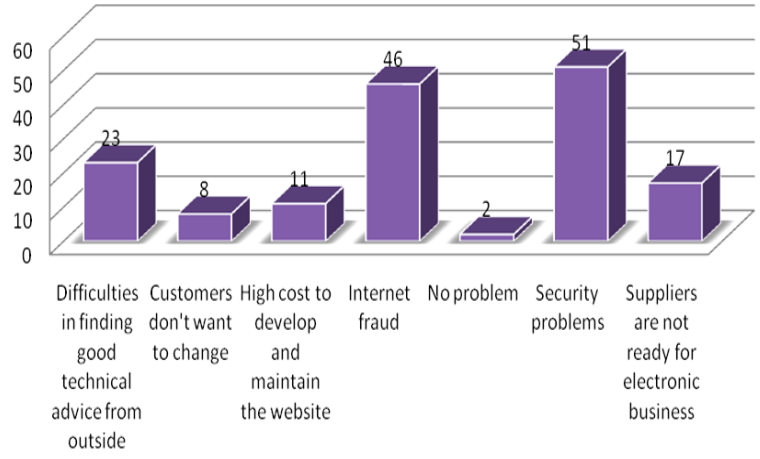


Fig. 7. Challenges in E-commerce implementation

4.8 Reported sources of IT advice

Here we attempt to find out where the firms get advice about ICT. As Figure 8 shows, about 80% of all surveyed firms highlighted the internet as source of advice. Further, the company itself is considered as source for IT advice by 35 firms. IT consultants and governmental/local authorities are also sources to get advice on ICT. Very few of SMEs get ICT advice from suppliers (6 firms) and the media (9 firms).

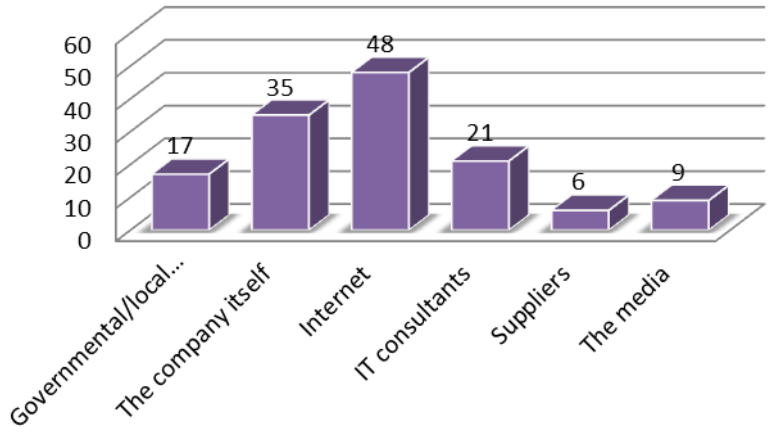


Fig. 8. Reported sources of IT advice

4.9 Internet advertisement of products/services

It is also important to know how many times firms are practicing internet advertisement. Our next figure (Figure 9) shows that majority of firms (34 firms) consider the Internet as inappropriate platform to advertise their products or services. After that, 31 firms advertise their products/services often and only 16 firms advertise many times.

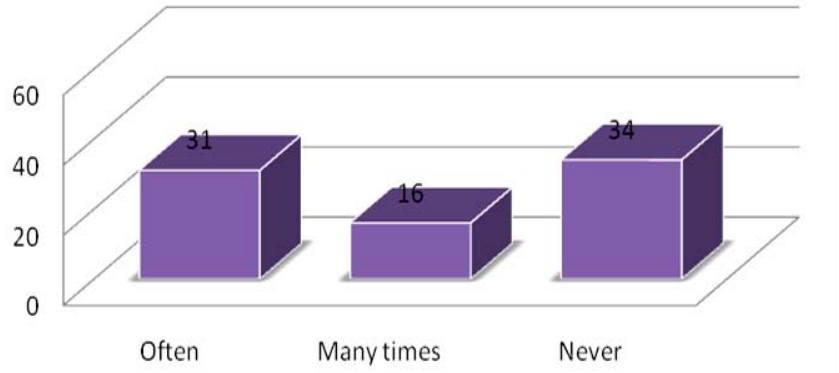


Fig. 9. Internet advertisement

5 Conclusions

This article attempts to discover the challenges and barriers that face SMEs in Macedonia while adopting and using ICT. It addresses the barriers of SMEs to access advanced technologies in order to open up business opportunities and increase productivity. The results of our study show that Macedonians SMEs are doing well in using common technologies such as e-mail, internet access and websites but are very li-

mited to more sophisticated technologies such as EDI, ERP, HRM and business intelligence systems. Our survey reports that SMEs, in general, are motivated to invest in ICT, whereby the main driving force for ICT investment was to improve relationships with clients, to keep up with rivals and to be updated with new technological trends. Nevertheless, our survey suggests that SMEs share concerns over cost of the ICT equipment and the security. With regard to awareness of the benefits of the internet, almost all of them use the internet for collecting information lacking the vision of internet and e-commerce opportunities.

6 References

1. Arendt, L. (2008). "Barriers to ICT adoption in SMEs: how to bridge the digital divide?" Journal of Systems and Information Technology, Volume 10; Issue 2.
2. Al-Qirim, N. (2006). Personas of E-commerce adoption in small businesses in New Zealand. Journal of Electronic Commerce in organizations, 4: 17-45
3. Blackburn, R., and Athayde, R. (2000). Making the connection: The effectiveness of Internet training in small businesses, Education and Training 42(4/5), 289-299
4. Baldwin, J.R., Jarmin, R.S., and Tang, J. (2001). "The trend to smaller producers in manufacturing in Canada and the US", Statistics Canada Working paper.
5. Caldeira, M.M., & Ward, J.M. (2002). Understanding the successful adoption and use of IS/IT in SMEs: an explanation from Portuguese manufacturing industrie. Information Systems Journal, 12: 121-152
6. Eriksson, L.T., & Hultman, J. (2005). One digital leap or a step-by-step approach? – An empirical study of e-commerce development amongst Swedish SMEs. International Journal of Electronic Business, 3: 447-460.
7. Gregor, S., Martin, M., Fernandez, W., Stern, S. and Vitale, M. (2006). The transformational dimension in the realisation of business value from Information Technology. Journal of Strategic Information Systems, 15, 249-270.
8. Hu, Q. and Plant, R. (2001). An empirical study of the causal relationship between IT investment and firm performance. Information Resources Management Journal, 17: 37-62.
9. Kapurubandara, M., and Lawson , R. (2006), "Barriers adopting ICT and E-commerce with SMEs in developing countries: An Exploratory Study in Sri Lanka", CollECTeR '06, 9 December, 2006, Adelaide, [online],
10. http://www.collector.org/archives/2006_December/07.pdf, Consulted: [23.12.2010]
11. Lange, T., Ottens, M., and Taylor, A (2000) "SMEs and Barriers to skills development: A Scottish Perspective", Journal of European Industrial Training, Vol. 24, No.1, pp.5-11
12. Powell, P.L. (1999). Evaluation of Information Technology investments: business as usual? In Beyond the IT productivity paradox, (ed. L.P. Willcocks and S. Lester), pp. 151-182. Wiley, Chichester

Specific Skill Set Training for Working Professionals by Faculties via e-Learning

Renata Petrevska Neckoska[1], Gjorgi Manceski[1]

[1] University "Sv. Kliment Ohridski" (UKLO) Bitola, Republic of Macedonia,
Faculty of Economics Prilep, Republic of Macedonia
renata.pe.ne@gmail.com; gmanceski@t-home.mk

Abstract. The e-Learning concept is a foreseeable tendency in many environments nowadays, with differences in the stage of e-Learning utilization and usage, as well as the paradigm shift that comes with it. In its varieties, it offers fruitful soil for developing systems such as distance and hybrid learning, as well as initiatives such as life-long learning and non-traditional learner education. We are proposing a model of Faculty certification program that trains non-traditional learners with various backgrounds for a specific skill set, needed by working professionals, delivered on demand, at the workplace, with obtained set of competences effective proximately after completion. The model is learner-oriented and specific-skill-set-oriented. It combines the related theoretical backgrounds of several faculty subjects along with practical exercises of what needs to be learned. The example skill set would be Financial Risk Analysis for an Entrepreneur, with brief elaboration of the benefits for the stakeholders in the process.

Keywords: Distance learning, e-Learning, e-Learning utilization, Faculty certification program, Non-Traditional learner, Working professional, Life-long learning, Education, Learner-oriented, Set of competences, Financial Risk Analysis, Entrepreneur

1 Introduction

The distance learning programs, offering complete studying on distance, as well as hybrid learning programs, offering also Face-to-Face activities for the parties involved in the educational process, are widening the access to education in terms of time, space and student profile dimension. With these models available, "students" become "learners", and learning becomes multidimensional activity and experience that overflows the traditional assessment expectations of the educational institutions, but of the students as well. Another aspect that has noted rapid sublimation is the expected timeframe for the learned theory and skills to become implementable in reality, which has reduced to: now, immediately.

"The social transformations caused by the emergence, development and direction of the new knowledge-based economy have led to multiple changes in the educational system. Most of the institutions concern the introduction of IT technology in the educational envi-

adfa, p. 1, 2012.

S. Markovski, M. Gusev (Editors): ICT Innovations 2012, Web Proceedings, ISSN 1857-7288

ronment as the main agent of transmission, dissemination and evaluation of information and their level of assimilation."[1]

The proposed model of education in the form of training for qualification and skills by wrapping the education with certification is based on the idea that there is a necessity for a mentality shift of the opinion that people who got employed have finished with the learning process for good. Number of reasons point in the opposite direction – the learning happens and should be life-long. On the working places, it is done for the purpose of certification, upgrade of qualification, change of qualifications or advancement in the career or expertise.

2 The Distance Learning Concept and Characteristics

The distance learning model, as a way of delivering instructions by teachers and learning theory and practice by learners, in different time and place from one to another, has many stakeholders, each of which investing certain inputs and gaining back valuable knowledge, experience and wisdom.

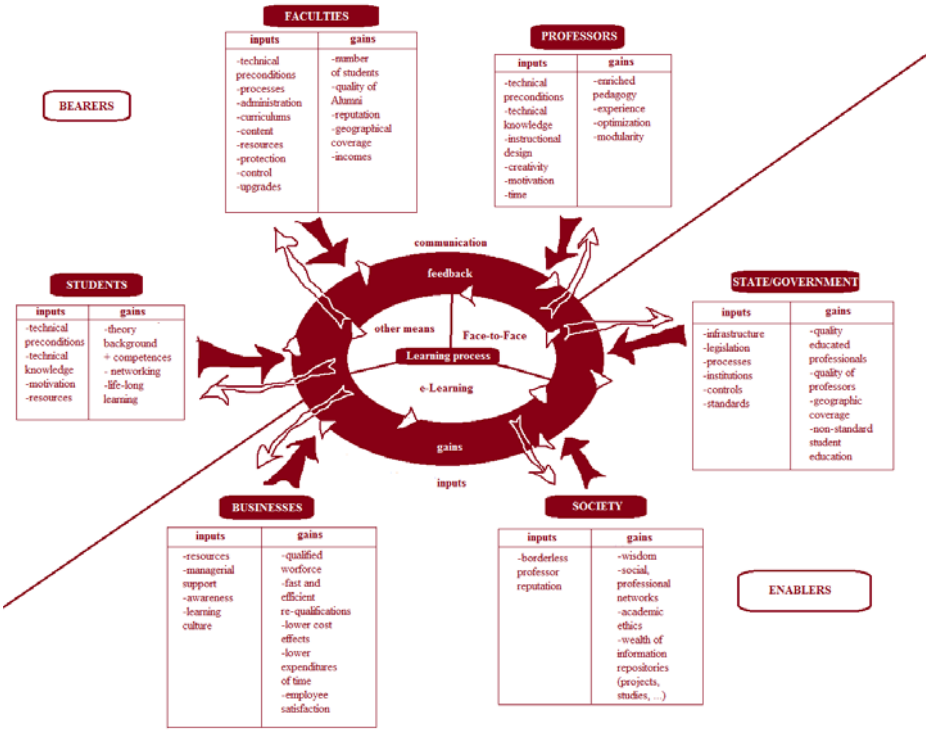


Figure 1. Stakeholder and Relational Model of Distance/Hybrid learning

The stakeholders in the process are categorized as:

- Bearers of the distance learning model of education
- Enablers of the distance learning

The bearers are the Students, Faculties and Teaching staff, who have to carry out the whole process of education and verification, while the enablers are the State/Government, the Society and the Businesses, who have the obligation to integrate and produce and support environment for the institutions to deliver distance learning, for the students to deliver results to the businesses, for the professors to deliver optimized knowledge and skills to the students. The investment each stakeholder has in the process, and the gains from it, are shown on Figure 1.

In this paper, for the purpose of condensed elaboration, we will be referring to the distance learning and hybrid learning models as equally applicable to the discussed topic about skill set trainings for working professionals.

In the context of distance learning prospects, we would like to bring up the term and profile of "non-traditional learner". Starting with the latter word, that signifies the major change of the concept of a student, since the "learner" does not necessarily go to Faculty campus, follow face-to-face lectures, but in most cases does the learning after working hours at the workplace, or in the late hours at home; the learner collaborates with the colleagues, is self-organized, proactive; the learner does not learn just for the sole purpose of being graded and assessed upon reproduction of materials, but learns with the purpose of implementing the skill or knowledge in the workplace or in life, considering that valuation as the ultimate assessment of what has been learned and shared. The term "non-traditional" on the other hand, expands the definition of the "traditional" student in diverse ways, different from each other in terms of geographical placement, in terms of age, in terms of professional background and in terms of goal of the learner. The distance learning, before all, opens the doors of education for students unable (from various reasons) to travel and follow face-to-face lectures or exams on a specific geographical location where the Faculty or its disperse office is. At the same time, the age limit of expected students (and learners in the broader sense), now is situated in a much wider range, very much out of 18-20 year old traditional student. This complies with the life-long characteristic of the learning process. The possibility to learn from distance, offers profiling of professions and expertise to working professionals who have been previously limited (by financial or other reasons) for their further education with multidisciplinary dimension. When learning after certain working experience, there are unexpected fruits of a new, more feasible and pragmatic way of viewing, perceiving and thinking about certain issues and problems, which add new quality to the theory. Finally, the purpose and the goal of education, when practicing distance learning model, shifts from getting a diploma and status, to getting a skill and knowledge applicable in practice as soon as possible, if not immediately. In this context, the reputation of the diploma or certificate has meaning more in terms of the quality and reference of the degree, rather than status symbol.

With the model of distance learning, a maximal interrelation among Faculties, Students and Businesses is being achieved. Their mutual relation paths are used for exchange and upgrade of knowledge, experience and theory, and in order for that to function in a proper way, there must be a certain shift in the centers of gravity of the educational programs. Here is brief elaboration of those particulars.

Student contribution: In programs such as this one, the knowledge gets rapid development and acceleration, because of the fact that students and learners of these programs are very often already employed persons, with some life and working experience, and have a lot of practical, theoretical and experiential understanding of the matters, and that is their special input in the educational process, compared to traditional students who have recently finished high school.

Instructor motivation: The threshold of the knowledge and skills expected by the professors and instructors is moving higher, in order to be able to survive in a system of multichannel creation and distribution of knowledge. The lecturers should have technical knowledge, new instructional design skills, ability to revise and adapt the materials to the goal of having lessons that are sufficient for a learner to learn by individual reading and engagement, as well as creating additional activities and practices that will support the goal of learning. This is far bigger challenge than the one in the traditional learning, where the lecturer presents the topics in the same way for all learners, in front of a wide audience of students, with generally one-way of communication, and assessing what has been learned

by grading the student's reproduction of the same material. In this model, the instructors are motivated to teach the learners a successful applicable skill, not just quotation of theory.

Student motivation: The motivation of the students in distance learning cases is the basic precondition for the whole process's success. Usually, the learner's profile is such that they are proactive to get engaged in a learning process, in order to obtain knowledge or skill, applicable in practice. This denotes that there is a high dosage of self-motivation. On the other side, the qualities coming to surface in this model are more than ever self-initiative, self-organization and self-discipline. In terms of busy schedules and long deadlines for task completion, anyone can be caught in the trap of procrastination and not meeting the expectations. The idea of "active learning" is underlined in every successful outcome of a distance learning program, delivered in various ways. "This is where the active teaching methods are strongly recommended. Experiments were based on innovative approaches to active learning, among others: problem-based learning, project-based learning and the case method."[2]. That is why, the above mentioned qualities, aside from the precondition of some technical knowledge, are the basic requirements for the learner in a distance learning skill set program to maintain his/her motivation on the expected level.

Multidimensional assessment of the learners: The traditional teaching model embodies lectures, seminar papers, home works, colloquia and exams, where the information flow is generally one-way (from professor towards student, in the case of lectures; or from student towards professor, in the case of assignments and exams. The distance learning model imposes engagements and situational environments where the learner is being placed, and where he/she should manifest various skills and knowledge, and learns at the same time. In this context, the student is encouraged to practice with tests, queries, quizzes, to step into communication and collaboration with colleagues, peers, and experts from the whole world, join actively the social, professional or other networks that exist on Internet, to participate in forums, discussions, wikis or glossaries, to collaborate on different projects (learning by doing) and to frequently use virtual data repositories, which are areas of additional (if not main) obtainment of knowledge and skills needed for the real world and business, and points of assessment by the instructors.

Virtual reality of the participants in the educational process: The environment in the traditional education is physical and situational – at the given moment, the professors and learners are in the same physical location. In the distance learning concept, neither the moment in time, nor the geographical position need to be the same. The term "distance" learning itself, specifies geographical displacement of the participants in the process, in the very least between the students and the instructors, and generally among the students themselves; while the term "asynchronous" learning means mismatch of the moment of creation of the lectures and their consumption from the students. This brings numerous advantages and setbacks at the same time. Usually, the same advantage for ones is disadvantage for others, so they should be observed in terms of student segmentation. With the "synchronous" distance learning, there is some similarity with the traditional "face-to-face" learning because the moment of lecturing and listening is the same, only there is geographical distance among the participants. Furthermore in the virtual surrounding, there is no prejudice or reluctance to speak, to compete in oratory skills in front of big auditorium, but everyone is in a virtual world, one can say with "virtual avatar" that is being presented before the colleagues and instructors with engagement in activities, texts, communication, contribution in projects, tutoring and assistance to peers, without the barriers of physical presence and interaction. The virtual environment rapidly enhances a person's technical skills of all involved parties, and also the skills of exploration, collaboration and communication, because everybody has the chance to participate in a discussion, given prior time for thinking, to place a comment or response, to explore the wisdom of the world accessible on the finger tips, and then estimate what of the gathered information is relevant for the topic and what is

not, for his/her specific requirements, and generate process of analysis and decision making upon it.

Communication between professor and student: In general case, the lectures mean one way communication from professors towards students, and in case of exams, it is vice versa. With the distance learning, the way of learning itself, by generating tasks and opinions, imposes questions with the student, and with that comes the need to communicate more frequently and more in-depth with the professors. This means that the professors are expected to be accessible throughout the whole day and night, or in other words, much longer than in the traditional educational system. This way of communication is enabled by direct communication tools such as e-mail, chat, forums etc.

Student **communication and collaboration:** In distance learning, the social, professional, and in this context, the academic networks have strong and sustainable communication links. In the case of traditional education, the students discuss certain related topics prior to the lectures, very rarely during the lectures, or out of the physical space of the educational institution. In distance learning, the communication and collaboration among the students is of immense importance for each one of them, because through sharing the way of thinking and viewing, and discussing viewpoints, is the way things work more effectively for all parties involved. This aspect gains momentum in the hybrid learning, where the students are placed in a setup that enables them to meet each other and the professors, on a few "face-to-face" occasions, and do most of the work on distance. "Individuals continually create and share information according to their interest and get into conversation. They collaboratively craft and animate an innovative "live" space in which they actively participate. Collaboration is the base of Web 2.0 (and e-learning 2.0) technologies."[3]

Networking: communication within social, professional, educational, academic, expert networks: The new "player" in the distance learning system, is the network. Actually, this means everyone present on the internet, as well as all the accumulated knowledge and wisdom from the world, accessible via internet. Every person stepping on the internet, eventually gets skilled in differentiating relevant from irrelevant or truthful from untruthful information coming from friends, colleagues, or experts, or total strangers. As the individual is learning from the networks, the same process is going on in the different direction. "In utilizing various pedagogical models in designing e-learning, however, e-learning leads to pedagogical reengineering, resulting in learning scenarios where students become more active participants in generating new knowledge. They refer to this as a participatory and contributions-oriented approach to learning." [4] Due to every person's communication with a network, every member is learning more too, at that moment, or in the future, when referring to past discussions, projects, repositories and other resources available on the internet.

Simulations and games: Until recently, the power of simulations and games in the process of studying has been, mildly said underestimated. With the development of software applications that can be used for these purposes, there is no limit for application of the creativity of the designers of learning blocks. In fact, the limit is their own creativity in terms of instructional design. From psychological point of view, when a person is playing a game or viewing or making a simulation, he/she is activating his/her brain in regions quite different than the regions used to store information while listening or reading a lecture. That is why, things observed or performed in a simulation or a game, have long lasting impact in the learning process and the memory of the person. Nicely structured game or simulation is one of the essential elements of distance learning. In addition to this, the games and the simulations in this context do not only mean visualization, but also textual quizzes, branching of activities and conditional activities in the learning process.

Virtual Information Repositories: The libraries have been seldom visited places, and when books have been read, this was done by hard working enthusiastic students, which

can't be generalized for the general student population. The virtual data repositories are inevitably accessible by anyone using the internet, with no reluctance, except maybe, due to financial reasons.

Helpdesk: In order to result with success, any distance learning project needs to have remote assistance and prompt troubleshooting of the every-day technical aspects of using the software, accessing resources and other handling needed by the users in order to put aside the technicality and focus on the learning process. This is the reason why any institution needs to keep and monitor log files, error logs, and be proactive in their solving, even before the end users realize their existence.

Mentors, Tutors, Study Partners: In the virtual environment, it is good to offer the student traditional support from real individuals, such as colleagues, assistants, senior colleagues and others who might assist the studying process and advance the knowledge on their side as well. This is world practice in the traditional educational systems, and has even more extensive role in the virtual world, where still real people are the main actors.

Minimum vital blocks for the success of distance learning are: motivation of all parties involved, engagement, hard-work, well comprised presentations, learning materials, exercises, tests, references, existence of final project, follow-up, reminders, deadlines. All of these lead to well used, effective and efficient thirty interactive minutes on a computer, confirming the benefit of withholding travel expenses, daily allowances, work absences, exclusivity in receiving trainings, and all other reasons distance learning is appropriate for business people and employees.

Expectations from the process: The employers are eager to feel the success of the trainings, sometimes even more than the employees. The expectations from this kind of trainings are not for a diploma or reputation, but for theoretical + practical education whose effects can be implemented at the workplace immediately.

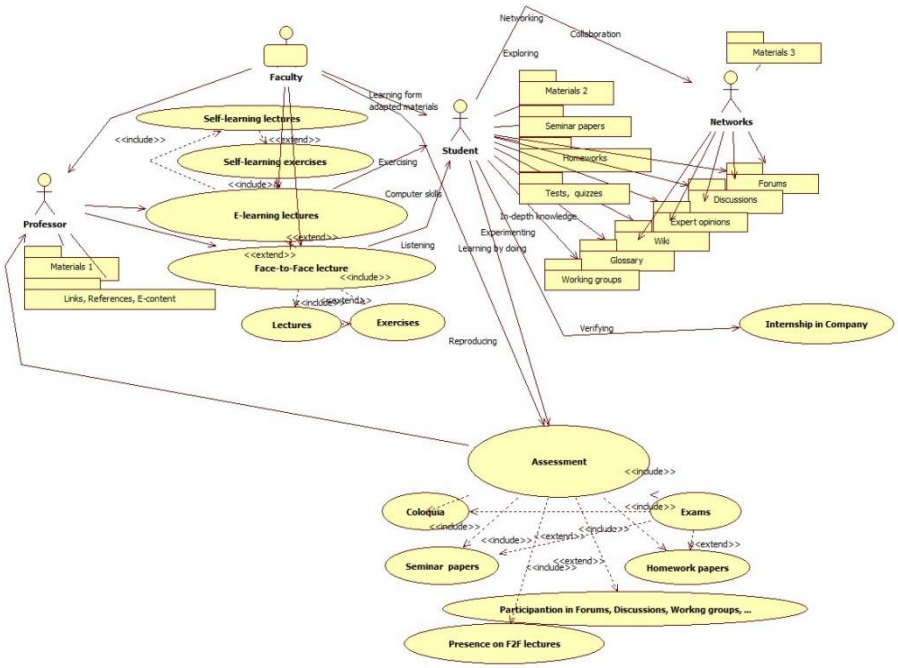


Figure 2. Functional Model of Distance/Hybrid Learning

The functional model of distance and hybrid learning, presented in Figure 2, describes the mechanism of the educational process starting from Professors preparing lectures and practical exercises by revising the materials and shaping them into optimized, self-learning form, as well as performing Face-to-Face, more or less, traditional lectures, and sharing those materials, references and links with the students via e-Learning system. The student is involved in self-learning, practicing, exploring, preparing Seminar thesis and Homework but also joining Discussions, Forums, Wikis, Glossaries, Repositories, Working groups, with the "new player" in today's educational process – the Networks (of professionals, friends, peers, experts, …). The range of skills obtained by the student placed in this setup is much more wider than the traditional, and gets represented in the process of student assessment, where not only reproduction of the learned material is being graded, but also participation in the above mentioned activities.

2.1 Distance Learning with e-Learning for Business Environment

When implementing or using e-learning systems in business environment there has to be a reference to the particulars that are differentiating the business-situated e-learning from the ones in the educational institutions.

For a distance learning program placed on e-learning platform to have a chance of success, what is needed is management support. The management should have comprehensive understanding of the idea, goal, the way the system is functioning, in order to provide the employees with the necessary support, paid attention and time consideration. The traditional way of making trainings means traveling, absences, replacements, exclusivity to receive training to a narrow group of employees, and many other setbacks that are bridged with the distance learning via e-learning system.

Another important component is alignment of the goals of the trainings, topics, materials and lectures in the e-learning system with the goals of the business and the specific employees using it. Whenever one talks about business, it is understandable that the base of the investments is its return, in this case, return on learning. The expected effect can be seen in increased productivity, competency, advanced skills, compulsory training deliverance with least expenses per employee, increased ICT skills, faster and better integration of new employees etc.

The motivation of the employees to learn is important moment also. If not motivated, employees will procrastinate, find excuses, will have reluctance, dissatisfaction and sense of failure.

Approaches of usage of e-learning in the workplace are:

- "e-learning tightly connected to the personal learning goals
- e-learning tightly connected to action projects."[5]

In the first case, the e-learning materials are tightly connected to the personal learning goals, and are used primarily for developing specific competences, introduction to a new employee, mentoring, coaching, collaboration, strengthened communication. However, these systems are used by a limited number of employees (e.g. newcomers, trainees, people changing working places etc.). The second model of e-learning in business environment is connected to action projects. In this case, the training needs to be completed with immediate application of the learned in practice. Here, beside the learners, mentors and instructors, is notable the existence of a project, as additional motivation (e.g. migration to a new software, radical change of a procedure, newcomer training, …) In these systems, the instructional design is enriched with tables, procedures, computer Print Screens, forms that need to be filled, all of which enable certain bigger control in the flow of learning, and its completion.

3 Specific Skill Set Training for Working Professionals by Faculties

Acquiring a skill instead of in-depth knowledge of certain topic (Skill Set Orientation): One displacement of the traditional methods of education is that in this certification model, the aim of the instructional design and the entire pedagogy and work of the learning blocks, is towards obtaining a certain skill, which is a combination of several, not necessarily related skills and knowledge. This is different from the traditional educational approach of obtaining in-depth knowledge of a certain scientific discipline.

In our model, the goal is to acquire a skill of analyzing the financial risk by an entrepreneur, without the in-depth background of entire scientific areas such as Risk Management, Financial Analysis, Calculation Software, Financial Reporting, Forecasting, Cash flow Analysis, Advanced Excel etc. In this way, the professors and the instructors put delicate and complex effort in designing the materials, exercises, tests and quizzes, and lectures in general, in a pragmatic way with ultimate goal of absorbing the optimized theory and sublimated practice and attaining a specific skill, possible to be implemented immediately after the completion of the training.

Professors collaborate and merge knowledge for student and training purposes, instead of students learning sole disciplines and expecting to be skilled (Student-oriented): The second difference of this model, compared to the traditional education, is interrelated with the previous one, and means cooperation of professors/instructors from different areas of expertise, towards a multidisciplinary training, in which the student is not interested so much what is the theoretical home of a certain knowledge, but is trying to learn the compilation of the lectures. This method provides knowledge in the form of "tip-of-the-ice-berg", which is segmented and not profound, but yet, covers the "need-to-know" basis and is designed by respective professors. With this approach, the material itself is a well of knowledge for the skill, and the student's side of the work is to try and implement it in real practice.

The specific example that we will try to develop over the distance learning model is training for qualification with earned certificate, that can be applicable in Faculties or Specialized Training Centers, and the skill set to be obtained from the training will be Financial Risk Analysis for Entrepreneurs.

Every entrepreneur has strong need of making decisions based on the financial performance of the company, viewed through its financial statements. These decisions often are accompanied by the ability to foresee and estimate the risk of the company, from third party, from the competition, from the banking system, the currency risk and many others. Some of these risks can be measured and seen in the financial statements of the company. When a business is asking for a loan in a Bank, the Bank is trying to estimate the risk for default of the loan, or collateral market according the financial statements of the applicant, which describe the past, as well as from cash flow and other projections. The banks do this by involving various profiles of employees such as: credit experts, risk analysis experts, collateral estimation, forecasting, ratios, business analysts, client relationship managers, references as well as statistics and reports from external institutions such as NBRM, Central Registry, Cadastre of Land and Property and others. Then, based on all these inputs, a Credit Committee Decision is made, where both sides of each case are reviewed, the business side and the risk side, along with the aspects of loan amount, purpose, ability to repay, and others.

The idea of the proposal for skill set training, is to equip the entrepreneur too, with some ability to overview their business, in a similar way that a bank does, so that even prior asking for a loan, or making a major step in the business, the person in charge can have good understanding of the company's strengths and weaknesses, numbers and capacity. With this step, the decisions have bigger chances of being successful, applicable and sustainable,

because the businessmen are those who know their business the best, and in fact, if they had the skill to make their own financial risk assessment, or at least become aware of what is to be analyzed, the right people would have the tool to use in the real world. This kind of training does not mean entire Faculty in Economics, Finance, Auditing or other, but a compilation of knowledge created by instructors, that functions towards independent financial risk analysis.

The particular areas that need to be incorporated in Specific Skill Set Training Financial Risk Analysis for Entrepreneurs are:

- Financial statements, basic entries and logic
- Auditing
- Collateral estimation
- Past and future cash flow
- Forecasting
- Cross-checking
- Advanced Excel
- Business software
- Risk assessment
- Financial institutions in the country
- Institutions and Credit Lines for development countries
- Basic entrepreneurship
- Mathematical calculations
- Basic financial terms, ratios and calculations (Interest, ROI, ROE, …)
- Capitalization
- Third party risk, Currency risk, …
- Customs and tariffs
- Decision making
- Customer mentality
- Basic computer skills

Only a brief overview of all these different elements that entrepreneurs need to be familiar with for successful financial risk analysis of the company, will draw the conclusion that in such educational process, that is not Business Management Faculty, but only limited training for skill set, there can be only building blocks on a need-to-know basis from the fields of economy, law, management, business, sociology, mathematics, statistics, informatics, quantitative methods etc. However, the skill set obtained, will be more than simple sum of all these areas, but a sublimation assisted by experts from theory, such as the professors, and experts from practice, such as lecturers, as well as various experts on the side of the students.

4 Conclusion

The few words describing the traditional educational systems are: presence, Face-to-Face lectures, reproduction of learned materials, a lot of theory while studying, almost insignificant practice, the student is the one who should gather all the knowledge and try to implement it at work, the student communicates with professors - rarely, with peers – occasionally, and last but not least, libraries are not “the place to be in” for most of the students. The new web era adds expansions to these words: presence is not necessary, Face-to-Face can happen, but not necessarily, reproduction and learning-by-doing, the professors gather the knowledge and combine it with simulations, examples, quizzes and then present it to students, the student communicates with everyone – frequently, and has another “player” the professional, social, academic and other networks, with whose assistance, the knowledge gets recorded and developed with incredible pace. The companies are attracted to the edu-

cational center of gravity, with the offer of flexible, affordable, grounded trainings with immediately applicable skills. And still, it is not a matter of shortcuts and omissions, but condensed, precisely built instructions for learners, so that the whole educational process has redefined goal and approach to its fulfillment.

Observed from the aspect of the society, even though the creation and production of learning materials and building blocks in the e-Learning systems and software are now market goods, not just exclusivity of the educational institutions, the Governments and the State Authorities, have the obligation to be far-sighted and to enable the life of distance learning and e-Learning systems in the educational institutions and businesses, since they are to be guarding the social aspect of the education. The lifelong learning combined with accessible education for as many citizens as possible, are the necessary preconditions for the future of noble definitions and theories to be put into practice. "Widening participation in education is seen by many as a means of including those who have hitherto been excluded from many of the benefits of modern society. "Education for all" is viewed as an imperative for world security, as an unconnected population suffering high unemployment leads to instability. Education, skills, ethics, and values lead to responsible citizens; an educated and competent people are the essential foundation for democratic societies and market economies" [8]

References

1. Eftimie Raluca Cristina, Avram Emanuela, Tufan Adriana: Educational Innovation and Consumer Behavior - A Study of Students' Perceptions on the use of e-Learning in Class. Annals of the University of Oradea Economic Science Series. M31 I25, 736-740 (2010)
2. Nerguizian, Vahé, Mhiri, Radhi & Saad, Maarouf: Active e-Learning Approach for e-Business. International Journal of e-Business Management. Vol. 5 No. 1, DOI 10.3316/IJEBM0501048, 48-60 (2010)
3. Sílvia Ferrão, Ramón Galván, Susana Rodrigues: e-Knowledge, e-Learning towards e-Competence – The Development of a Model that Illustrates the Acquisition of Competences on Virtual Learning Environments, Proceedings of the European Conference on Intellectual Capital, 200-209 (2010)
4. Doo Hun Lima, David Ripley, Billy O'Steen: E-learning methodologies in practice: similarities and differences between North American countries and New Zealand, Human Resource Development International, Vol. 12 No. 2, 209–224 (2009)
5. Jean Adams: A Four-Level Model for Integrating Work and e-Learning to Develop Soft Skills and Improve Job Performance, IUP Journal of Soft Skills, Volume IV Number 4, 48-68 (2010)
6. Nikolaos Antoniadis, Dimitrios Konetas: Correlation Between Awareness of Blended Learning Techniques and Participation Rate in E-Learning: A Case Study, International Journal of Advanced Corporate Learning (iJAC), Volume 4 Issue 3, 5-9 (2011)
7. Mirjana Radović Marković: Education through e-Learning: Case of Serbia, Journal of Business Economics and Management, DOI: 10.3846/1611-1699.2009.10.313-319, 313-319 (2009)
8. Julian Sims, Richard Vidgen, Philip Powell: E-Learning and the Digital Divide: Perpetuating Cultural and Socio-Economic Elitism in Higher Education, Communications of the Association for Information Systems CAIS, Volume 22, 429-442 (2008)

Data Mining Application for Real Estate Valuation in the city of Skopje

Zoran Gacovski, Josip Kolic, Rosica Dukova, Marko Markovski

FON University, Bul. Vojvodina, bb, Skopje, Macedonia
{zoran.gacovski,josip.kolic, rosica.dukova,
marko.markovski}@fon.mk

Abstract. In this paper we present an application of data mining techniques in order to make price prediction of the real estate properties in the city of Skopje. The current research on the real estate data is insufficient, resulting in miss-understandings between the key players - local government, construction companies, real estate agencies and the potential clients. A dataset from over 1000 transactions in the past three years was used. Five variables (attributes) for each apartment were taken into consideration. We have used SQL Server database and Microsoft Business Intelligence tool (three different algorithms – decision trees, neural networks and logistic regression) in order to perform the price prediction. Also, it can be useful for the prediction of future trends in the urban development of city of Skopje.

Keywords: Data mining techniques, Real estate property value, Microsoft Business Intelligence.

1 Introduction

This paper examines the factors that determine housing prices in a sample of over 1000 home sales in Skopje's region during the period of 2009-2011. Our analysis can be used in real-estate Agencies, and can help the potential buyers to estimate whether the property price is in accordance with the existing market trends.

This paper is organized as follows: in Section 2 – a problem of price prediction is described and data mining algorithms (neural networks, decision trees and logistic regression) are presented. In Section 3- an overview of the Business intelligence software is given, and in Section 4 – simulation results are presented.

2 Problem description and algorithms

The problem we have investigated is the real estate market in Macedonia, particularly - apartment sales in Skopje. According to available dataset for about 1000 transactions from the previous three years (2009-2011), and by applying the Business Intelligence and data mining, the task is to predict the price of an apartment with known attributes (characteristics).

For population of apartment sales database (training set) we have used data from Macedonian real-estate agencies for transactions from the past 3 years. For each apartment we have used the following data: suburb (settlement) of the apartment, quadrature/ surface area (in m^2), floor of the apartment, number of rooms, heating

(central, electricity, fossils, etc.) and apartment price (in Euros). Finally, our database consisted of 1200 apartments, which was the training set for our predictions.

We have used Business Intelligence Studio from Visual Studio 2008. For solution of our problem (price prediction) we have applied 3 different techniques: Decision Trees, Neural Network and Logistic Regression.

After creation of the data mining model, it can be evaluated in Business Intelligence Studio, in order to make the predictions for the price (for each apartment within the table). We have shown the predicted prices graphically.

3 Overview of the Business Intelligence tool

The Business Intelligence software offers useful tools which can be applied in the process of strategic planning and management in the companies. This software enables the companies to discover their critical operations via different reporting and analyzing tools. BI deliverables can incorporate different components, like tabular reports, shared lists, diagrams and graphs. Although traditional BI systems were developed via host terminals and printed reports, the current development of BI applications is performed via Web (Internet). It is also possible to develop interactive BI applications, which are optimized for mobile devices, smart-phones and e-mail usage.

The BI environment is well integrated in Microsoft Visual Studio.NET, in order to enable faster development of BI applications. The data mining project developed in a BI environment is known as a solution.

4 Simulation Results

In our research – we have investigated and simulated the problem of price prediction – with the three algorithms offered in Microsoft Business Intelligence. Bellow, the diagrams for comparison of the apartment selling prices with the predicted prices obtained by these algorithms, are given.

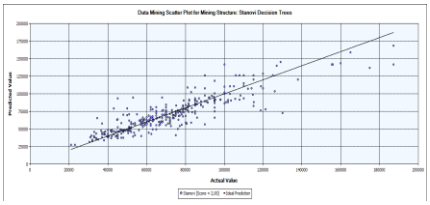


Fig. 1. Diagram of predicted prices by Decision Trees algorithm.

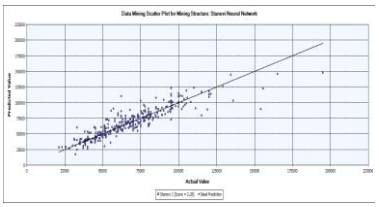


Fig. 2. Diagram of selling/predicted prices obtained by Neural Network.

5 Conclusion

In this paper we have used dataset from over 1000 transactions in the past three years. Five variables (attributes) for each apartment were taken into consideration. We have used SQL Server database and Microsoft Business Intelligence tool (three different algorithms – decision trees, neural networks and logistic regression) in order to perform the price prediction. Our analysis shows that logistic regression algorithm gives better (closer) prediction of the home prices than other two algorithms.

The Ethics of Artificial Intelligence

Mersiha Ismajloska[1], Jane Bakreski[1]

[1]University of Information Science and Technology "St.Paul the Apostle", Ohrid, R.Macedonia
mersiha.ismajloska@uist.edu.mk, jane.bakreski@uist.edu.mk

Abstract. The issue of the ethic of artificial intelligence came, maybe as the most important for the future development of AI, according to the thesis which concern humanity, humans and their interaction in the world, especially the world of technique. The issue connects philosophy and technique and therefore is present not only in scientific journal, but in SF literature, visual arts that often interferes with movie industry and of course with the principles of aesthetics. Starting from the three Asimov laws of robotic, describe in his book I, Robot, 1950s, we can develop a study for the ethic of artificial intelligence, always having on mind that subject is complex and changeable during time.

Keywords: ethic, artificial intelligence, aesthetic, laws of robotic

1 Introduction

To define and explain the subject of the ethics of artificial intelligence first we must describe and analyze the three mentioned Asimov's laws. First of them reads: *A robot may not injure a human being or through inaction, allow a human being to come to harm.* This law is in logical interaction with the others two: *A robot must obey orders given by human beings except where such orders would conflict with the First Law* and *A robot must protect its own existence as long as such protection does not conflict with the First or Second Laws.* That is a continuation of a humanistic line of philosophy present since ancient times. The way of thinking is similar, but the technique has changed during time.

2 Homo sapiens / AI

Implementation of the three Asimov laws in the field of robotic and philosophy put the human in the center similar to Leonardo da Vinci, Vitruvian Man, and we can say that the humanism as principle is present in science abstraction and in some way we can talk about science philosophy. The romanticism idea of the man as the center of Universe found its re-actualization in robotic in a very original

way. First of all physical similarity of man and robots implies a wish for creating friendly mechanical companions to humans that has of course led to existential and essential consequences provocative for human thought and way of leaving.

As it is said in *Artificial Intelligence A Modern Approach:* we call ourselves Homo sapiens-man the wise-because our intelligence is so important to us. For thousands of years, we have tried to understand the thinking process; that is, how a mere handful of matter can perceive, understand, predict, and manipulate a world far larger and more complicated than itself. The field of artificial intelligence, or AI, goes further still: it attempts not just to understand but also to build intelligent entities. The most important questions is related to interaction of humans and robots and almost all variations as part of narration of Asimov *I Robot. I Robot,* originally play with the identity, starting with the title, and connecting human identity with that of a robot, and speaking of robot identity we are getting closer to the question of consciousness that directly leads to field of ethics. The impossibility of the robot to get consciousness about its existence, in the real science world, give to science fiction provocation to make creative combination with possibility of interaction of robot and men. Having in mind SF literature and the presence of mechanization in some visual art works give us the privilege to speak with terms of imagination when explaining of machine-human relation(ship). That connection is an abstract way of speaking about human ethic, which originated from the first moment when man become conscious about himself.

3 Definitions and possibilities

In that manner of speaking machine ethics is the field of research concerned with designing Artificial Moral Agents (AMAs), robots or artificially intelligent computers that behave morally or as though moral. Asimov proposed the Three Laws of Robotics to govern artificially intelligent systems. His intelligent manipulation with boundaries of his three laws to see where they would break down, or they would create paradoxical or unanticipated behavior create interesting play with Susan Calvin, Alfred Lanning, Robbie, Stephen… His work suggests that no set of fixed laws can sufficiently anticipate all possible circumstances. That's the way science give coordinates to artistic thought and every artistic thought starts from scientific attempt to make precise definition of the issue. Of course art is playing with definitions, but always having in mind that definitions exist. In the case of Artificial Intelligence according to *Artificial Intelligence A Modern Approach,* the starting point is that it can't be said what it is AI in first place. Book suggests two ways of making a definition, one with *though processes* and *reasoning,* and the other with *behavior.* For example:

Thinking Humanly: The exciting new effort to make computers think…*machines with minds,* in the full and literal sense. (Haugeland, 1985)

Acting Humanly: The art of creating machines that perform functions that require intelligence when performed by people. (Kurzweil, 1990)

Thinking Rationally: The study of mental faculties through the use of computational models. (Charniak and McDermott, 1985)

Acting Rationally: Computational Intelligence is the study of the design of intelligent agents (Poole, *et al.,* 1998)

(*Artificial Intelligence A Modern Approach*, page 3)

Implementation of this kind of definition in art brings speculations about rational and human making question about identifying rational with human and rational with machines. Identifying rational-human identifying go principles of ethic arise, but with rational-machines identifying principles of ethic must be carefully input in the manner of robots making. Asimov creates an interesting concept of that issue, and fact that stories originally appeared in the American magazines *Super Science Stories* and *Astounding Science Fiction* between 1940 and 1950. The stories are spoken with Dr. Susan Calvin voice and a narrator who re-tells them is a reporter in 21 century, ironically, this current century. Though the stories can be read separately, they share a theme of the interaction of humans, robots and morality, and when combined they tell a larger story of Asimov fictional history of robotics.

4 Interpretation of ethics and AI through art

Visual interpretation of that issue and the three laws of robotic came with Alex Proyas film, too, adaptation of Asimov book. The book has been adapted in a movie in Hollywood manner in 2035 year when, techno-phobic cop investigates a crime that may have been perpetrated by a robot, which leads to a larger threat to humanity. Speaking in terms of aesthetics we can say that this time Hollywood made not only a popular film which is expected for Hollywood, but at same time original visual work where the original Asimov message is sent to viewers. Here, we can talk about the presence of machines and AI in the aesthetic world with all consequences which provoke artists to use that symbiosis in their work. One of them of course is Hans Rudolf "Ruedi" Giger, Swiss surrealist. His initiation was meting with Salvador Dalí, to whom he was introduced by painter Robert Venosa. This introduction to Giger artistic work lead to re-definition of surrealism and connecting him with other artistic works where presence of technique (machines or robots) is more or less obvious. Questions of technique thinking, leads again to *Artificial Intelligence A Modern Approach,* and subject of thinking humanly i.e. the cognitive modeling approach. Section speaks about human thinking and technical approach for making a program to think like a human. With a sufficiently precise theory of the mind, it becomes possible to express the theory as a computer program. If the program's input-output behavior matches corresponding human behavior, that is evidence that some of the program's mechanisms could also be operating in humans. But remain question, would be enough humanely ?!

Contradictory to Giger, the art of Macedonian painter Vasko Tashkovski as a whole is created in accordance with surrealist tradition but with a tendency, art experiences from the past to be fed with the most important changes resulting from the modern living. So, this painting enters the sphere of fantasticality and science fiction, as a result, but also as confirmation of his belonging to the modern lifestyle.

S. Markovski, M. Gusev (Editors): ICT Innovations 2012, Web Proceedings, ISSN 1857-7288

In an era of a modern technological and technical civilization, in a time of more numerous, more blurred and less certain relations between man and nature, man and society, man and machine, the artist is gifted to perceive the big self-delusion and confusion by the deep changes and processes, turmoil and tremors – as carriers of fear but, at the same time, hope as well.

Basically, Vasko Tashkovski's art is a synthesized vision of our time with existential themes of modern man (fear from the war machinery, form the dimension of the universe, from biological decay, destruction and pollution, the fated situations of mankind in the chaos of natural and mechanical phenomen) now reshaped into exciting themes like rebellion and revolt, a warning, conscience, or hope in the future.

5 Conclusion

Again way of questioning the thinking and ethics going back to Isak Asimov *I Robot:*

Then you don't remember a world without robots. There was a time when humanity faced the universe alone and without a friend. Now he has creatures to help him; stronger creatures than himself, more faithful, more useful, and absolutely devoted to him. Mankind is no longer alone. Have you ever thought of it that way?"

Giving advantage to ethics of robots, make Asimov an optimistic thinker who also believe that people know what they are doing but we can tell that he is forgetting to critize the expanding mechanization of the world which make people to fell alienation. That specific issue of ethic of artificial intelligence and human alienation is a concept used by Pink Floyd in only in context to warn not about possibilities, but about consequences which are not always predictable. That way constant awareness of humanity in first place makes us better visionaries of the future and better people of the. Having in mind Asimov's words at the end of the book: *And that is all," said Dr. Calvin, rising. "I saw it from the beginning, when the poor robots couldn't speak, to the end, when they stand between mankind and destruction. I will see no more. My life is over. You will see what comes next."(128)* we must share hope for ethic interaction of all subjects, no matter if they are humans or machines. In expectation of what is "coming next" we must be sure that human life always will be highest ethical value and priority in any kind interaction of humans and machines.

References

1. Andova – Foteva Vera. Tashkovski V. Museum of Skopje, Skopje, 2005
2. Asimov, I.: I, Robot, Oxford University Press, 2000
3. Velickovski V. Tashkovski V. Skopje, 1990
4. Russel, S. Norvig P.: Artificial Intelligence A Modern Approach (Third Edition)

Computer aided translation – the cloud approach

Veno Pachovski[1], Eva Blazevska[1]

[1] University American College Skopje, III Proleterska brigada br. 60,
1000 Skopje, Republic of Macedonia
{pacovski, blazevska}@uacs.edu.mk

Abstract. The need for automated translation grows very fast and so is the demand for creating tools to meet it.
The research presented represents a possible semi-automated solution i.e. a concept of cloud aided translation tool based on aligned corpora. Namely, the concept gives an opportunity to translate a text by keeping its structure (paragraph-wise) intact, using all other same language texts as context sources.
The final goal is pairing the server and a user application in such a manner that the server which hosts the corpora (translation pairs of texts) is in communication with an application in which a source document is paired with the one user is working on.

Keywords: Natural language processing, translation, corpora

1 Introduction

The need for automated translation grows very fast and so is the demand for creating tools to meet it. Current situation in EU (60+ languages in Europe, 5 main languages and 23 of other member countries) and its commitment to multilingualism dictates further efforts and investments to be made. Within Horizon 2014-2020 which estimates more than 80 billion Euros, part of ICT and specifically for LT remains to be defined depending on the activity in the field. [1]

Macedonian language faces additional difficulties because of its relatively small population (less than 2 million), its specific grammar as well as alphabet, lacking corpora and basic language processing tools. All these mean that there is a real need for some original research. Also, any automatic solution requires rich context (the richer, the better), which means corpora upon which the translation system could base its decisions. This is an effort to do both.

1.1 The Motivation and Problem Approach

As a result of previous research and techniques and technologies involved ([2][3]), the approach adopted in this project is based on two important requirements.

The first one is that the translations are done paragraph-wise meaning that the translator is required to translate the text by paragraphs. Thus, the basic structure of the text is kept and even if (for various reasons) the sentences are not aligned (within the paragraph), at least the search for similar meaning for a word (or a phrase) i.e. the context (in future) is (will be) limited to that paragraph. Also, while the user is working on a paragraph, the search option is available, so that the user can search for possible translations of a particular word, or similar paragraphs, so that the sense of a word to be translated can be clarified.

The second, as a direct consequence of the first, is that the aligned pairs of translated texts result in an aligned corpora. Here, the working assumption is that as the corpora grow, the context will become richer and WSD when choosing an appropriate word (or phrase) will be easier. In time, the system may become (hopefully) a valuable platform for testing software and algorithms for automatic translations, considering that there will be paired paragraphs in various languages.

Also, considering that the system requires authentication of its users, it preserves the authorship, so the work can be copyright protected

Finally, inspired by the latest approaches to cloud storage (like DropBox), the idea appeared that a local application could be paired with the web server, so that a user can work on text in-directly, using a local application (which will again, in the background, use system resources).

2 Conclusion and future work

The system looks promising as a platform for future research. The first step is to make the web site popular among translating community (or least within University), so that the corpora can grow faster.

Next, the optimal protocol for communication between a desktop application and the web server (the cloud) should be devised, in order to follow the process of translation on the local machine, so that users can create and optimally coordinate their local resources with the cloud application.

References

1. European Commission : CORDIS : FP7 : ICT : Language technologies, http://cordis.europa.eu/fp7/ict/language-technologies/
2. Paskaleva, E., Pacovski V,. Aligning the translations – a possible strategy for creation of aligned corpora (for South-Slavic languages). In: Proceedings of the Conference on Formal Approaches to South Slavic and Balkan Languages, pp. 113--117, (2006)
3. Nakov, P., Pacovski, V., Acquiring “False Friends” from Parallel Corpora: Application to South Slavonic Languages. In: Readings in Multilinguality. Selected Papers from Young Researchers in BIS-21++. Galia Angelova, Kiril Simov, Milena Slavcheva (Editors). Incoma Ltd. Shoumen, Bulgaria. (2006)

Metrics for Service Availability and Service Reliability in Service-oriented Intelligence Information System

Jugoslav Achkoski[1],Vladimir Trajkovik[2]

[1] Military Academy "General Mihailo Apostolski",
str. Vasko Karangelevski bb, 1000 Skopje, Macedonia
jugoslav_ackoski@yahoo.com
2 Faculty of Computer Science and Engeneering
str. Rugjer Boshkovikj 16, P.O. Box 393 1000 Skopje, Macedonia
vladimir.trajkovik@finki.ukim.mk

Abstract. This paper gives contribution in definition of metrics for service reliability and service availability in terms of their usage by the end-user.

Keywords: SOA, metrics, service availability, service reliability, information system, Intelligence, QoS.

1 Introduction

Contemporary Intelligence models should be based on information-commutation systems. Usage of contemporary ICT technology gives opportunity for more effective implementation of Intelligence function in terms of collecting information, planning information, analyzing information and dissemination.

2 Service Availability

Availability is service attribute, whether or not service is active or available after received request by a user.

Presumption that information system or services in certain period of time are founded in one of numerous service states whether or not services are unavailable or available allows implementing Markov' models.

Function for service availability in certain time moment is presented by following equation:

$$A(t) = \left(1 - \frac{\lambda_{iz}\lambda_{jz}}{r_2 r_1}\right) - \frac{\lambda_{iz}\lambda_{jz}}{r_1 - r_2}\left(\frac{e^{r_1 t}}{r_1} - \frac{e^{r_2 t}}{r_2}\right) \quad (1)$$

S. Markovski, M. Gusev (Editors): ICT Innovations 2012, Web Proceedings, ISSN 1857-7288

3 Service Reliability

Function of service reliability represents probability of service processing in certain time interval [0,t]. Intensity when service is not available for using can be presented with constant value λ=const.
Function for service reliability in certain time is presented by following equation:

$$R(t) = \frac{r_1 + \mu_{zi} + \lambda_{zj} + \lambda_{iz}}{(r_1 - r_2)} e^{r_1 t} - \frac{\mu_{zi} + \lambda_{zj} + r_2 + \lambda_{iz}}{(r_1 - r_2)} e^{r_2 t} \quad (2)$$

Common characteristic for previously mentioned services is probability that refers to service availability in certain time during Intelligence operation. Also, zones (green, yellow, red) for determining functions of probability can be introduced (see figure 1).

Service availability	Zones
service is available for using	
service can be available for using	
service cannot be available for using	

Fig. 1. Service availability that is related to appropriate zones

Probability value of service availability allows selecting services that can be exploited in certain Intelligence operation in certain time. Introduced zones contribute to select services that can respond on the most appropriate manner.

4 Conclusion

In this paper we present estimation of probability for services in certain time moment by determination of service reliability and service availability.

References

1. Maheswari, S., Karpagam G., R.: QoS Based Efficient Web Service Selection. European Journal of Scientific Research, vol. 66, pp: 428-440 (2011)
2. Ramović, M. R.: Skripta - Pouzdanost sistema elektronskih, telekomunikacionih i informacionih, Katedra za Mikroelektroniku I tehnicku fiziku, Univerzitet u Beogradu, Elektrotehnicki fakultet (2005)

Integration of EuroGeoss Applications to Enhance the Research Methods in the Region

Sanja Stefanova , Marina Ivanova, Igor Stojanovic, Zoran Zdravev
Goce Delcev University – Shtip

{sanja.stefanova, marina.ivanova, igor.stojanovik, zoran.zdravev}@ugd.edu.mk

Abstract. This paper represents research related to the climate and climate changes in Macedonia and worldwide. The paper described climate events in the past and today, emphasizing the efforts of researchers and institutions dealing with such climatic changes and disasters. Macedonia, as a developing country has not yet achieved significant results related to the climate, so this study actually represents the introduction and integration of new tools and services in Macedonia which are used by worldwide institutions for research, prediction and reduction of those change. The key points in our research are: to review the available applications and explain the process and methods with which the data will be edited, stored, standardized, multimedia displayed and published. Our main goal is to describe the services and tools to standardize data from surveys and use all opportunities that can be use by these organizations. The creation of such a work in environment, improves the climate picture in Macedonia, and will be driving force of climate researchers and institutions.

Keywords: Climate, Change, Climate organizations, Geoss, Biodiversity, European Union, Applications, Standardization, mapping.

1 Introduction

The climate represents set of meteorological elements in the atmosphere such as temperature, precipitation, wind and others. The climate is very important geographical factor in each state. It affects the amount of precipitation, vegetation, hydrograph and of demographics. On the climate affects: geographical location, relief, atmospheric currents, and in recent times the people through technology and activities. Through the years there have been many climate changes, warming and vulnerabilities that led a change in flora and fauna in some areas, their disappearance up to the human sacrifices and devastation. The consequences and the vulnerabilities of these changes are continuation with global warming, the warming of the Arctic and Antarctica, melting glaciers, changing the seasons, spreading diseases, snow storms, droughts, floods and etc. The world's leading economic forces are making many efforts to tackle with climate change. In addition the leading forces are creating applications, do monitoring on the all events, organize conferences in order to inform the public about the dangers

S. Markovski, M. Gusev (Editors): ICT Innovations 2012, Web Proceedings, ISSN 1857-7288

of climate change. The European Union is the main driving force in this region for the climate change. On the last summit the European Union come forward with 7.2 billion Euros in aid for adaptation to developing countries to new technologies.

2 National platforms and strategies to reduce climate change

Republic of Macedonia is characterized by a striking manifestation of history and modern destructive processes, among other things resulted in occurrences of intense seismic activity in many areas and regions. The floods are natural disasters that often occupy the territory of the Republic of Macedonia and they result from the particularities of relief, topographic, geomorphologic and climatic conditions, and unequal regime of flow of natural watercourses.

One of the major problems that facing our country are losses in biodiversity. Although this sector depends on several segments, however, the key factors that give a sign of declining biodiversity in Macedonia are: anthropogenic land use in the past in general, recent economic collapse, inadequate spatial planning and inappropriate land use. Just assumed it would come losing and disturbing the ecology of different species. Until now are produced several significant projects that emerged significant documents to preserve biodiversity here. Although these climate changes are not treated directly, however they highlight their importance. The Government of the Republic of Macedonia has implemented several action plans consisting of a sequence of necessary actions to mitigate the negative consequences of climate change.

3 Strategies and plans to reduce climate change in Republic of Macedonia

Our main objective in this research is to describe the biodiversity in Macedonia and globally and finally offer solution to long-term problems in our region. The solution is a combination of already available applications and would give a draft program of which will be designated for our researchers. We analyzed the results and applications, and in this article we have described their functionality and ability as they use them and upgrade. These applications give us the opportunity to update their content and customization depending on the region. Most of the functionality would have taken the applications and would complement the programming part that will be only for researchers in Macedonia region. So they would set up according to the requirements and needs of the most vulnerable areas in Macedonia. . From the very beginning I adopted standard professional terminology that will be written in English language. During the program would add a map with the territory of the Republic of Macedonia, which will be specially prepared the draft of the territory, and the coordinates of the regions. From the research that we have done so far, we came to the conclusion that you need to build a store that would be an archive of previous research papers and storage of future results. Researchers will be able on Macedonian language to perform data standardization. If they want their data to be accessible in the world,

they will synchronize in the main application, as described below, and if they meet the required standards will be published in the pages of international organizations. The next section will give a brief description of the functionalities of applications and their global application.

3.1 Guidelines for creating and updating metadata in the drought metadata catalogue

One of the applications created by EuroGeoss is CatMDedit, and its main goal is to create and update the metadata. This program facilitates the documentation of resources, with particular focus on the description of geographic information. On the internet are available instructions for installation and a description of the main functionalities of the tools. In this section we will try to briefly give a description of this application to discern its strengths and to conclude that the benefits of application, how applicable is in our situation, whether to enable researchers new concept of work that will move to world institutions.

When metadata is created there is an opportunity for them to be added to the appropriate catalog and how the export of metadata as XML files in accordance with ISO 19139 specifications for coding. If metadata is properly updated, they are exported as XML file in alignment with the standards. If you want to delete given records with confirmation of your verification you can remove the selections. EuroGeoss and CatMDEdit have created a template and instructions for using the application, where are described in detail all steps and procedures of the functionalities of the system. The last upgrade of the application was made 8 months ago and it has the ability to upgrade with an appropriate vocabulary that is simple and can be easily integrated with CatMDEdit accommodation of the files in the application directory. Also upgrading GEMET is very simple and enables new privileges and features of CatMDEdit.

3.2 EUOSME: European Open Source Metadata Editor

The European Open Source Metadata Editor (EUOSME) is a web application written in Java and based on Google Web Toolkit (GWT) libraries. More specifically, this implementation allows to describe a spatial data set, a spatial data set series or a spatial data service compliant with the standards ISO 19115:2003 (corrigendum 2003/Cor.1:2006)ii and ISO 19119:2005iii. It is therefore an implementation of the INSPIRE Metadata Technical Guidelines based on these two ISO standards, and published on the INSPIRE web site3. This editor builds on the experience acquired in the development of the INSPIRE Metadata Implementing Rules, and includes the INSPIRE Metadata Validated Service available from the INSPIRE EU Geo-portal (http://www.inspire-geoportal.eu/).

Identification of resource is composed of two elements: law and namespace and annotation data in this field is mandatory. This program is more advanced than previously described and provides detailed tagging of each imported data with the above described functions. In the functionality will enumerate the possibilities for the the

topic categories, resource locator, a tab with keywords, panel with a choice of language etc. The implementation of these two applications of the European Union and GEOSS and any other applications that will be available in future will enable us to move to these global institutions, and that means using the same involves financial, technical and educational assistance.

4 Conclusion

In this research examined part of climate change which for years has been most vulnerable and has suffered major losses. Due to persistent climate change biodiversity faces transformations, losses and even extinction of some species. In Macedonia it is this area's most threatened because the competent institutions have so far neither achieved significant results, nor have conducted monitoring to discover the causes of such phenomena. Our main goal is the presentations to be the start to new work environments in our region, to reduce losses from year to year are bigger and more devastating. Through their application and standardization of research would have moved to global institutions and their work commitments and efforts to overcome and alleviate the situation of different regions of the the planet earth.

References

1. B,Klement: Climate change scenarios in Macedonia
2. H, Richard.,O,Angelica.: Informational Governance of Climate Change Organizations, (2011)
3. Second National Report of the Republic of Macedonia to the UN Framework Convention on Climate Change
4. Second National Plan for climate change, December (2008)
5. Open source code, www.catmdedit.sourceforge.net/
6. The European approach to GEOSS!, http://www.eurogeoss.eu/Documents/D-5.2a_ EuroGEOSS_guidelines_updating_metadata_in_catalogue.pdf
7. Enhancing access to European spatial data, www.inspire-geoportal.eu/
8. Group on earth observations, www.earthobservations.org/geoss.shtml
9. European Commission Joinup, www.osor.eu
10. World Meteorological Organization, www.wmo.int/pages/index_en.html
11. Infrastructure for Spatial information in the European Community, www.inspire.jrc.ec.europa.eu/index.cfm/newsid/10281
12. National Hydro meteorological Service of Republic of Macedonia, www.meteo.gov.mk/
    The World Bank, www.climatechange.worldbank.org/

Personalizing mobile user subscription services using data mining

Aleksandar Karadimce[1], Dijana Capeska Bogatinoska[1]

[1]University of Information Science and Technology "St.Paul the Apostle", Ohrid, R.Macedonia
aleksandar.karadimce@uist.edu.mk,
dijana.c.bogatinoska@uist.edu.mk

Abstract. Mobile companies today offer diversity of user subscription services to their subscribers in order to attract their attention. In order subscribers effectively to use that service they have to choose appropriate subscription service package. Choosing the appropriate subscription service is not always simple decision, using the knowledge discovery process can help customers to make the right choice. We suggest applying different data mining techniques to already aggregated user traffic from different types of services, stored in the data warehouse system. With the extraction of the useful information we can provide offers to subscribers that will be optimal with their used services (voice, SMS, MMS and Internet). Personalization of the user subscription services will contribute to more objective and transparent process of billing the subscribers.

Keywords: Personalized mobile services, data mining, mobile subscribers, Business Intelligence.

1 Introduction

Telecommunication companies, in battle to enlarge their coverage on the market, propose variety of user subscription services to their subscribers in order to attract their attention. When users sign a contract with telecommunication provider for mobile subscription services they have to choose from a predefined tariff models, either pre-paid or postpaid subscribers. These predefined pattern based subscription tariff models contain predefined amount of internet traffic, number of SMS or MMS events and specific amount of time for free voice calls in different mobile zones. Users should be able to use entirely their benefits defined with the subscription tariff model, or they have underutilized use of the benefits provided with the assigned subscription model. Proposed solution offers mobile subscribers appropriately to use the chosen subscription tariff model, providing them with personalized user subscription service.

Choosing the appropriate subscription service is not always simple decision, using the knowledge discovery process can help customers to make the right choice. Customizing the user subscription services according to personal requirements, it asks for knowledge to understand and analyze user data flow behavior. Everyday use of mobile devices generates different type of data traffic that can be used to analyze the user

S. Markovski, M. Gusev (Editors): ICT Innovations 2012, Web Proceedings, ISSN 1857-7288

requirements. The main contribution of this paper is these gathered user data traffic to be reused for personalizing mobile user subscription services that are offered, which will contribute to more objective and transparent process of billing the subscribers.

In section II of this paper we present personalized mobile user subscription services. The Section III gives overview of results from comparison of different data mining techniques. Section IV concludes the paper and presents some future work.

2 Personalized mobile user subscription services

Mobile customers in order to take advantage of the new benefits have to sign new long term loyalty contracts with mobile provider, as a subscriber to only one tariff model. On the other side the mobile provider itself cannot offer customized mobile subscription tariff model to every subscriber, as stated in its request for mobile services. Existing subscribers have to choose the predefined subscription tariff model and state the limit of monthly subscription with its appropriate charging rate plan. This way mobile provider knows subscribers predefined limit for monthly spending for mobile services. Existing research areas have already used the mobile events history records, call detail records (CDR), for marketing and fraud detection applications [1], [3]. Also data mining of large database systems has been a major challenge in the telecommunication companies, such as the survey of different available data mining techniques [2]. There have been different methodologies of data mining used for customer churn prediction based on either demographic features or billing or usage features, as shown in [4].

The importance of personalization is given in [5] that give an overview of issues that must be considered to leverage future technologies that can support more advanced personalization. It also highlights the distinction between customization and personalization, where customization is thought of as user controlled modifications of a service and personalization is machine-controlled. In [6] is given technical sense of profiling that using data mining we have certain degree of probability in order to customize individual decisions.

The process of creation subscriber personal profile enables businesses to provide highly individualized services for their subscribers and targeted advertising for their customers. This way mobile subscription service is more subscriber oriented, provides increased user experience and offers flexibility that the limit will be used more optimally, instead of the current pattern based mobile subscription services. Disadvantage in the current subscription service is that the existence of shorter events fragment the subscription limit, instead the calls that have long duration to be included to the subscription limit. The implementation of personalized mobile subscription services will provide new type of subscription service that will bring revenue increase on long term basis. Instead of applying different kind of promotion packages for the mobile tariff models, we suggest creating subscriber personal profile, where there will be modular discrete packages for different kind of service. This way only one discrete package for specific service can be assigned to a particular subscriber personal profile, which shall decrease the total amount of the user invoice.

3 Comparison of different data mining techniques

The advantage of using the data mining, as technique to extract knowledge, performed on already aggregated data relevant to the particular subscriber. We have done research using four different algorithms. The Microsoft Decision Trees algorithm is a classification algorithm that works well for predictive modeling [7]. On the other hand the Microsoft Clustering algorithm uses iterative techniques to group records from a dataset into clusters containing similar characteristics [7]. The Microsoft Logistic Regression algorithm is a regression algorithm that works well for regression modeling. The Microsoft Neural Network algorithm uses a gradient method to optimize parameters of multilayer networks to predict multiple attributes [7]. That dataset for the research are subscriber previous calls and data events that are being deposited by the mobile provider. Using the Microsoft SQL Server Management Studio we have conducted preparation for data mining research, on dataset sample of 10.000 CDR records, randomly generated and stored in MS SQL Server table called [CDR_traffic].

Especially important is the column Promotion that is populated indirectly using database update procedure, with discrete values 'LONG', 'MEDIUM' and 'SHORT', based on the values recorded in the column Duration. Estimation for the prediction value is saved in column Discount, receives values from 1 to 3 that represent affinity weight for discount. The highest value for discount means it should be applied highest priority distribution of the subscription limit. Microsoft Business intelligence development studio is other tool that was used to perform the task of conducting different data mining techniques. In order to create testing set that will be used in the data mining training, we have assigned 50% of 10.000 CDR records to be reserved for model testing. Using the Microsoft SQL Server Analysis Services Designer the process continued with comparison of the four different data mining models under the name "MINING AGG Traffic Data", see Fig 1.

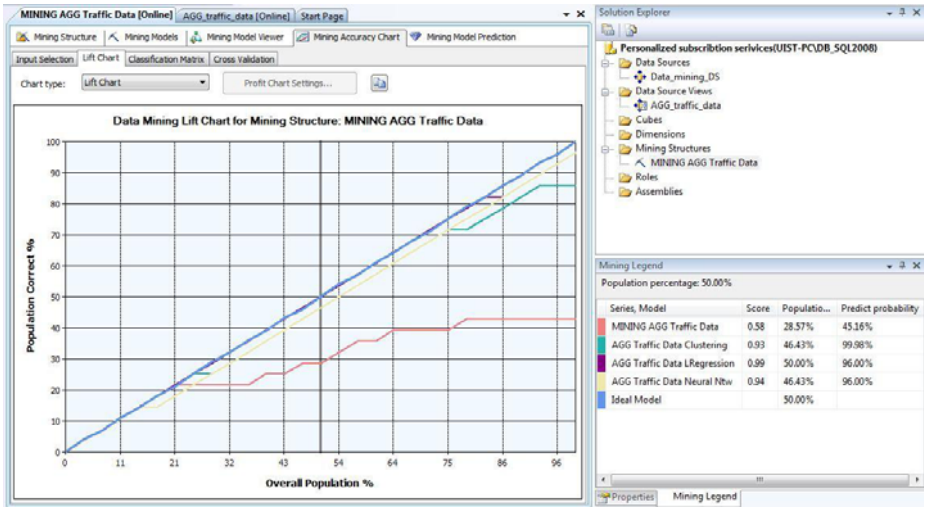


Fig. 1. Comparison of data mining models – lift chart.

Results from the compared four different data mining models in Mining Accuracy Lift Chart have shown that the best results, with score 0.99 closest to Ideal Model, are provided by using the Microsoft Logistic Regression. On the other hand the Microsoft Decision Trees algorithm, with score of 0.58, is not appropriate in the research. Observing the results from this research we can conclude that Microsoft Logistic Regression provides best results in determining the state of the predictable column for continuous and discrete input values.

4 Conclusion and future work

Personalization of the user subscription services means distribution of subscription according to user needs, depending on the type of service, zone or duration of the calls. It will contribute to more objective and transparent distribution of the subscription limit that gives customized charging to the subscribers. Used different data mining techniques to already aggregated user traffic from different types of services, provides new customized and flexible way subscribers to gain optimal charge for their used services that points to user oriented personalized subscription. The relatively small sample of data used in this research has provided results instantly; otherwise it should be considered that in real-time database systems, where there is thousand times more data, it would require more demanding resources. Upcoming research should involve implementation of realistic dataset, to overcome the real-time limitations of resource intensive environment, we suggest using the benefit of cloud computing.

References

1. Weiss G. M.: Data Mining in the Telecommunications Industry. In: O. Maimon and L. Rokach (.eds), *Data Mining and Knowledge Discovery Handbook: A Complete Guide for Practitioners and Researchers*. Kluwer Academic Publishers, 1189-1201. (2005)
2. Chen M.S., Han J., Yu P. S.: Data mining: An Overview from a Database Perspective. In: *IEEE Transactions on knowledge and data engineering*, Vol 8, No 6, Dec 1996. 866- 883. (1996)
3. Qayyum S., Mansoor S., Khalid A., Halim K. Z., Baig A.R..: Fraudulent Call Detection For Mobile Networks. In: International Conference on Information and Emerging Technologies (ICIET). 14-16 June 2010. 1 – 5. (2010)
4. Khan A. A., Jamwal S., Sepehri M.M.: Applying Data Mining to Customer Churn Prediction in an Internet Service Provider. In: International Journal of Computer Applications (0975 – 8887) Vol 9, No.7, Nov 2010. 8-14. (2010)
5. Jorstad I., Thanh D. V., Dustdar S.: Personalisation of Mobile Services: In: IEEE International Conference on Wireless And Mobile Computing, Networking And Communications, 2005. (WiMob'2005), 22-24 Aug. 2005. Vol. 4. 59 – 65. (2005)
6. Nancy K. J., Wegner J. P.: Profiling the Mobile Customer – Privacy Concerns When Behavioural Advertisers Target Mobile Phones – Part I. In: Computer Law and Security Review. Computer Law and Security Review, 26(6), 595-612. Elsevier Publishing. (2010)
7. Microsoft Data Mining Algorithms (Analysis Services - Data Mining). http://msdn.microsoft.com/en-us/library/ms175595(v=sql.105) <29.07.2012>

GeoGebra as e-Learning Resource for Teaching and Learning Statistical Concepts

Dijana Capeska Bogatinoska[1], Aleksandar Karadimce[1], Aneta Velkoska[1]

[1]University of Information Science and Technology "St.Paul the Apostle", Ohrid, R.Macedonia
dijana.c.bogatinoska@uist.edu.mk,
aleksandar.karadimce@uist.edu.mk, aneta.velkoska@uist.edu.mk

Abstract: Understanding probability and statistics is essential in the modern world, where the print and electronic media are full of statistical information and interpretation. The probability and statistics lessons should provide to the students the ability to collect, organize and analyze numerical data, and to understand chance. Appropriate use of technology allows more students access to mathematical concepts in general and also to access statistical concepts. A number of software tools are available for solving and visualization of mathematical problems. GeoGebra, as dynamic mathematics open-source software, is attracting a lot of interest in the mathematical community. Spreadsheet, which enable statistics calculations, and probability calculator are features of GeoGebra that is not found in other dynamic mathematics software. In this paper, using GeoGebra will be created instructional materials that will solve several practical problems from the area of probability and statistics. Use of GeoGebra applets had a positive effect on the understanding and knowledge of the students.

Keywords: GeoGebra, e-learning, probability, statistics

1 Introduction

There is no doubt that probability and statistics are very important in variety of sciences as well as daily applications. Using statistics, the data are turned into knowledge. In the modern digital world, everything is about data. But, data without knowledge are useless. So, for every student regardless of his future profession, is essential to understand basic concepts of mathematical probability and statistics. This will be useful in technology fields as well as business in general, for reporting and understanding results.

In the modern educational process, use of technology is inevitable. There are a number of applications that can be used for solving statistical problems. Among them, we choose the application GeoGebra, because it is free, open-source and is very user-friendly. To demonstrate the real-life application of some of statistical parameters and methods, we use real data from the comparative analysis of handwriting among students with harmonious handwriting and their compeers with dysgraphia in elementary schools in Ohrid [1].

S. Markovski, M. Gusev (Editors): ICT Innovations 2012, Web Proceedings, ISSN 1857-7288

2 Practical Examples in Teaching and Learning Statistics with GeoGebra

In our study, we want to conclude is there a significant difference between the students' ability for writing (is their handwriting good or they have dysgraphia), depending on their gender and their age. For that purpose, we choose 37 students from the total of 238 students (this number statistically represents the *sample* of the population), which during the testing process (this process statistically represent the collection of data) shown that have dysgraphia (they got more than 14 point during testing), and we call this group of students an **experimental group**. This group consists of 33 boys and 4 girls. 13 of them are in second grade, 17 in third and 7 in fourth grade. On those data, using GeoGebra, we performed different statistical methods to make conclusion about data.

The solution of these problems consists of two parts. In the first solution we calculated means, medians and standard deviations and then performed appropriate tests to make conclusions about data. Second solution is a graphical representation of data, which consists of frequency tables and relative frequency histograms (we work with relative frequencies because the number of elements in the lists are not equal; for example we have 33 boys and only 4 girls).

2.1 Example 2 - Dysgraphia by Age in the Experimental Group during Dictation

The statistical question is: "Does the **age** have influence on the handwriting of the students in experimental group (the students that have dysgraphia)?"
To answer this question, we divided the students into three populations: in the first populations are students from second grade, in the second population students from third, and in the third population students from fourth grade. The sample data are grouped and are shown on Table 1:

Table 1. Grouping sample data according to age

grade	**Handwriting points**				total
	<10	**10-13,5**	**14-19**	**>19**	
second	0	0	3	10	13
third	0	0	17	0	17
fourth	0	0	7	0	7
total	0	0	27	10	37

Because in this case we have more than two populations, the *ANOVA* test is used. Using GeoGebra list command three lists are created, and for each list the mean, median and standard deviation are calculated. After that, the ANOVA test is run with following syntax: *ANOVA[<List>, <List>, ...]*. The obtained results are shown on Fig. 1.

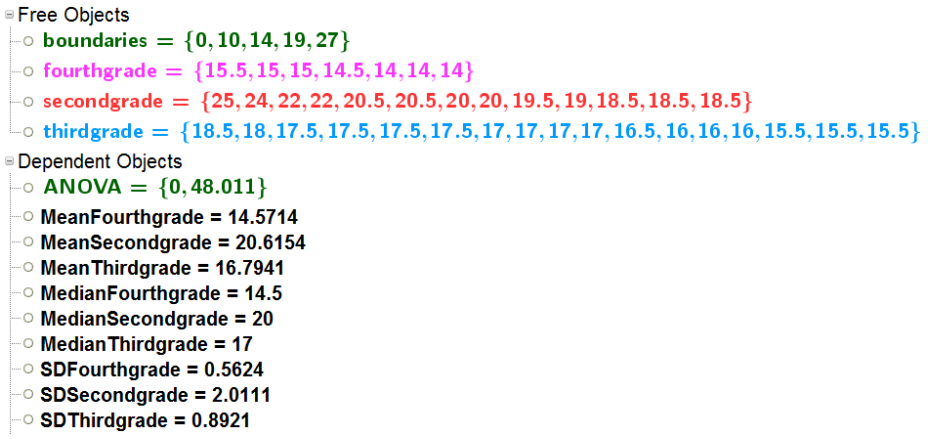


Fig. 1. GeoGebra algebra window with ANOVA test result

Result is returned in list form as {P value, T test statistic}. In our case the result is ANOVA={0, 48.011}. From this result we can conclude that by conventional criteria (P value = 0), this difference is considered to be **statistically significant**, which means that there is big difference between mean values of populations or, with other words, the age has significant influence on the handwriting of the students.

To confirm this result, we drew frequency tables and histogram.

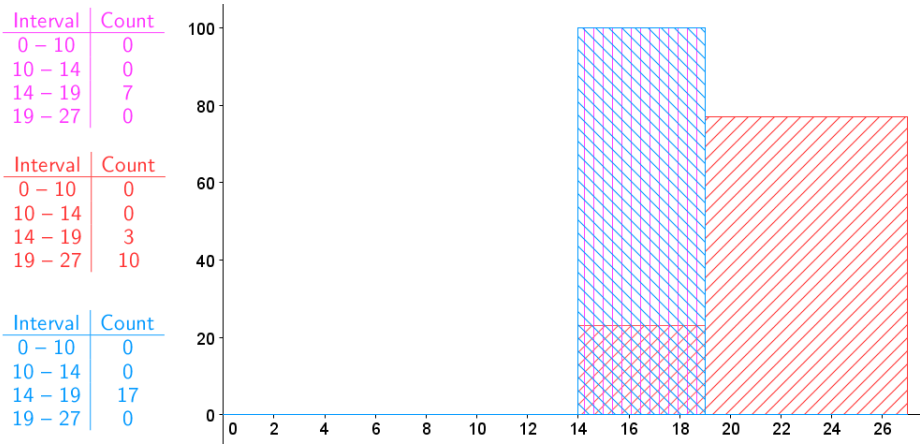


Fig. 2. Frequency tables and Relative frequency histogram in GeoGebra for three populations

By the histogram (Fig. 2) is confirmed the result obtained by ANOVA test. It is obvious that there is significant difference between three populations (second grade is represented with red color, third with blue and forth with pink color), which means that age have significant influence on the handwriting of the students.

The GeoGebra applet with short explanation is uploaded on: http://www.geogebratube.org/student/m15074

3 Conclusion

Learning probability and statistics is sometimes hard for students, because they need mathematics with many new and abstract concepts. It is very hard for them to make relationship between statistical theory and its application in solving different real-life problems. So, the way of teaching statistics is very important. This study is only a small contribution to students' understanding of basic statistical concepts in a new way, through visualization of statistical data, which contribute to better interpretation of obtained results.

GeoGebra has implemented a lot of statistical tools that offer possibility for solving a wide range of statistical problems. Also, the GeoGebra applet can be uploaded on GeoGebra tube together with the short explanation, which means that is available to the students for self-learning.

References

1. Крстеска, А.: Компаративна анализа на ракописот помеѓу учениците со складен ракопис и нивните врсници со дисграфичен ракопис во редовните училишта од општина Охрид. Магистерски труд, Универзитет Св. Кирил и Методиј, Филозофски факултет Скопје, Институт за дефектологија, Скопје (2008).
2. Стојаноски, З., Велкоска, А.: Теорија на веројатност и статистика, Прв Приватен Европски Универзитет Р. Македонија. pp 351. (2009), ISBN 978-608-4574-02-6
3. Малчески, Р.: Статистика за бизнис, Факултет за општествени науки, Скопје, pp. 187-218. АЛФА 94, Скопје (2006). ISBN 9989-936-11-0
4. Freun, R.J., Wilson, W.J.: Statistical methods. Rev ed. Orlando. FL: Ac-ademic press, (1977)
5. Bernstein, S., Bernstein, R.: Schaum's outlines of elements of statistics I: descriptive statistics and probability. McGraw-Hill, New York, (1998)
6. Bernstein, S., Bernstein, R.: Schaum's outline of elements of statistics II: inferential statistics. McGraw-Hill, New York, (1998)
7. Pitler, H., Hubel, E., Kun, M., Malenkoski, K. (2007). *Application of technology in effective school education*. Bitola, Macedonia. Translated to Macedonian (LFS, MT, DCB, CJ, EK), Bitola 2009. ISBN: 978-1-4166-0570-6
8. Dikovic, Lj.: Applications GeoGebra into Teaching Some Topics of Mathematics at the College Level. In: ComSIS Vol.6, No.2, December 2009. (2009)
9. http://geogebrawiki.pbworks.com
10. http://www.geogebra.org/help/documk/index.html

Some fixed point theorems for cyclic contractions in ultra metric spaces

Elvin Rada

Department of Mathematics and Informatics, Faculty of Natural Sciences,
"Aleksandër Xhuvani" University, Elbasan, Albania
elvinrada@yahoo.com

Abstract: In this paper we study some fixed point theorems in an uniformly convex Banach space. We see these results for some contract mappings with cyclic operators of Banach type and Kannan type. We give some results on convergence of Picard operators on a spherically complete ultra metric space. Our intention is to give some existence results for approximation of the fixed points for cyclic contractions using comparison functions that can be used in algorithms.

Keywords: fixed point, ultra metric space, Picard operator, cyclic contraction, comparison function.

1. Introduction and Preliminaries

One of the most important results used in functional analysis is the well-known Banach's contraction which in 1922 asserts that:

If (X,d) is a complete metric space and $T: X \to X$ is a mapping such that

$$d(T(x),T(y)) \le \lambda d(x,y)$$

for all $x, y \in X$ and some $\lambda \in [0,1)$ then T has a unique fixed point in X.

Definition 1. Let (X,d) be a ultra metric space. A mapping $T: X \to X$ is called a ϕ-contraction if there exists a comparison function $\phi: R^+ \to R^+$ such that $d(T(x),T(y)) \le \varphi(d(x,y))$ for all $x, y \in X$.

Definition 2 Let (X,d) be a ultra metric space, m a positive integer $A_1,...,A_m$ nonempty closed subsets of X and $Y = \bigcup_{i=1}^{m} A_i$ an operator $T: Y \to Y$ is called a cyclic ϕ-contraction if

(i) $\bigcup_{i=1}^{m} A_i$ is a cyclic representation of Y with respect to T

(ii) There exists a (c)-comparison function $\phi : R^+ \to R^+$ such that $d(T(x),T(y)) \le \varphi(d(x,y))$
for any $x \in A_i$, $y \in A_{i+1}$ where $A_{m+1} = A_i$

Definition 3. A function $\phi : R^+ \to R^+$ is called a (c)-comparison function if it satisfies:
(i) ϕ is monotone increasing;
(ii) there exist $k_0 \in \mathrm{N}, a \in (0,1)$ and a convergent series of nonnegative terms $\sum_{k=1}^{\infty} v_k$ such that
$\varphi^{k+1}(t) \le \alpha\varphi^k(t) + v_k$ for $k \ge k_0$ and any $t \in R_+$. Let us denote this family with $\mathcal{F}$

2. Main Results

Theorem 4. Let (X,d) be a ultra metric space, m a positive integer $A_1,...,A_m$ nonempty closed subsets of X and $Y = \bigcup_{i=1}^{m} A_i$, a (c)-comparison function $\phi : R^+ \to R^+$, an operator $T : Y \to Y$
Assume that
(i) $\bigcup_{i=1}^{m} A_i$ is a cyclic representation of Y with respect to T
(ii) T is a cyclic ϕ-contraction.

Then T has a unique fixed point $x^* \in \bigcap_{i=1}^{m} A_i$ and the Picard iteration $\{x_n\}$ converges to x^* for any initial point $x_0 \in Y$.

Now we will prove that the Picard iteration converges to x^* for any initial point $x \in Y$. Let $x \in Y = \bigcup_{i=1}^{m} A_i$, there exists $i_0 = \{1,...,m\}$ such that $x_0 \in A_{i0}$. As $x^* \in \bigcap_{i=1}^{m} A_i$ it follows that $x^* \in A_{i_0+1}$ as well. Then we obtain: $d(T(x),T(x^*)) \le \varphi(d(x,x^*))$

By induction, it follows that: $d(T^n(x),x^*) \le \varphi^n(T(x,x^*))$ $n \ge 0$

Since $d(x^*, x^*) \le d(T^n(x), x^*)$ we have $d(x^*, x^*) \le \varphi^n(d(x, x^*))$

Now letting $n \to \infty$ and supposing $x \neq x^*$ we have $d(x^*, x^*) = \lim_{n\to\infty} d(T^n(x), x^*) = 0$

Definition 5. Let (X, d) be an ultra metric space, m be a positive integer, $A_1, A_2, ..., A_m$ be nonempty subsets of X and $X = \bigcup_{i=1}^{m} A_i$ .An operator $T : X \to X$ is a *cyclic weak* $(\varphi - \psi)$*-contraction* if

(i) $X = \bigcup_{i=1}^{m} A_i$ is a cyclic representation of X with respect to T

(ii) $\phi(d(Tx, Ty)) \le \phi(d(x, y)) - \psi(d(x, y))$ for any $x \in A_i$ $y \in A_{i+1}, i = 1,2,...,m$, where

$A_{i+1} = A_1$ and. $\phi, \psi \in F$

An important result based on Karapinar, Sadarangani is the following.

Theorem 6. Let (X, d) be a complete metric space, m be a positive integer, $A_1, A_2, ..., A_m$ be nonempty subsets of X and $X = \bigcup_{i=1}^{m} A_i$. Let $T : X \to X$ be *a cyclic* $(\varphi - \psi)$*-contraction* with $\phi, \psi \in \mathcal{F}$. Then T has a unique fixed point $z \in \cap_{i=1}^{m} A_i$

References

1. Kannan, R: Some results on fixed points. Bull Calcutta Math Soc. 60, 71–76 (1968)
2. Reich, S: Kannan's fixed point theorem. Boll Unione Mat Ital. 4(4):1–11 (1971)
3. Matthews, SG: Partial metric topology. Papers on General Topology and Applications
4. C.Petalas, F.Vidalis: A fixed point theorem in non Archimedean vector spaces, Proc.Amer.Math.Soc., 11 8 (1993), 819-821.
5. Ljiljana Gajic: On ultra metric spaces, Novi Sad J.Math., 31, 2 (2001),69-71.
6. Altun, I, Sadarangani, K: Corrigendum to generalized contractions on partial metric spaces. Topol Appl 158, 1738–1740 (2011).
7. M. P¸acurar,I.A. Rus: Fixed point theorems for cyclic ϕ -contractions, Nonlinear Analysis: Theory, Methods and Applications, Vol 72, Issues 3-4, 1 February 2010,
8. E. Karapınar, K. Sadarangani: Fixed point theory for cyclic (ϕ - ψ)-contractions, Fixed Point Theory and Applications 69 (2011),
9. V. Berinde: Iterative Approximation of Fixed Points, Springer, Berlin, 2007 Rev. Roumanie Math. Pures Appl.,50 (2005), nos 5-6, 443-453

10. M.A. Petric: Some remarks concerning Ciric-Reich-Rus operators, Creative Math. and In., Vol 18(2009), no. 2, 188-193
11. W.A. Kirk, P.S. Srinivasan , P. Veeramani: Fixed Points For Mappings Satisfying Cyclical Contractive Conditions, Fixed Point Theory, Volume 4 No. 1(2003), 79-89

Adaptive Applications: Formal and Informal Definition

Ammar Memari[1] and Jorge Marx Gómez[1]

Carl von Ossietzky University of Oldenburg, Ammerländer Heerstr. 114-118 D-26129 Oldenburg, Germany
ammar.memari@uni-oldenburg.de,
jorge.marx.gomez@uni-oldenburg.de,
WWW home page: http://vlba.uni-oldenburg.de

Abstract. Even though different aspects of adaptivity play a major role in today's software, the term "Adaptive Applications" is not well defined in literature, and is usually confused with other terms referring to other sorts of applications. In this paper we try to de-fuzzify the term and capture common properties of these applications into an informal definition. Then come up with a formal definition through mapping application architecture to a suggested reference model.

Keywords: Adaptive applications, reference modeling, rough sets

1 A Formal and an Informal Definition

For the informal definition we start by listing some well-known applications in different application areas and then start clustering them up to reach several categories thereby defining categories such as collaborative filtering, context aware, recommendation systems etc. Distinguishing in the process between adapted, adaptable and adaptive applications [5]. Following also the model of [1] and [2]. Ending up eventually with a matrix of applications and their different attributes related to adaptation from different dimensions: Sensitivity to: user, content, context and neighborhood Object of adaptation: content, presentation or navigation [3]. Matchmaking and relations: content-content, user-content, user-user, content-context and user-context relations [4]. Depth of adaptation: adapted, adaptable or adaptive [5] both on the first and second order Conceptual space: ability of application to navigate the conceptual space using a single-faceted or a multi-faceted hierarchy or ontology Business constraints: respect of best practice business constraints by the adaptive application like privacy and scrutability [6, 7] Adaptability of application architecture: a modular application architecture is more adaptable than a non-modular one.

For the formal definition we start by listing related formal models especially in the field of adaptive hypermedia and adaptive computing. Such as the AHAM model of [8], its enhancement by [9] and further extensions done by [2]. Other non-Dexter-based models will also be discussed such as the LAOS model by [1]. Sub-models used commonly for building these models will be listed and

S. Markovski, M. Gusev (Editors): ICT Innovations 2012, Web Proceedings, ISSN 1857-7288

defined extending thereby the works of [10]. We will then discuss needs for future generations of adaptive applications trying to capture their properties as well into our final model. Ending up with a set of requirements. In the end we describe a formal reference model using UML (Unified Modeling Language) following the notation of [11] with which we capture the defining attributes of adaptive applications. The model will be used for determining the lower approximation of the adaptive applications set by determining the minimal components/attributes that should be mapped into the application component in order for it to be classified as an adaptive application. Moreover, a higher approximation will be given determining thereby a model for a full-fledged adaptive application that would be exemplified in the end with the www.jinengo.com integrative adaptive navigation application. This application will serve as a proof of applicability and viability of proposed model and ideas, and will help evaluate the model in the end.

References

1. Mooij, A.D., Cristea, A.I.: LAOS: layered WWW AHS authoring model and their corresponding algebraic operators. In: WWW03 The Twelfth International World Wide Web Conference, Alternate Track on Education, Budapest, Hungary (2003)
2. Knutov, E.: Generic Adaptation Framework for Unifying Adaptive Web-based Systems. PhD, Technische Universiteit Eindhoven, Eindhoven NL (2012)
3. Knutov, E., De Bra, P., Pechenizkiy, M.: AH 12 years later: a comprehensive survey of adaptive hypermedia methods and techniques. New Review of Hypermedia and Multimedia **15**(1) (2009) 5 – 38
4. Memari, A., Wagner vom Berg, B., Marx Gmez, J.: An agent-based framework for adaptive sustainable transportation. In: 20th IEEE International Workshops on Enabling Technologies: Infrastructures for Collaborative Enterprises, WETICE 2011, Paris, France, 27-29 June 2011, Proceedings, Paris, France, IEEE Xplore (June 2011)
5. Lenz, C.: Empfaengerorientierte Unternehmenskommunikation Einsatz der Internet-Technologie am Beispiel der Umweltberichterstattung. PhD thesis, Eul, Lohmar; Koeln (2003)
6. Kasanoff, B.: Making It Personal: How To Profit From Personalization Without Invading Privacy. 1st edn. Basic Books (November 2001)
7. Kay, J.: Scrutable adaptation: Because we can and must. In: Adaptive Hypermedia and Adaptive Web-Based Systems. Springer Berlin Heidelberg (2006) 11–19
8. De Bra, P., Houben, G., Wu, H.: AHAM: a dexter-based reference model for adaptive hypermedia. In: HYPERTEXT '99 Proceedings of the tenth ACM Conference on Hypertext and hypermedia : returning to our diverse roots: returning to our diverse roots, Darmstadt, Germany, ACM (1999) 147–156
9. Balík, M., Jelínek, I.: Modelling of adaptive hypermedia systems, Bulgaria (June 2006)
10. Ghali, F., Cristea, A.I.: Social reference model for adaptive web learning. In Spaniol, M., Li, Q., Klamma, R., Lau, R.W.H., eds.: Advances in Web Based Learning ICWL 2009. Volume 5686. Springer Berlin Heidelberg, Berlin, Heidelberg (2009) 162–171

11. Favre, J., NGuyen, T.: Towards a megamodel to model software evolution through transformations. Electronic Notes in Theoretical Computer Science **127**(3) (April 2005) 59–74

S. Markovski, M. Gusev (Editors): ICT Innovations 2012, Web Proceedings, ISSN 1857-7288

S. Markovski, M. Gusev (Editors): ICT Innovations 2012, Web Proceedings, ISSN 1857-7288

Implementation of Robust Digital Watermarking Algorithms using SVD and DCT Techniques

Albian Fezollari[1], Betim Cico[2]

[1, 2] Department of Computer Engineering, FIT, UPT, Tirana, Albania
albianfezollari@gmail.com, betim.cico@gmail.com

Abstract. Copyright risks in multimedia are increasing as a result of large numbers of computers in network. A technique used to preserve rights is watermark implementation to authenticate authorship. This information which is embedded in the original image is a "Digital Watermark", which might be visible or invisible. For this technique to become more efficient, the Watermark should be robust so that its removal from the original image by the hackers becomes difficult. In time, various Watermarking algorithms are developed. However, each of them has its own advantages and limitations. Recent research shows that SVD (Singular Value Decomposition) algorithms are used because of their simple scheme and mathematical function. Another algorithm is DCT (Discrete Cosine Transform). Each of such algorithms has its own advantages, a combination of their advantages is the RST (Rotation, Scaling, Translation) algorithm. This paper shall attempt to compare algorithms and provide a conclusion on their performance and finally implementation in Hardware, in Xilinx Virtex II Pro

Keywords: algorithms, data compression, Gaussian noise, Benchmark testing; cryptographic protocol, Embedded Systems.

1 Introduction

Digitalization of multimedia has brought reliability, speed, big storage that contains space in TB, but their modifications and duplications are simplified. In order for the digital watermark to become efficient in protecting copyrights, it should be robust, retrievable from the document, ensure original information, and impossible to be removed by unauthorized persons. Robust Watermarks are difficult to be removed from original data where embedded. Attackers try with techniques such as JPEG data compression, scale change, rotations, translations, to remove the place where the Watermark is.

S. Markovski, M. Gusev (Editors): ICT Innovations 2012, Web Proceedings, ISSN 1857-7288

2 Digital Watermarking Interface

Watermark images are the methods which have been applied for many years. The watermarking scheme consists of 3 parts: Watermark signal, Encoder and Decoder

2.1 Watermarking Attacks

Based on the watermarking jargon, the attack is a method to find and remove watermarking from content. It can very well be called a process. The watermarking information process is called "attacked information". Robustness is an elimination scheme of various attacks. Attacked data can be easily detected by the watermark quality and channel capacity, from bit errors [1]. The largest category of attacks may be further divided into four distinct groups: removable attacks, geographic attacks, cryptographic attacks, and cryptographic protocol attacks.

2.2 Algorithm

SVD (Singular value decomposition) is an efficient tool to analyze lengths. In SVD transformations a matrix may be decomposed into three matrices of the same size has the original one. Based on Algebra, it can be said that one image is a matrix with out negative and scale numbers. Without losing generality, if A is an image and part of R, $A \in R^{n \times n}$ where R is a real number, then SVD and A are determined as in formula (1):

$$A = USV^T \tag{1}$$

$U \in R^{n \times n}$ and $V \in R^{n \times n}$ are orthogonal matrix and $S \in R^{n \times n}$ is diagonal with σ value. Diagonal coefficients σ have unique values and the conditions are:

$$\sigma_1 \geq \sigma_2 \geq \cdots \sigma_\tau \geq \sigma_{\tau+1} \geq \cdots = \sigma_n = 0$$

The property of SVD transformation in proportional scale is N2 in the original image [2]. S => N "Degree of freedom"

There exists a hybrid method based on DCT (Discrete Cosine Transform) [3, 4] and SVD proposed by Sverdlov [5]. To start with, DCT application in the entire image and DCT coefficients are mapped in 4 boxes using a zigzag sequence. These four boxes actually present the frequency band from the lower to the highest. SV of DCT transformation is used to modify SV of each box. In this paper, the image is divided into four blocks and as a result, the watermark size is one fourth of the image. This shows that the information embedded in low frequency is protected by a set of attacks and when embedded in a high frequency, it is protected by another set of attacks. It should be underlined that such modification is not resistant against translation attacks.

3 Watermark Process

3.1 Embedding Process of Watermark

Embedding process consist on below formulas (2), (3), (4), (5). Application of DCT in image A and mapping of DCT coefficients in 4 blocks B1, B2, B3, B4 Application of SVD in this block where k belongs ,Blocks B1-B4. The parameter k is k=1, 2, 3, 4 and i =1,..., n

$$A^{k}=U_{A}^{k}S_{A}^{k}V_{A}^{kT} \tag{2}$$

Application of DCT in all visual contents of watermark W and application of SV in DCT Transformation:

$$W = U_{W}S_{W}V_{W}^{T} \tag{3}$$

Modification of SV value in each block B from DCT Transformation:

$$\lambda_{i}^{*k} = \lambda_{i}^{k} + \alpha_{k}\lambda_{Wi} \tag{4}$$

Find all four modified DCT coefficients:

$$A^{*k}=U_{A}^{k}\Sigma_{A}^{k}V_{A}^{kT} \tag{5}$$

Mapping of modified DCT coefficients in original positions. Applying the procedure inverse DCT on the image to watermark.

3.2 Processing before extraction

These steps increase robustness of attacks; the steps are as follows:

- Extraction of the size n x n matrix corresponding to the image with watermark
- if n is odd then n = n +1
- Numbers required in the matrix for values that are not infinite number
- Replace the values of the matrices that are not numbered with zero.

3.3 Extraction of Watermark

Application of DCT in image with A* Watermark and mapping of DCT coefficients on Blocks B1-B4. Application of SVS in this quadrant as shown (6):

$$A^{*k}=U_{4}^{k}\Sigma_{4}^{*k}V_{4}^{kT} \tag{6}$$

Extraction of SV from B quadrant:

$$\lambda_{Wi}^{k}=(\lambda_{i}^{*k}-\lambda_{i}^{k})\big/\alpha_{k} \tag{7}$$

Building of DCT coefficients in four visual quadrants by using SV:

$$Wk = W^{k}=U_{W}^{k}\Sigma_{W}^{k}V_{W}^{kT} \tag{8}$$

Applying inverse DCT on the set of coefficients to build the four blocks of watermark

4 Experimental Results

This watermarking scheme has been tested against attacks such as rotation, scaling, translation, Gaussian blur, Gaussian noise, JPEG compression, Equalization histogram, etc.

Table 1. Value of correlation coefficient

Attacks	Sverdlov Method	Modified Method
Blue Gaussian	0,9894	0,997
Gaussian Noise	0,9942	0,9844
JPEG Compression	0,9998	0,995
Histogramsof the equation	0,9148	0,972
Rescaling 256>128>256	0,9957	1
Rotate 20’	0,7617	0,770
Rotate 75’	not applicable	0,843
Rotate 135’	not applicable	0,81
Scale 200%	not applicable	0,997
Translations 25,35	not applicable	0,619

Table 1 provides watermarking images after some attacks and on the right gives the extracted image after the attack. The tool that is used for testing algorithm is Stirmark Benchmark. Based on the correlation values obtained above, it is ensured that the watermarking scheme is RST invariant and resistant against the abovementioned attacks. The scheme provides the same results even if the rotation angle alters. Table 1 provides a comparison of correlation values between Sverdlov and the modified method.

5 Hardware Implementation

Device in Figure 1 are Power cable, Jtag Cable, Monitor, Video Cable, Video Generator, VDEC1, Xilinx XC2VP30 development board. The Verilog language has been chosen.

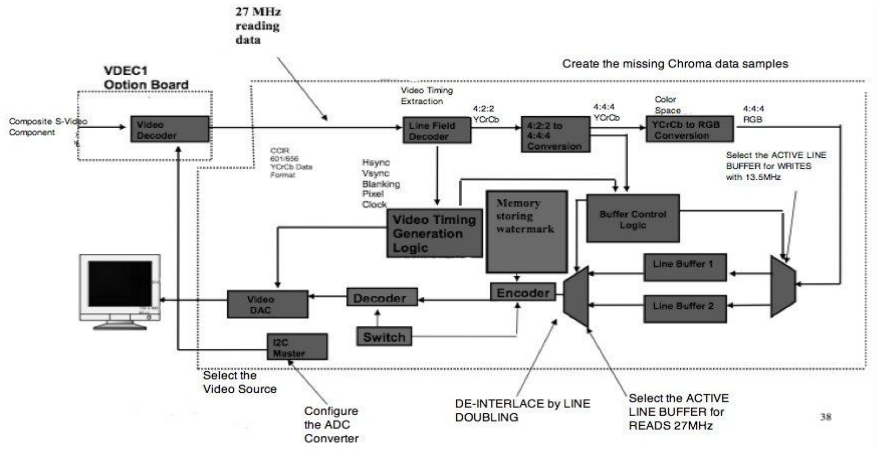


Fig. 1. Watermark Hardware Scheme

6 Conclusions

Some of these algorithms are robust against some attacks but not against some others. SVD based algorithms are new and demand continuous improvements. In this paper we provided some SVD based algorithms and furthermore suggested an idea of a DCT SVD based algorithm. The modified method is applicable for some attacks compare with Sverdlov Method. Moreover a digital watermarking scheme was implemented in the Xilinx Virtex II Pro platform in which the alternation of the watermarking algorithm from encoder to decoder is sufficient for the scheme to be dynamic.

Reference

1. Peter Meerwald, "Digital Image Watermarking in the Wavelet Transform"
2. Zheng, D.Liu,Y., Zhao, J., and El Saddik, A. "A survey of RST invariant image watermarking algorithms", ACM Computing Surveys, Volume 39, No. 2.
3. Barni, F.Bartolini and A. Piva. A DCT domain system for robust image watermarking. IEEE Transactions on Signal Processing. 66, 357-372, 1998.
4. W.C.Chu, DCT based image watermarking using sub sampling. IEEE Trans Multimedia
5. A. Sverdlov, S. Dexter, and A. M. Eskicioglu, Robust DCT-SVD Domain

S. Markovski, M. Gusev (Editors): ICT Innovations 2012, Web Proceedings, ISSN 1857-7288

Applying semantically adapted vector space model to enhance information retrieval

Fisnik Dalipi[1], Ilia Ninka[2], Ajri Shej[3]

[1,3]Department of IT, Faculty of Math-Natural Sciences, Tetovo State University
fisnik.dalipi@unite.edu.mk, ajri.shej@gmail.com
[2]Department of IT, Faculty of Natural Sciences, University of Tirana
ilia.ninka@fshn.edu.al

Abstract. While most enterprise data is unstructured and file based, the need for access to structured data is increasing. In order to reduce the cost for finding information and achieve relevant results there is a need to build a very complex query which indeed is a serious challenge. Data volumes are growing at 60% annually and up to 80% of this data in any organization can be unstructured. In this paper we focus on describing the evolution of some modern ontology-based information retrieval systems. Further, we will provide a brief overview of the key advances in the field of semantic information retrieval from heterogeneous information sources, and a description of where the state-of-the-art is at in the field. Finally, we present and propose a novel use of semantic retrieval model based on the vector space model for the exploitation of KB (Knowledge Base) to enhance and support searching over robust and heterogeneous environments.

Keywords: ontology, information retrieval, semantic web, knowledge base.

1. Introduction

The phrase “information retrieval - IR” dates back to the 1950s [1], but the concept was firstly used at the library catalogues. Initial opinions on the subject emerged from librarianship and information science. Originally, this opinion had philosophical nature, dealing with how information should be classified and organized. Various schools held various positions and an ongoing debate was evident between them on philosophical and anecdotal rather than empirical grounds [2].

However, with the increasing volume of publication, and specifically of scientific literature, after the Second World War, practical concerns of how to effectively access this literature became urgent [3,4]. One of the most influential methods was described by H.P. Luhn in 1957, in which (put simply) he proposed using words as indexing units for documents and measuring word overlap as a criterion for retrieval [4]. In recent times, ontologies are widely used in IR systems.

Nevertheless, its main use has to do with query expansion, which consists in searching for the terms in the ontology more similar to the query terms, to use them together as a part of the query. In this work, we present and propose a novel use of semantic retrieval model based on the vector space model to enhance and support searching over robust and heterogeneous environments.

S. Markovski, M. Gusev (Editors): ICT Innovations 2012, Web Proceedings, ISSN 1857-7288

2. Semantic retrieval from heterogeneous environments

Semantic retrieval from distributed and heterogeneous environments is quite new concept and current ontology based retrieval technologies are very hypothetical, without having any well defined framework on applying ontology based search to the web as whole, which is consisted by unlimited number of domains. Some attempts have been made by [5], but they lack to address the potential use of ontology search beyond the organizational data corpus, as their models have difficulties to deal with the heterogeneity of Web and are limited to a predefined set of ontologies. The proposed architecture in Figure 3 reflects the concept of heterogeneity assuming large amount of semantic metadata online without having a pre-defined range of domains. We assume that the external element is not only a single knowledge base but involves online semantic web information.

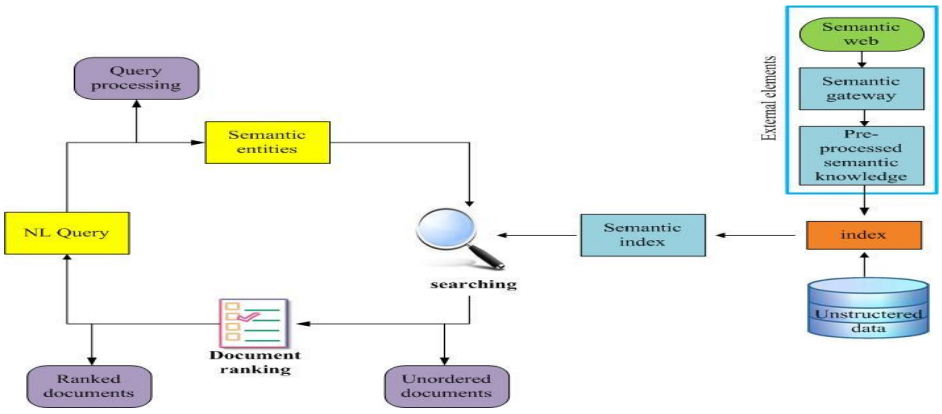


Fig. 3 Semantic information retrieval framework

This model does not require users to know special purpose query language; rather, the system expects queries to be expressed in natural language. Another relevant aspect is that the set of unstructured (web) information is not needed to be adapted into conventional fragments of ontological knowledge. In order to answer the queries, the system uses available semantic data and other information from standard web pages. When dealing with such a large amount of semantic information, we need a semantic gateway which will pre-process, gather, store and access the online distributed semantic web information. One of the most popular semantic way gateways currently available in the state of the art are: Watson [6], and Swoogle [7]. Once the user poses the query, that query can further be processed by any ontology based system which ensures access to the online ontologies and that translates generic natural language queries into SPARQL. Such systems of choice could be AquaLog, proposed by [8,9], Querix [10], or QASYO [11]. After returning the fragments of relevant ontological knowledge as an

answer, the system will perform a second step which includes retrieving and ranking by their probability the documents which contain the needed information. The ranking process can apply the concepts of vector space model ranking algorithm.

References

[1] S. Robertson. On the early history of evaluation in IR. In J. Tait, editor, Charting a New Course: Natural Language Processing and Information Retrieval – Essays in Honour of Karen Sp¨arck Jones, pages 13–22. Springer, 2005.

[2] S. Robertson. On the history of evaluation in IR. Journal of Information Science, 34 (4):439–456, 2008a

[3] W. Cleverdon. The significance of the Cranfield tests on index languages. In A. Bookstein, Y. Chiaramella, G. Salton, and V. V. Raghavan, editors, Proc. 14th Annual International ACM SIGIR Conference on Research and Development in In-formation Retrieval, pages 3–12, Chicago, Illinois, USA, Oct. 1991.

[4] H. P. Luhn. A statistical approach to mechanized encoding and searching of literary information. IBM Journal of Research and Development, 1(4):309–317, 1957.

[5] Maedche, A., Staab, S., Stojanovic, N., Studer, R., & Sure, Y. (2003). SEmantic portAL: The SEAL Approach. Spinning the Semantic Web. MIT Press , 317-359

[6] D'Aquin, M., Gridinoc, L., Sabou, M., Angeletou, S., & Motta, E. (2007). Characterizing
Knowledge on the Semantic Web with Watson. 5th International EON Workshop at International Semantic Web Conference (ISWC'07). Busan, Korea.

[7] Ding, L., Finin, T., Joshi, A., Pan, R., & Cost, S. (2004). Swoogle: A Search and Metadata Engine for the Semantic Web. 13th Conference on Information and Knowledge Management (CIKM 2004), (pp. 625-659). Washington, DC, USA.

[8] V. Lopez, M. Pasin, and Enrico Motta, "AquaLog: An Ontology-Portable Question Answering System for the Semantic Web," Lecture Notes in Computer Science, Vol. 3532, Springer, Berlin, pp. 546-562, 2005.

[9] V. Lopez, and E. Motta, "Ontology-Driven Question Answering in AquaLog," Lecture Notes in Computer Science, Vol. 3136. Springer-Verlag, Berlin, pp. 89–102, 2004.

[10] E. Kaufmann, A. Bernstein, and R. Zumstein., "Querix: A natural language interface to query ontologies based on clarification dialogs," In proceeding 5th International Semantic Web Conference (ISWC 2006), pp 980–981, 2006.

[11] A. M. Moussa and R. F. Abdel-Kader. QASYO: A Question Answering System for YAGO Ontology. International Journal of Database Theory and Application Vol. 4, No. 2, June, 2011

S. Markovski, M. Gusev (Editors): ICT Innovations 2012, Web Proceedings, ISSN 1857-7288

A Literature Review of Data Mining Techniques Used in Healthcare Databases

Elma Kolçe (Çela) [1], Neki Frasheri [2]

[1,2] Department of Computer Engineering, Polytechnic University of Tirana, Albania
[1] elmakolce@yahoo.com, [2] nfrasheri@fti.edu.al

Abstract. In this paper we present an overview of the current research being carried out using the data mining techniques for the diagnosis and prognosis of various diseases. The goal of this study is to identify the most well-performing data mining algorithms used on medical databases. The following algorithms have been identified: Decision Trees, Support Vector Machine, Artificial neural networks and their Multilayer Perceptron model, Naïve Bayes, Fuzzy Rules. Analyses show that it is very difficult to name a single data mining algorithm as the most suitable for the diagnosis and/or prognosis of diseases. At times some algorithms perform better than others, but there are cases when a combination of the best properties of some of the aforementioned algorithms together results more effective.

Keywords: Data Mining (DM), Decision Tree (DT), Support Vector Machine (SVM), Artificial Neural Network (ANN), Naïve Bayes, Genetic Algorithm, Logistic Regression, Healthcare Database, Diagnosis, Prognosis

1 Introduction

Data mining is defined as "a process of nontrivial extraction of implicit, previously unknown and potentially useful information from the data stored in a database" by Fayyad [1]. Healthcare databases have a huge amount of data but however, there is a lack of effective analysis tools to discover the hidden knowledge. Appropriate computer-based information and/or decision support systems can help physicians in their work. Efficient and accurate implementation of an automated system needs a comparative study of various techniques available. In this paper we present an overview of the current research being carried out using the DM techniques for the diagnosis and prognosis of various diseases, highlighting critical issues and summarizing the approaches in a set of learned lessons. The rest of this paper is organized as follows: First we show the methodology of research used in this study in chapter two, we classify them with different criterions in chapter three, then we identify the most used

S. Markovski, M. Gusev (Editors): ICT Innovations 2012, Web Proceedings, ISSN 1857-7288

algorithms for disease diagnosis and prognosis, and finally we show the conclusions of our work.

2 Methodology

The methodology used for this paper was through the survey of journals and publications in the fields of computer science, engineering and health care. European Journal of Scientific Research, International Journal on Computer Science and Engineering, Expert Systems with Applications, Data Science Journal are some of these journals. In order to obtain a general overview on the literature, book chapters, dissertations, working papers and conference papers are also included. The research is focused on most recent publications.

3 Literature review

There are different kinds of studies for DM techniques in medical databases. We identify the following categories:

1. Studies that summarize reviews and challenges in mining medical data in general [6], [24], [25], [31], [32]
2. Studies of DM techniques used for diagnosing and/or prognosing of specific diseases, which can be further classified into three other categories: those which use DM techniques for disease diagnosis [3],[7],[9],[14],[22],[37], for disease prognosis [4],[10],[26],[29],[42],[43], or both diagnosis and prognosis.[13],[36]
3. Studies to investigate factors which have higher prevalence of the risk of a disease[5],[12],[28]
4. Studies that present new technologies and algorithms [18-21], [40], [41] and studies that present new techniques improving old ones, such as [8],[11],[30],[39]
5. Studies that present new frameworks, tool and applications in medicine and healthcare system [2],[15-17],[23],[33-35],[38]

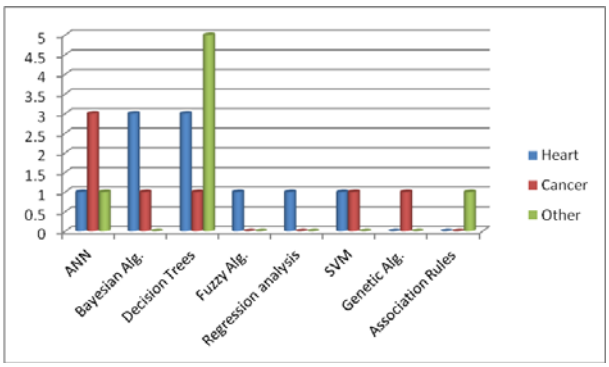


Fig. 1. Efficient Algorithms for Disease Diagnosis

4 Well-performing dm algorithms used for disease diagnosis and prognosis

The graphs in Figures 1 and 2 show the most well-performing algorithms used for disesase diagnosis and prognosis respectively, resulting from the studies in Chapter 3 (excluding studies of categories 1 and 4). We have classified the diseases in Heart Diseases (Cardiovascular disease, Heart Attack, Coronary Arthery Disease, Hypertension), Cancer Diseases (Breast, Prostate, Pancreatic Cancer) and Other Diseases (Asthma, Diabetes, Hepatitis, Kidney Disease, Nerve Diseases, Chronic Disease, Skin Diseases).

As we can see in Fig.1, ANNs are the most well-performing in diagnosing Cancer Diseases, Bayesian Algorithms and Decision Trees in Heart Diseases, and DTS in diagnosing other diseases. On the other side in Fig. 2 we can see that for Cancer and Heart Disease Prognosis, ANNs are the most well-performing and also Bayesian Algorithms the most well-performing in Heart Diseases Prognosis.

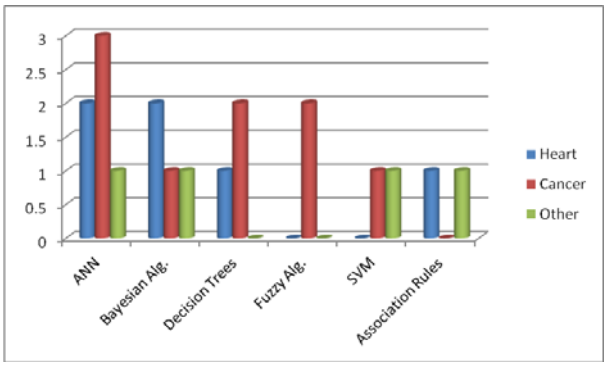


Fig. 2. Efficient Algorithms for Disease Prognosis

5 Conclusions

In this paper we identified and evaluated the most commonly used DM algorithms resulting as well-performing on medical databases, based on recent studies. The following algorithms have been identified: Decision Trees (DT's) C4.5 and C5, Support Vector Machine (SVM), Artificial neural networks (ANNs) and their Multilayer Perceptron model, Bayesian Networks and Naïve Bayes, Logistic Regression, Genetic Algorithms (GAs), Fuzzy Rules, Association Rules.

Analyses show that DTs, ANNs and Bayesian Algorithms are the most well-performing algorithms used for disease diagnosis, while ANNs are also the most well-performing algorithms used for disease prognosis, followed by Bayesian Algorithms, DTs and Fuzzy Algorithms. But it is very difficult to name a single DM algorithm as the best for the diagnosis and/or prognosis of all diseases. Depending on concrete situations, sometime some algorithms perform better than others, but there are cases

when a combination of the best properties of some of the aforementioned algorithms results more effective. The follow-up of our work will aim at dealing with algorithms that have wider spectra of application for groups of diseases.

References

1. Fayyad, U. M. , Piatetsky-Shapiro, G., Smyth, P., Uthurusamy , R. G. R.: Advances in Knowledge Discovery and Data Mining. AAAI Press / The MIT Press, Menlo Park, CA. (1996)
2. Shantakumar B.Patil, Y.S.Kumaraswamy: Intelligent and EffectiveHeart Attack Prediction System Using Data Mining and Artificial Neural Network, European Journal of Scientific Research ISSN 1450-216X Vol.31 No.4 (2009), pp.642-656 © EuroJournals Publishing, Inc. 2009.
3. M.Kumari, S. Godara: Comparative Study of Data Mining Classification Methods in Cardiovascular Disease Prediction, IJCST ISSN : 2229- 4333 Vol. 2, Iss ue 2, June 2011
4. K.Srinivas , B.Kavihta Rani, Dr. A.Govrdhan: Applications of Data Mining Techniques in Healthcare and Prediction of Heart Attacks (IJCSE) International Journal on Computer Science and Engineering Vol. 02, No. 02,(2010),pp 250-255
5. M. Karaolis, J.A. Moutiris, L. Papaconstantinou, C.S. Pattichis: Association Rule Analysis for the Assessment of the Risk of Coronary Heart Events (2009)
6. R.D. Canlas Jr., Data Mining in Healthcare:Current Applications and Issues (2009)
7. J.Soni, U. Ansari, D. Sharma, S. Soni: Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction (2011)
8. K.S.Kavitha , K.V.Ramakrishnan , M. K. Singh: Modeling and design of evolutionary neural network for heart disease detection, IJCSI International Journal of Computer Science Issues, Vol. 7, Issue 5, September 2010, ISSN (Online): 1694-0814, pp. 272-283 (2010)
9. Chi-Ming Chu, Wu-Chien Chien, Ching-Huang Lai, Hans-Bernd Bludau, Huei-Jane Tschai, LuPai, Shih-Ming Hsieh, Nian-Fong Chu, Angus Klar, Reinhold Haux, Thomas Wetter: A Bayesian Expert System for Clinical Detecting Coronary Artery Disease, J Med Sci 2009; 29(4), pp. 187-194 (2009)
10. A.A. Aljumah, M. G.Ahamad, M.K.Siddiqui: Predictive Analysis on Hypertension Treatment Using Data Mining Approach in Saudi Arabia, Intelligent Information Management, 3, (2011), pp. 252-261
11. S.H.Ha, S.H.Joo: A Hybrid Data Mining Method for the Medical Classification of Chest Pain, International Journal of Computer and Information Engineering 4:1,pp 33-38 (2010)
12. C. Yang, W. N.Street, Der-Fa Lu, L. Lanning: A Data Mining Approach to MPGN Type II Renal Survival Analysis(2010)
13. S.Gupta, D. Kumar, A.Sharma: Data Mining Classification Techniques Applied For Breast Cancer Diagnosis And Prognosis (2011)
14. B.D.C.N. Prasad, P.E.S.N. K.Prasad, Y. Sagar: A Comparative Study of Machine Learning Algorithms as Expert Systems in Medical Diagnosis (Asthma) (2011)
15. A.Shukla, R. Tiwari, P. Kaur: Knowledge Based Approach for Diagnosis of Breast Cancer, IEEE International Advance Computing Conference (IACC 2009)
16. E. Savic, J.Potic, Z. Babovic, G. Rakocevic, V. Strineka, M. Dobrota, V. Milutinovic: Sensor Nets and Data Mining in Medical Applications (2011)
17. L. Duan, W. N. Street & E. Xu: Healthcare information systems: data mining methods in the creation of a clinical recommender system, Enterprise Information Systems, 5:2, pp169-181 (2011)

18. T.H. McCormick, C. Rudin, D.Madigan: A Hierarchical Model For Association Rule Mining Of Sequential Events: An Approach To Automated Medical Symptom Prediction
19. S. CHAO, F.WONG: An Incremental Decision Tree Learning Methodology Regarding Attributes In Medical Data Mining (2009)
20. S.Chao , F. Wong: A Multi-Agent Learning Paradigm for Medical Data Mining Diagnostic Workbench
21. I.Ullah: Data Mining Algorithms And Medical Sciences (2012)
22. C. S. Dangare, S.S. Apte: Improved Study of Heart Disease Prediction SystemUsing Data Mining Classification Techniques (2012)
23. D.S.Kumar, G.Sathyadevi, S.Sivanesh: Decision Support System for Medical Diagnosis Using Data Mining (2011)
24. N.Satyanandam, Dr. Ch. Satyanarayana, Md.Riyazuddin, A.Shaik: Data Mining Machine Learning Approaches and Medical Diagnose Systems : A Survey
25. F.Hosseinkhah, H.Ashktorab, R.Veen, M. M. Owrang O.: Challenges in Data Mining on Medical Databases IGI Global pp. 502-511(2009)
26. D.Delen: Analysis of cancer data: a data mining approach (2009)
27. E.Dincer, N.Duru: Prototype of a tool for analysing laryngeal cancer operations
28. Acute Coronary Syndrome Prediction Using Data Mining Techniques- An Application, World Academy of Science, Engineering and Technology 59 pp.474-478 (2009)
29. A.O. Osofisan ,O.O. Adeyemo, B.A. Sawyerr, O. Eweje: Prediction of Kidney Failure Using Artificial Neural Networks (2011)
30. R. Parvathi, S. Palaniammalì: An Improved Medical Diagnosing Technique Using Spatial Association Rules, European Journal of Scientific Research ISSN 1450-216X Vol.61 No.1 pp. 49-59 (2011)
31. F.I.Dakheel, R.Smko, K. Negrat, A.Almarimi: Using Data Mining Techniques for Finding Cardiac Outlier Patients (2011)
32. S.K. Wasan, V. Bhatnagar , H.Kaur: The Impact Of Data Mining Techniques On Medical Diagnostics, Data Science Journal, Volume 5, pp. 119-126 (2006)
33. S.Palaniappan, R. Awang: Intelligent Heart Disease Prediction System Using Data Mining Techniques (2008)
34. M.G. Tsipouras, T.P. Exarchos, D.I. Fotiadis,A.P. Kotsia, K.V. Vakalis, K.K. Naka, L. K. Michalis: Automated Diagnosis of Coronary Artery Disease Based on Data Mining and Fuzzy Modeling (2008)
35. M. L.Jimenez , J. M. Santamarı, R. Barchino, L. Laita, L.M. Laita, L. A. Gonza´lez, A. Asenjo: Knowledge representation for diagnosis of care problems through an expert system: Model of the auto-care deficit situations, Expert Systems with Applications 34 pp.2847–2857 (2008)
36. M.-J. Huang, M.-Y.Chen, S.-C. Lee: Integrating data mining with case-based reasoning for chronicdiseases prognosis and diagnosis, Expert Systems with Applications 32 pp.856–867 (2007)
37. K.Aftarczuk: Evaluation of selected data mining algorithms implemented in Medical Decision Support Systems (2007).
38. T.Sakthimurugan, S.Poonkuzhali: An Effective Retrieval of Medical Records using Data Mining Techniques, International Journal Of Pharmaceutical Science And Health Care. ISSN: 2249-5738. 2(2), pp 72-78 (2012)
39. J.Gao, J. Denzinger, and R.C. James: A Cooperative Multi-agent Data Mining Model and Its Application to Medical Data on Diabetes (2005)
40. A.Habrard, M.Bernard, F. Jacquenet: Multi-Relational Data Mining in Medical Databases, Springer-Verlag (2003), LNAI 278

41. A.Kusiak, Decomposition in Data Mining: A Medical Case Study , B.V. Dasarathy (Ed.), Proceedings of the SPIE Conference on Data Mining and Knowledge Discovery: Theory, Tools, and Technology III, Vol. 4384, SPIE, Orlando, FL, April 2001, pp. 267-277
42. S.Floyd: Data Mining Techniques for Prognosis in Pancreatic Cancer (2007)
43. A.Kika, B.Cico, R.Alimehmeti: Using Machine Learning for Preoperative Peripheral Nerve Surgical Prediction (2010)

WiMAX Technology and Coverage in Kosovo

Vigan Raça[1], Betim Çiço[2]

[1]Department of Computer Sciences, SEEU, CST, Tetovo, Macedonia
viganraca@gmail.com
[2] Department of Computer Engineering, FIT, UPT, Tirana, Albania
bcico@abcom.al

Abstract - Analysis of telecommunication systems of particular network generation, is a wide area that includes various mechanisms for research. Achieving coverage is one of the main objectives that affects network quality and directly the customer. The main reason of this analysis is to offer services for customers regardless their location. Coverage structure is organized by the cells that are created through the sectors that are located in the base stations (BTS). The different frequency bands enables the achievement of higher efficiency coverage. By increasing coverage various problems become present including monitoring and system maintenance. The WiMAX 3G system and combined services, voice and data, offers a good solution that is part of the future. As any system that requires financial justification, WiMAX has its own specific system unless services that offers it increase economic welfare. Finally part of this paper are obstacles faced through the phase of coverage; weaknesses that affect the system performance; opportunities and ideas for further development and advancement of mobile telecommunication technologies.

Keywords: Wireless, Coverage, Network, Systems.

1 Introduction

Communication and research on the internet is becoming the solution of most problems faced world-wide. Nowdays wireless communication system has become easy to implement. As part of wireless networks, WiMAX technology for which is analyzed on this paper, is a good opportunities toward achieving the success in the world of telecommunications. System is built in a way to offer subscribers satisfactory services like voice and internet regardless location [1]. Modulation technique (OFDMA)[1] that used WiMAX system allows many subscribers to connect in cell simultaneously. Urban areas are more complicated and pose a problem during designing the network coverage due to the great density of buildings and large numbers of subscribers. While rural areas are more simplified compared with urban. Further development is done by WiMAX forum and manufacturing companies of WiMAX devices.

[1] **Orthogonal frequency-division multiplexing (OFDM)** is a method of encoding digital data on multiple carrier frequencies

2 Basis concepts and characteristics of WiMAX

WiMAX Interoperability for Microwave Access is a technology whereby transmitted data can be voice (VoIP) or information data. It uses wireless channels on large distances along different paths based on Point-to-Point connections [2]. This is based on IEEE 802.16 Standard as we mentioned above, otherwise known as Wireless MAN. WiMAX technology enabled Internet users to surf on web via laptop or computer, without the need to physically connect to router, hub or switch. The WiMAX name was created by WiMAX Forum [7] that was established on January 2001. WiMAX mobility[3] is wireless broadband solution that provides coverage for both types of networks; mobile and fixed through radio waves propagation. It uses OFDMA and SOFDMA techniques, and is based on NLOS[2]. The working principle of WiMAX technology is based on these elements[3]:

- Base Station BS is designed to operates on 3.5 GHz frequency band
- System Management and Monitoring of Network BWA which provides and configure normal functioning of system work.
- Call Session Controller CSC as software part of devices
- Session Border Control that provides security and QoS functionality
- Voice Gateways enables flexibility of PSTN migration to NGN network.

Table 1. Wimax Characteristics

Technology	WiMAX (802.16a/revD)	WiMAX (802.16e)
Bandwidth	*1.75, 3.5, 7, 14, 20 MHz*	*1.75, 2.5, 7, 14, 20 MHz*
Downlink Speed	*> 70 Mbit/s, 20 MHz channel*	*> 70 Mbit/s, 20 MHz channel*
Uplink Speed	*BTS Capacity – 4 Mbit/s*	*BTS Capacity – 4 Mbit/s*
Latency	*Not known, low*	*Not known , low*
Mobility	*Fixed*	*Limited mobility*
Speech	*VoIP*	*VoIP*
Availability	*Fall 2005*	*2007*
Cell Radius	*5 – 10 Km*	*2 – 5 Km*
Standard	*Ready*	*Fall 2005*

3 WiMAX coverage in Kosovo

The current state of telecommunications infrastructure in Kosovo shows that this is ideal time to implement WiMAX technology. Based on the current situation of the population, around 70% of population lives in cities whereas around 30% lives in villages. The following table will reflect the state's residential population.
The houses in Kosova towns are built very near to each other. Also in the smaller villages the buildings typically form clusters of 5 to 10 houses. This makes it possible to share the WLL/BWA Customer Premises equipment between various households. Also, there houses are separated from other houses that CPE sharing is not feasible.

[2] **Non-line-of-sight** (**NLOS**) is radio transmission across a path that is partially obstructed

3.1 Rural Area

For rural areas we have made the following assumptions [8]:

- 8% of the households are separated houses
- 66% of households belong to clusters where one Customer Premises Equipment can be shared between five households
- 26% of household belongs to clusters where one Customer Premises Equipment can be shared between ten household

3.2 Country Town

For Country Town we have made the following assumptions [8]:

- 2% of the households are separated houses.
- 58% of households belong to clusters where one Customer Premises Equipment can be shared between five households.
- 40% of household belong to clusters where one Customer Premises Equipment can be shared between ten households.

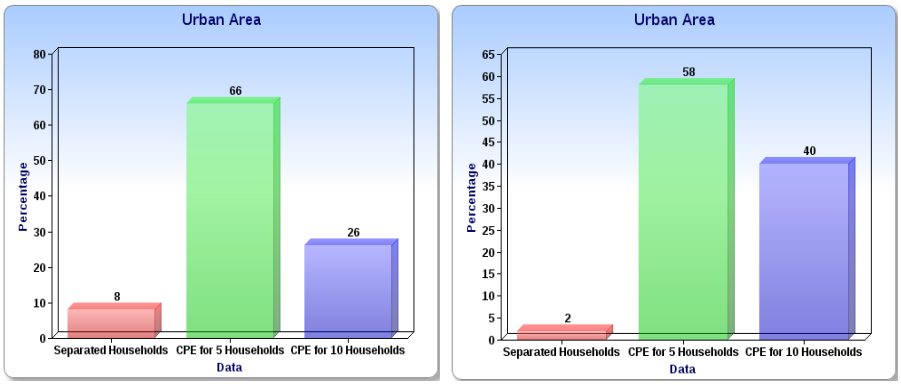


Fig1. Rural and Urban Area Households

4 Why WIMAX in Kosovo

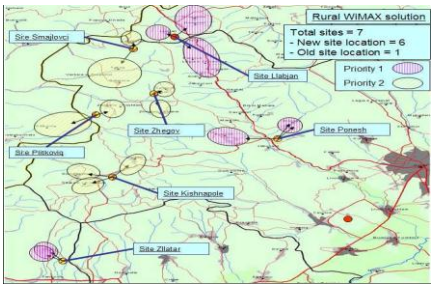
Many responses argue WiMAX system implementation. The current technological infrastructure in Kosovo, is the main reason justifying WIMAX system implementation. Services that are offered today in Kosovo through telephone lines via DSL or cable technology (cable modem) or wireless through Wi-Fi access are not satisfying customers, both, financially and quality.

4.1 Coverage and Capacity Dimensioning

The radio network dimensioning methodology is based on desk-top studies on geographical maps, terrain profile analysis, benchmarking of existing GSM coverage and area surveys for verification of the desk-top studies.

- Coverage dimensioning
- Capacity dimensioning.

Radio network dimensioning takes the existing towers as the preferred location to place base stations to cover the desired areas. New towers for base stations are considered only as a secondary option. Based on the above subscriber profile assumptions and required capacities per subscriber, the capacity dimensioning for the radio network was performed as presented in the figure [2].



Site Name	Households	Traffic Mbit/s	Priority	Expansion need	Site status
Llabjan	123 (can be more)	1,788	1	No	Existing
Ponesh	554	8,060	1	No	Not built
Zllatar	232	3,373	1	No	Not built
Kishnapole	152	2,210	2	No	Not built
Smajlovci	120	1,744	2	No	Not built
Plitkoviq	90	1,302	2	No	Not built
Zhegovc	29	0,421	2	No	Not built

Fig. 2. The coverage and capacity dimensioning in certain areas

5 Conclusion

Based on given analysis on this paper we conclude that WiMAX system meets all possible criteria and requirements that today's human civilization needs for communication. Planning coverage including (base stations, cells capacity and dimensioning etc) and challenges during the coverage, remains a key objective in implementing of this technology. Achieving capacity was another position which should be fulfilled for more effective coverage. These analysis that is reported on this paper, provides a good solution for Kosovo, and their implementation will solve a problem that will be faced on the future.

References

1. Radha Krishna Rao G. Radhamani , "WiMAX A Wireless Technology Revolution, October 19, 2007
2. Maode Ma Editor, "Current Technology Developments of WiMax Systems, Januaray 30, 2009
3. Bachir Belloul1, Simon R. Saunders3 "Measurements and Comparison of WiMAX Radio Coverage " , 09, november 2007
4. Bharathi Upase, Sunil Vadgama "Radio Network Dimensioning and Planning for WiMAX, 7 May, 2007
5. Kai Dietze Ph.D, & Ted Hicks "WiMAX Uplink and Downlink Design Considerations", 8 May 2008
6. Anna Ha´c , "Mobile Telecommunication Networks for Data", 15 November 2002.
7. www.wimaxforum.org
8. Consultancy Services for Technical, Legal and Financial Support to Wireless Local Loop Project for Post and Telecommunication of Kosovo (PTK), 30 September 2005.

Weather Forecast Web-Site weatherforecast.tk : History and Development Perspectives

Dmytro A. Zubov, Ile Dimitrievski

University for Information Science and Technology, Ohrid, Macedonia

{dmytro.zubov , ile.dimitrievski}@uist.edu.mk

Abstract. In this paper, the short history of the long-term weather forecast inductive method was shown. Its application was realized in the web-site weatherforecast.tk on the ASP.NET technology basis. The operation experience showed next main development perspectives: cloud computing (Windows Azure's web role), partial-page update (e.g., update panel), weather forecast's calculation for the definite set of the geographical coordinates, the automatic algorithm's development for the parameters of the above long-term weather forecast inductive method. The average daily air temperature forecast's mistake is up to 6.5 % for Skopje Airport (half year ahead).

Keywords: weather forecast, inductive simulation, cloud computing.

1 Introduction

Data mining for weather forecast in general and for weather inductive forecast in particular have been of long standing interest. Several approaches for long term and short term weather forecast have been suggested. However, precise forecast of the weather conforming to localised environment conditions is fraught with difficulties due to computational complexity of long-term forecasts requiring computing grid and/or clouds (SaaS-technology) to compute the forecast values in reference points using the classical inductive analogue method for more than 80 % forecast quality at the average. The problem focusing on enhancement in quality of the long and short term forecast of weather processes, therefore, has drawn considerable attention. As may be seen in Internet that rise to this situation can be attributed to weather processes and noosphere interplay. It is well known that classical hydrodynamic equations are used for continuous modelling of these events. But, this approach depends on the variables variation (the theory catastrophe's well-known effect). In addition, discrete modelling is applied on the basis of analogue complexing algorithm. In addition, experts note that the forecast's quality (up to 75 % at average) and forward (up to two weeks mainly) are not enough nowadays. Furthermore, standard software and hardware combined solutions with user-friendly interface have not been developed as yet.

S. Markovski, M. Gusev (Editors): ICT Innovations 2012, Web Proceedings, ISSN 1857-7288

2 Realization of Some Development Perspectives

Partial-page update. Creating a web form that performs partial-page updates to the server (without a full-page refresh) is the top-ranked approach of the web-application improvement. It was realized with the AJAX technology (UpdatePanel control) – first element for short-term forecast part and second one for long-term forecast.

Cloud computing. It is well known that Windows Azure's web-role is reflection of the classic ASP.NET application. One positive feature can be underlined – all processor's time is used for web-application's computation. Therefore, original ASP.NET application was converted to Windows Azure's web-role application.

Language facilities. It is clear that the multilingual interface is necessary (English and Russian languages were realized). Weather forecast's calculation for the definite set of the geographical coordinates, and the automatic algorithm's development for the parameters of the above long-term weather forecast inductive method. This item is realized on the basis of next stages: 1. Correlation analysis of the different cities' data. This stage's result is a list of cities with the correlation function's argument more than defined value (0.7 and above in absolute value). 2. Forecast model's synthesis on the basis of the inductive modelling. We will describe these stages in details on the basis of Skopje Airport's air temperature long-term forecast model's synthesis.

67 places took part in the correlation analysis initially (names were written according to the www7.ncdc.noaa.gov; places were chosen with 2 criterion: one country – one representative place; time series have to be continuous from 1st January, 1973): Nwso Agana, Aarhus Lufthavn, Abbeville, Aeropuerto Pettiros, Amman Airport, Amsterdam AP Schiph, Annaba, Ashgabat Keshi, Athinai al Helliniko, Auckland Airport, Bangkok Metropolis, Beijing, Ben-Guron International Airport, Beograd-Surcin, Bogota-Eldorado, Brasilia-Aeroporto, Bratislava-Letisko, Bruxelles National, Bucuresti INMH-Bane, Budapest-Ferihegy I, Busan, Cairo Airport, Canberra Airport, Caracas-Maiquetia, Damascus International Airport, Geneve-Cointrin, Gibraltar, Guernsey Airport, Helsinki-Vantaa, Hengchun, Jersey Airport, Kiev, Kingston-Norman Man, Kisinev, Kwajalein-Bucholza, La Paz-Alto, Lima-Callao Airport, Lisbon, London, Luqa, Luxembourg, Minsk, Moscow, Nassau Airport New, New Delhi-Safdarjun, Noumea-Nlle-Calledo, Nuuk, Oslo-Gardermoen, Paphos Airport, Praha-Libus, Rabat-Sale, Rarotonga, Reykjavik, Riga, Roma-Ciampino, Skopje Airport, Tallin-Harku, Tashkent, Tokyo, Torshavn, Tripoli, Tunis-Carthage, Ulaanbaatar, Vaduz, Warszawa-Okecie, Washington National, Wien-Hohe Warte.

20 time series were selected as the most correlated to Skopje Airport with half-year (approximately) delay (correlation function is normalized and centralized): 0. Skopje Airport (delay 181, autocorrelation function's value -0.832742597188605). 1. Aeropuerto Pettiros (delay 187, correlation function's value 0.623045698552278). 2. Ashgabat Keshi (delay 184, correlation function's value -0.851641283820963). 3. Auckland Airport (delay 175, correlation function's value 0.725715828489775). 4. Canberra Airport (delay 182, correlation function's value 0.798801309832135). 5. Gibraltar (delay 164, correlation function's value -0.823452379725064). 6. Guernsey Airport (delay 163, correlation function's value -0.782844557722741). 7. Jersey Airport (delay 167, correlation function's value -

0.782198050491036). 8. Lisbon (delay 165, correlation function's value -0.761637258660198). 9. London (delay 174, correlation function's value -0.781332533443968). 10. Nassau Airport New (delay 165, correlation function's value -0.767634036673984). 11. New Delhi-Safdarjun (delay 199, correlation function's value -0.840312570729328). 12. Noumea-Nlle-Calledo (delay 166, correlation function's value 0.782559778633274). 13. Nuuk (delay 168, correlation function's value -0.740870261967304). 14. Paphos Airport (delay 166, correlation function's value -0.858819233109924). 15. Rabat-Sale (delay 165, correlation function's value -0.784343051359234). 16. Reykjavik (delay 172, correlation function's value -0,734883892104582). 17. Tashkent (delay 183, correlation function's value -0.833431861928644). 18. Tokyo (delay 168, correlation function's value -0.855582219493774). 19. Washington National (delay 179, correlation function's value -0.837538640603375).

Average daily air temperature long-term forecast model has next linear structures:

$$\frac{X[i]}{\max\{X[i]\}} = k_0 + k_1 \frac{X_{j_1}[i]}{\max\{X_{j_1}[i]\}}, \quad (1)$$

$$\frac{X[i]}{\max\{X[i]\}} = k_0 + k_1 \frac{X_{j_1}[i]}{\max\{X_{j_1}[i]\}} + k_2 \frac{X_{j_2}[i]}{\max\{X_{j_2}[i]\}}\Bigg|_{j_2 \neq j_1}, \quad (2)$$

$$\frac{X[i]}{\max\{X[i]\}} = k_0 + k_1 \frac{X_{j_1}[i]}{\max\{X_{j_1}[i]\}} + k_2 \frac{X_{j_2}[i]}{\max\{X_{j_2}[i]\}}\Bigg|_{j_2 \neq j_1} + k_3 \frac{X_{j_3}[i]}{\max\{X_{j_3}[i]\}}\Bigg|_{\substack{j_3 \neq j_1 \\ j_3 \neq j_2}}, \quad (3)$$

where $X[i]$ – air temperature data; i – data position's number in time series, i=1, 2, 3, ..., 14198 (July19, 1973 – June 1, 2012); $X_{j_1}[i], X_{j_2}[i], X_{j_3}[i]$ – biased (with appropriate delay) time series for the appropriate places; k_0, k_1, k_2, k_3 = [-2;+2] – weighting coefficients (this range allows to find model with physical meaning); j_1, j_2, j_3=0,1,2,...,20 – number of the place in the above list.

The main task is to find weighting coefficients k_0, k_1, k_2, k_3. We will use combinatorial (step is equal to 0.01) inductive modelling with next criterion (minimum of the regularity plus displacement):

$$\alpha_1 \frac{\sum_{\substack{i=1 \\ i \in B_1}}^{13680} |X^*[i] - X[i]|}{6840 \max\{X[i]\}} + \alpha_2 \frac{\sum_{\substack{i=1 \\ i \in B_2}}^{13680} |X^*[i] - X[i]|}{6840 \max\{X[i]\}} + \alpha_3 \frac{\left| \sum_{\substack{i=1 \\ i \in B_1}}^{13680} |X^*[i] - X[i]| - \sum_{\substack{i=1 \\ i \in B_2}}^{13680} |X^*[i] - X[i]| \right|}{6840 \max\{X[i]\}} \to \min \quad (4)$$

where $|.|$ – absolute value, B_1 – first learning sample (odd numbers); B_2 – second learning sample (even numbers); $X^*[i]$ – forecast values; $\alpha_1 = \alpha_2 = 1, \alpha_3 = 10$ criteria's weighting coefficients.

Thus, a formula (1) has next view (temperature measures Fahrenheit degrees):

$$X^*[i] = 92.6\left(1.08 - 0.8\frac{X_2[i]}{100.8}\right). \qquad (5)$$

A criterion (4) has next view on the learning sample (i=1, 2, ..., 13680):
0.0726133559941771 + 0.0726451595651796 +
+ 10 · | 0.0726133559941771 – 0.0726451595651796 | = 0,145576551269382 .

A criterion (4) has next view on the training sample (i=13681, 13682, ..., 14198):
0.0696785289719569 + 0.0671852780183542 +
+ 10 · | 0.0696785289719569 – 0.0671852780183542| = 0.161796316526338
A formula (2) has next view:

$$X^*[i] = 92.6\left(1.29 - 0.49\frac{X_2[i]}{100.8} - 0.54\frac{X_{11}[i]}{103.7}\right). \qquad (6)$$

A criterion (4) has next view on the learning sample:
0.0648588883227404 + 0.0649056479360714 +
+ 10 · | 0.0648588883227404 – 0.0649056479360714 | = 0.130232132392122

A criterion (4) has next view on the training sample (i=13681, 13682, ..., 14198):
0.0638972281191006 + 0.0630384066284932 +
+ 10 · | 0.0638972281191006 - 0.0630384066284932 | = 0.135523849653668

It is clear that criteria's values are identical. Model (3) is not discussed because of high computational complexity (the development perspectives).

Results' analysis shows that forecast model (6) is more precise than (5).

3 Conclusion

In this paper, the short history of the long-term weather forecast inductive method was shown. Its application was realized in the ASP.NET web-site weatherforecast.tk. The realization of some development perspectives are: 1. Partial-page update. Creating a web form that performs partial-page updates to the server is the top-ranked approach of the web-application improvement. It was realized with the AJAX technology (UpdatePanel control) – first element for short-term forecast part and second one for long-term forecast. 2. Cloud computing. It is well known that Windows Azure's web role is reflection of the classic ASP.NET application. One positive feature can be underlined – all processor's time is used for web-application's computation. Therefore, original ASP.NET application was converted to Windows Azure's web role application. 3. Weather forecast's calculation for the definite set of the geographical coordinates is realized on the basis of found simple linear equation. E.g., we have the average daily air temperature forecast's mistake up to 6.5 % for Skopje Airport (half year ahead). 4. The multilingual interface was realized (English and Russian languages).

Recognizing E-Learning Quality in Global Market

Suzana Loshkovska[1], Marjan Miloshevic[2], Danijela Miloshevic[2]

[1]Faculty of Computer Science and Engineering, RugjerBoshkovikj 16, Skopje, Macedonia
suzana.loshkovska@finki.ukim.mk
[2]Technical Faculty Čačak - University of Kragujevac, 65, Svetog Save St., 32000 Čačak Serbia
{marjan, danijela}@tfc.kg.ac.rs

Abstract. The key changes that have taken place in higher education and the working market during the last decades lead to increasing number of institutions that provide e-learning. Issues that are significant for both providers and consumers include the quality assurance (QA) of the e-learning and the recognition of qualification. Generally there are three levels for obtaining QA: institutional, national and international. In the growing globalization and the student exchange schemas, the international QA is becoming extremely important. In that context, the number of bodies and organizations that provide international recognition of e-learning has increased too, and each of them definesits own QA procedures for evaluation and quality recognition. The paper provides an overview of international quality labels and their quality evaluation shemas.

Keywords: e-learning, quality of e-learning, international QA labels.

1 Introduction

The growing availability of educational technologies, expansion in e-learning adoption by institutions, changing of learning paradigma and life-long learning initiatives lead to increasing diversity of student population and offering of e-learning outside higher education institutions or schools [1]. The question of quality is raised and standard quality assuranceprocedures connected only tonational accreditation boards and/or institutional QA bodies are not sufficient. Even more, the growing globalization and establishment of different students' exchange schemas require international recognition of e-learning.

Several surveys are written on quality assurance of e-learning especially that in higher education [2]. One conclusion of these surveys is that QA in e-learning is a non issue for many, especially for the quality assurance agencies. Some reports even suggest that the same criteria for quality should be applied to e-learning and traditional campus-based education. The accreditation, audit and assurance process of e-learning should be integrated in the national framework and not be evaluated separately. This is especially valid for Western Balkan countries where e-courses or e-programs are not differentiated from standard ones in all national and institutional documents.

Initiatives on QA in e-learning that are running for some years noware still restricted to some interested universities. The QA agencies put QA in e-learning only

recently on their agenda and are searching for the expertise for setting the specific criteria and indicators. The expertise and responsibility for QA in e-learning is however in first instance within the universities.
Numerous international projects were developed to form a comprehensive, yet usable framework for quality assurance. Frameworks and accompanying tools that came as result were related to various extents of e-learning, starting from learning units to institution infrastructure. However, it is indicative that many quality schemes developed through european projects had lack of sustainability and are no longer active, nor applicable.

Specialized organizations have developed their own benchmarking procedures and tools and established labels as brands, well recognized and sustainable.First question raised is whether it is possible to establish a unified QA framework. There are few reasons why it is not very realistic to expect such a scenario. First of all, there is a diversity in quality definition, such as described by Donabedian [3].Additionally, there is no unified solution among e-learning standards too, but rather we deal with several different specifications (IEEE, IMS, Ariadne...) and that fact scaffold the claim that we cannot expect a QA in e-learning to be unified soon.
The aim of the paper is to provide an overview of currently active international quality labels for e-learning and reveal potential trends for Balkan countries in this area.

2 Quality of E-Learning

To define quality of e-learning and related standards and procedures that assure e-learning quality, we should start with a definition of e-learning itself. Most widely, e-learning is defined as learning using both a computer and the Internet. Under this definition, we can distinguish different forms of products and services, like single courses and/or entire programs, entire courses and/or course units, lessons or components or elements of an e-learning package (LMS).

Going further, there exist several definitions of qualityof learning systems [4]. Some of them like that in the ISO/IEC 19796-1:2005 standard [5] are too wide and should be adapted to be applied forassessment of e-learning quality. Generally, the definition of the quality of e-learning depends on its scope, objectives, focus and the methodology of the quality approach. For the purpose of the paper we consider that quality of e-learning is related to all: processes, products and services for learning, education and training supported by the use of information and communication technologies.

3 International Quality Labels and Evaluation Schemas

Table 1 presents the active international schemas for quality of e-learning.
UNIQUe. The UNIQUe evaluation schema consists of criteria which are divided into three areas, each with its own criteria, and sub-criteria [6]. The following list of criteria and sub-criteria are evaluated by the reviewers: learning/institutional context, learning resources and learning processes. Additionally several sub-criteria are considered as to be critical to any quality learning experience. Some of them are: available

evidence that eLearning/TEL is an integral part of the institutional strategy; selection of course delivery methods; employed systemic collaborative working procedures and tools to share knowledge developed with the community; available course design and delivery guidelines; flexible pedagogic and learning delivery models; tools and procedures for evaluation of the outcomes of the learning process; continuous promotion of an optimal learning environment; both formative and summative assessment are used; availability of training services and materials for the staff in charge.

Table 1. International Quality Labels

Label name	Level	Review process	Type	Reviewers
UNIQUe	Universities	Quality grid is online	Quality mark	Internal
E-xcellence[7]	University program	Publicly documented; on-line and on-site (optionally)	Benchmark	Internal
ECBCheck	University program	Documents and accompanying web-site are in construction phase	Quality Mark	Internal
SEVAQ+	Universities, Vocational studies	Supporting documents and web-site are are available	Self-evaluation tool	Stakeholders (teachers, students, ...)
Procert [10]	Courseware for IT specialists	Documents are not publicly available	Benchmark	N/A
epprobate	General courseware	Quality grid is public; review conducted on-line	Quality Mark	Internal and external

ECBCheck.ECBCheck[8] evaluates the institution according seven distinct criteria areas: information about and organization of program; target audience orientation; quality of content; program/course design; media design; technology; evaluation and review.Each of these criteria are evaluated according several characteristics. For example, the program or course design the following aspects are evaluated learning design and methodology, students' motivation and participation, learning materials, eTutoring, assignments and learning programs and assessment and tests.

epprobate. To award a courseware with the quality label, epprobate reviewers check the following characteristics [11]:course design(provision of course information, learning objectives and instructional guidance and constructive alignment);learning design(the courseware fulfill the following criteria learner needs, personalization and instructional strategies);media design(media integration, interface, interoperability and technological standards); and content(accuracy and values of content, intellectual property rights, legal compliance). It is important to mention that there is no intention that any single criterion be essential, but rather that a courseware supplier should in their self assessment document indicate to what extent they meet a specific criteria.

SEVAQ+. The SEVAQ+ is based on EFQM™ quality framework and Kirkpatrick evaluation model [9].The EFQM Excellence Modelis used as a basis for self-assessment of organizations. Each organization grades itself against the nine criteriaThrough the nine criteria the organization can understand and analyses the cause and effect relationships between what the organization does and the results it achieve. Five of these criteria are denoted as 'Enablers' and four as 'Results'. The 'Enabler'

criteria cover what an organization does and how it does it. The 'Results' criteria cover what an organization achieves.This model is modified in the part of 'Results' using the evaluation model of learning elaborated by Kirkpatrick to be applied in a context of education. The Kirkpatrick model evaluates: the students' reaction or feelings of the students during learning; the learning result, or the increase in the knowledge of the learner by taking part in the course; the impact on the learner's functioning in the workplace, or transfer of new knowledge to skills; and the impact on the business results as a consequence of skilled people.

4 Conclusion

Taking into account that there is a not unified quality standard for e-learning, the paper presents several quality labels for international recognition of e-learning. Labels are distinguished by the context and the scope, objectives, focus, perspective, methodology and metrics. UNIQUE, e-xcellence, ECBCheck and SEVAQ+ are focused on institutional evaluation, while Procert and epprobate are focused on courseware. Finally, although, the labels are focused on different forms of e-learning, most of them as important aspects of quality consider: information about the organization, target group of learners, design of learning, quality of content, media design and technology, and evaluation and assessment.

Acknowledgements. The research for this paper was done in the framework of DL@WEB project (511126-TEMPUS-1-2010-1-RS-TEMPUS-SMGR) funded by the European TEMPUS program.

References

1. McLoughlin, C., Visser, T., "Quality e-learning: Are there universal indicators?", 16th ODLAA Biennial Forum Conference Proceedings 'Sustaining Quality Learning Environments', 2003 (1)
2. -, E-learning Quality (ELQ) – Aspects and criteria for evaluation of e-learning in higher education. Report 2008:11R, Swedish National Agency for Higher Education, 2008(2)
3. Donabedian, A ., The Definition of Quality and Approaches to Its Assessment, Ann Arbor: Health Administration Press, 1980 (6)
4. Ehlers,U.D., Pawlovski, J.M., Handbook on quality and standardization in e-learning, Springer Berlin - Heidelberg 2006(7)
5. International Organization for Standardization/International Electrotechnical Commission,*ISO/IEC19796-1:2005. Information Technology - Learning, Education, and Training - Quality Management, Assuranceand Metrics - Part 1: General Approach*. International Organization for Standardization, 2005 (8)
6. http://unique.efquel.org/(Accessed in July 2012)
7. http://www.eadtu.nl/e-xcellencelabel/ (Accessed in July 2012)
8. http://efquel.org/(Accessed in July 2012)
9. http://sevaq.efquel.org/(Accessed in July 2012)
10. http://www.procert.com/ (Accessed in July 2012)
11. http://epprobate.com/(Accessed in July 2012)

Artificial Neural Networks in CRM

Rexhep Rada[1] and Bashkim Ruseti[2]

[1]Department of Mathematics and Informatics, Faculty of Natural Sciences,
"Aleksandër Xhuvani" University, Elbasan, Albania
rexhep.rada@gmail.com
[2]Department of Mathematics, Statistics and Applied Informatics,
Faculty of Economics, University of Tirana, Tirana, Albania
bruseti@gmail.com

Abstract. In recent years, Data Mining becomes a very important technique supporting in business decisions. From the modern techniques, artificial neural networks have improved a lot the companies in their success. Artificial neural networks allow creating and using complex functions in a natural mode, to make useful prediction. In this paper we will evaluate the use of ANN in a Customer Relationship Management (CRM) database. Our goal is to define the target group, who spends more in mobile communication in Albania.

Keywords: Artificial Neural Networks, Customer Relationship Management (CRM), Data Mining

1 Introduction

The process of determining the target group interested more in a certain product is on the top of the agenda of every business company. Technical implementation of these methods in Western countries is widespread, but in Albania doesn't exist. Companies use as prediction tools, the descriptive statistics, without entering in more sophisticated methods. [2]

In this article we will try to give an overview of the application of ANN in mobile communication. It consists in analyzing an old database of customers, connecting the characteristics of the clients with the amount of money that they spend in phone calls. The data are taken from a survey in 930 users. [3]

2 Methodology

The data are dividing in two parts: Data Training Set and Data Validation Set. This case study is a based in supervised learning, more specifically in back-propagation algorithm.

We try to regulate the weights in ANN for improving prediction from training set. [4]

S. Markovski, M. Gusev (Editors): ICT Innovations 2012, Web Proceedings, ISSN 1857-7288

The training data consist in a large database with more than 620 records, with respective input and output. In this case the variables are: age, marital status, social status, personal income, family income, number of persons in family, and the outcome is the monthly expenses for mobile communication. (Table 1)

Table 1. Database sample input

Name	Surname	Age	marital status	Social status	Personal income	Family income	Nr. of persons in family	Nr. of brothers / sisters	Monthly expenses
Ina	Sina	19	single	student	0	50,000	4	2	1,500
Miri	Hoxha	28	single	employed	100,000	200,000	3	3	5,000
Refit	Sulmina	20	single	employed	10,000	50,000	4	3	4,000
Blerta	Teta	35	married	employed	10,000	40,000	5	7	2,500
Valbona	Filja	39	married	employed	17,000	100,000	8	7	2,000

2.1 Variables

We have to make every variable numerical. But beyond this, the numbers have to been normalized. We do this because we don't want that big numbers of not weight variables affect more in the final prediction. [1]

We can use from dummy variable, which equals 0 and 1 in two value variable domain to proportional value in the]0;1[. Let see a version of numerical table (table 2).

Table 2. Database sample input with numerical variables.

Name	Surname	Age	marital status	Social status	Personal income	Family income	Nr. of persons in family	Monthly expenses	Nr. of brothers/ sisters
Ina	Sina	19	1	0.5	0	50,000	4	1,500	2
Miri	Hoxha	28	1	1	100,000	200,000	3	5,000	3
Refit	Sulmina	20	1	1	10,000	50,000	4	4,000	3

Blerta	Teta	35	0	1	10,000	40,000	5	2,500	7
Valbona	Filja	39	0	1	17,000	100,000	8	2,000	7

After the numeric step, we can perform other processing methods like min-max standardization, to normalize the values with the aim that the highest number doesn't affect more in prediction. [5]

2.2 Results

For each exemplar of 310 records of validation dataset (remaining from the database with 930 records) we make the prediction with several ANN models. After that we compare the outcome of the prediction with the outcome in dataset. We see the percentage of the right predictions. Table 3 displays the results.

Table 3. Full results

Architecture (with the number of the hidden layers)	**Dataset training**			**Dataset validation**		
	MSE	**1-MSE**	**Good classification rate**	**MSE**	**1-MSE**	**Good classification rate**
Net_00	9.6	90.4	89.9	20.6	79.4	74.5
Net_01	3.7	96.3	96.4	4.54	95.46	82.5
Net_02	1	99	95.35	2.64	97.36	86.8
Net_03	0.5	99.5	97	0.8	99.2	90.1

We notice from tables, that the use of hidden layers improves substantially the model.

In training sample, the best version (Net 03) has the lowest mean square error (0.5%) and allows us the classification rate of 97%. In validation set we notice that the best network is the same, the classification rate jumped to 90.1% with MSE=0.8%.

3 Conclusion and future work

On finding the right target group of a product depends the profit of every company. This article gave an overview of Artificial Neural Networks and its applications. This technique have to be used even in Albania to improve the nowadays analyzes. We described a case study in mobile communication. Is used a database from a survey of

930 records. Comparing the prediction after training with the real results in database we got an amazing result. In validation set we notice that the best network gave a classification rate of 90.1% with MSE=0.8%.

The future direction of this study will be on improvement of the activation function. These helpful results will encourage the Albanian company for a fast introduction of Data mining techniques in their company.

References

1. Classification with Artificial Neural Networks And Support Vector Machines: Paolo Valigi, Vidas Gulbinas, Rainer Westphal, Khaled Mohamed Almhdi And Rainer Reuter.
2. CRM through DM: A Case Study Angelo M. Cister, Nelson F. F. Ebecken.
3. Bankruptcy Prediction for Credit Risk Using Neural Networks: Amir F. Atiya.
4. Neural Networks By Christos Stergiou And Dimitrios Siganos.
5. Artificial Neural Networks (Ann)Goodchild, Longley.

Development overview of TTS-MK speech synthesizer for Macedonian language, and its application

Slavcho Chungurski [1], Sime Arsenovski [1], Dejan Gjorgjevikj [2]

[1] Faculty of informatics, FON University – Skopje
{chungurski, sime.arsenovski}@fon.edu.mk
[2] Faculty of Computer Science and Engineering, University of "Sv. Kiril i Metodij" - Skopje
dejan.gjorgjevikj@finki.ukim.mk

Abstract. This paper shows the current results of development of TTS-MK – a speech synthesizer for Macedonian language. The basic principles for projecting and building of speech synthesizer for Macedonian language, based on concatenation of speech segments, are shown. Every language has its respective and specific speech norms and characteristics that should be observed during the speech synthesis. The Macedonian language is phonetic; hence the normative pronunciation does not contain great difficulty, except in some special cases that should be taken into consideration. The presentation also focuses on the accent in the Macedonian language, which is dynamic and positioned on the third syllable. The rules and regulations for the accent positioning in the Macedonian language can be easily derived, with some deviations that should be resolved. There are two versions of the system based on different segments corpora. Both of them are presented, as well as their application.

Keywords: Text-To-Speech, Macedonian Language, TTS, TTS-MK, Orthoepy, Speech API – SAPI

1 Introduction

The systems which synthesize speech with connection of previously recorded speech segments take significant place among the systems for text to speech conversion (TTS systems). These TTS systems are called concatenative speech synthesizers. These systems are simple and they do not require deep knowledge of phonetic transitions and co-articulation effects, which is the case with other kinds of speech synthesizers based on rules defined by linguists.

There were some attempts for development of quality concatenative speech synthesizer for Macedonian language, but these developments were based on speech corpora for other Slavic languages, which resulted in unnatural intonation of the synthetic speech in Macedonian language. This paper includes a brief overview of the development of TTS-MK synthesizer for Macedonian language. Concatenative speech synthesizers require setting of serious task for definition and recording of speech and its processing for extracting convenient speech segments. Consequently,

S. Markovski, M. Gusev (Editors): ICT Innovations 2012, Web Proceedings, ISSN 1857-7288

this paper presents appropriate definition and development of speech corpus for Macedonian language, used in TTS-MK.

The general functional organization of speech synthesis system for Macedonian language is shown in Fig.1. As in [3], it is made of two modules:

- Natural Language Processing (NLP) module, that gets text as input, makes analysis of the text, create its transcription into phones and recognizes the prosodic elements of the input. The output of this module is symbolic information for the phones and the prosody for the input text.

- Digital Signal Processing (DSP) module, that gets the symbolic information for phones and prosody from the NLP module and after certain processing of the input gives synthetic speech as output.

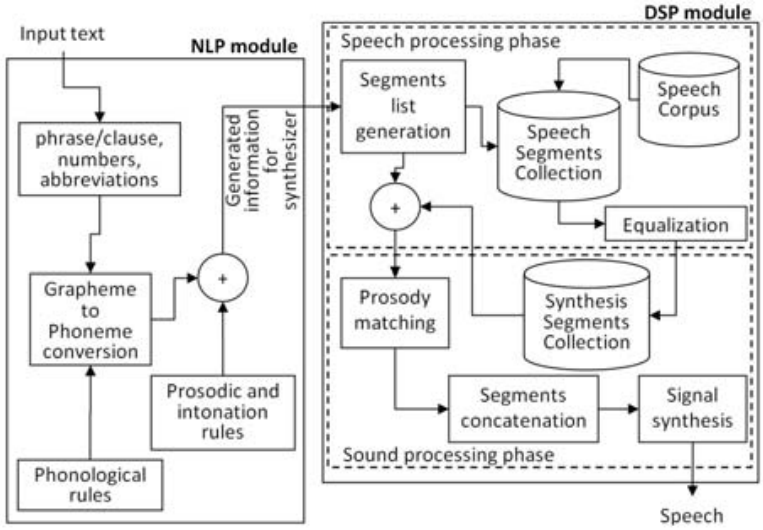


Fig. 1. Structure of general TTS system

The NLP module consists of three main parts: text analyzer, grapheme-to-phoneme unit and prosodic generator. Text analyzer is made of pre-processing unit where the input sentences are transformed into word arrays. This unit also identifies the numbers and abbreviations and transforms them into words. There is a module for morphological analysis that performs morphological analysis of the text for recognition of the affixes (prefixes and suffixes) that are added to the basic forms of the words. This module also confirms the correct accent of the words. This module of TTS systems for Macedonian language seems to be very complicated because in Macedonian language, unlike other languages, a plural of the words as well as words with an articles, adds whole syllables at their end. This can look very difficult to implement and it suggests to some complications in morphological analysis of the text. Thanks to antepenultimate word stress in Macedonian language and the fact that this analysis is strictly for speech synthesis, this procedure is very simple. This is supported with the fact from [4], that in the Macedonian literature language there is no reduction of voc-

als, which means that the vowels sound very similar in their accented and unaccented form. As a result of this fact, this considerably simplifies the whole speech synthesis process for Macedonian language, because the vowels have only one sound variant, and not multiple, as in French for example (described in [3]). One of the main features of Macedonian language is its very simple rule for grapheme to phoneme transcription. This means that in Macedonian language, like in other Slavic languages, each letter represents one phone. So, the phonetization task for Macedonian language is reduced to trivial checks so solution techniques based on dictionaries and morphophonemic rules, like in English or French for example, are not necessary. The task of the prosodic generator is to supply naturalness to the synthesized speech. Apart from the fact that synthesized speech with natural intonation is more pleasant for listening, it is easier for understanding. Also, during a speech which is practically uninterrupted, the listener can easily recognize the bounds between the words. It is very important for a speech synthesizer to cover prosodic elements and proper intonation, which thanks to accentuation rules is not very complicated in Macedonian language.

The execution in DSP module is divided into two phases: speech processing and sound processing. Speech processing is one of the most important phases in TTS synthesis in general. It means that in this phase it is mandatory to make definition and recording of the speech corpus for Macedonian language, as well as its segmentation. In this phase, the symbolic information for the phones and the prosody for the input text are applied on the recorded speech corpus. The segments from the corpus and their concatenation order are also established in this phase. The collection of segments for speech synthesis obtained from the previous phase is the actual input for the next phase of DSP module - the sound processing. In this phase an adjustment of prosodic elements is performed. Here, as it was described in the part for text analysis, an accentuation is performed. After reconstruction of accentuation aggregate, f0 codebook is implemented in this phase. f0 codebook holds information for the f0 curves for each accentuation aggregate obtained by empirical way and the selection of appropriate curve is made on base on the position of the accentuation aggregate in the sentence (beginning, neutral, before comma, end).

2 Orthoepy of Macedonian language

Orthoepy is normative pronunciation of some standard language and its proper investigation leads to a better speech synthesis for the language. Macedonian language consists of 5 vocals (а /a/, е /ɛ/, и /i/, о /ɔ/, у /u/) and 26 consonants (б /b/, в /v/, г /g/, д /d/, ѓ /ʤ/, ж /ʒ/, з /z/, ѕ /dz/, ј /j/, к /k/, л /l/, љ /ʎ/, м /m/, н /n/, њ /ɲ/, п /p/, р /r/, с /s/, т /t/, ќ /c/, ф /f/, х /x/, ц /ts/, ч /ʧ/, џ /dʒ/, ш /ʃ/).

Every language has its respective and specific speech norms and characteristics that should be observed during the speech synthesis. The Macedonian language is phonetic, hence every voice corresponds to particular grapheme, so normative pronunciation is not difficult. However, deviations do occur in written and spoken part of the language (discharging, substitution, inserting, voice replacement)[4]. These deviations must be taken into consideration according to the rules of the normative pronun-

ciation during the speech synthesis. The mentioned aspects are considered within the NLP module of TTS-MK. These are some cases which are considered in TTS-MK:

- Double vocals pronunciation
  - Case: when adding a prefix of a word (Example: poodi → po|odi)
    - Solution: dictionary of prefixes
  - Case: when doubling is in the middle or the end of the word, and the accent is not on any of these vocals (Example: vikaat → vikāt; vakuum → vakūm, Exception: E, both vocals are pronounced separately)
    - Solution: processing module
- Vocal R
  - Case: R, where it is the leading grapheme followed by a consonant, is noted with ‘R (Example: ‘rž ; ‘rgja)
    - Solution: phoneme inventory contains @ (schwa)
  - Case: Adding a prefix, which ends with consonant, to words from the previous case leads to discharging of the sign (‘) in the notation, but not in the pronunciation (Example: srska, srža)
    - Solution: R → @R replacement
  - Case: R, where it has a role of a vocal when it is surrounded by consonants (Example: srce → s@rce; brza → b@rza)
    - Solution: R → @R replacement
- Consonants pronunciation
  - Case: Consonant sonority at the end of the word B → P; V → F; D → T; Dz → C; Z → S; Ž → Š; Dž → Č; G → K (Examples: grob →grop; gluv → gluf; led → let; bez →bes; nadež →nadeš; Džordž → Džorč; plug →pluk)
    - Solution: processing module
  - Case: TS → C; DS → C; SS → S; ZZ → Z; ŽD → ŠT (Examples: gradski → gracki; bratski → bracki; bessovesen → besovesen; bezzimen → bezimen; glužd → glušt)
    - Solution: processing module
  - Case: other isolated cases (Examples: ovca → ofca; vkluči →fkluči; gladta → glatta)
    - Solution: None

Macedonian accent is observed in the text analyzer and it can be processed with ease because it is dynamic and positioned on the third syllable (when more than two syllables), but there are some deviations. These cases are observed in TTS-MK:

- Adopted foreign words (Examples: foajè; birò; *ìzam, *ìst etc)
  - Solution: dictionary
- Accent aggregates (Examples: Kisela voda → Kiselàvoda; pri toa → prìtoa)
  - Solution (bad): dictionary, and then third syllable rule

3 Structure of the speech corpus of TTS-MK

Several continuous steps were performed for definition and creation of the speech corpus for TTS-MK[1]:

- Selection of speech segment types needed for the synthesis
- Definition of the set of the phonemes that covers all sound occurrences in Macedonian language
- Selection of the set of speech segments to be used that cover the whole phoneme set from the previous step
- Selection of the texts that cover whole speech segment set from the previous step

As a result of this procedure the following results are achieved:

- Segment types (Diphones, disyllables, whole accented words)
- Phoneme set (TTS-MK phonetic inventory adds ŋ /ŋ/ (allophone of n), ь /ə/ (schwa) and _ (silence))
- Selection of speech segments (34x34 = 1156 diphone units and 150 frequent disyllables)
- Selection of texts for recording of the segments (existing words from a text corpora and logatoms)

The elements required for speech synthesis in the TTS-MK are stored in several files shown in Table 1.

Table 1. Storage of TTS-MK corpus

Filename	Protection	Contents
diphone.mk	Encrypted	Diphone set
disyllable.mk	Encrypted	Disyllable set
words.mk	Encrypted	Words set
abbreviations.mk	Open	Abbreviation list
accentAggregates.mk	Open	Accent aggregates list
accentExclusions.mk	Open	Accent exclusions list

There are two versions of TTS-MK speech corpora.
TTS-MK Emma is the first experimental version. It includes 1156 diphones, 180 disyllables and 392 words from total 1196 recorded segments. It has poor dictionaries for abbreviations, accent exclusions and accent aggregates, contains lot of noise and includes low prosody elements. The whole corpus requires 56 MB of disk storage.

TTS-MK Lence is the second version and it is currently in development. It includes all 1156 diphones, plus 1161 disyllables and 392 words from a total of 1144 recorded segments. Because of its quality, and the quality of the speaker and the tone master, it can be commonly assumed that it is upgradeable. It includes rich dictionaries for abbreviations, accent exclusions and accent aggregates, without noise and with some prosody elements. It requires 50 MB of disk space. This version is also SAPI 5.3

compatible with rate, pitch and volume adjustments and it is expected to be deployed as a basic tool for blind computer users. It also has some UI settings like Latin texts (spell or phonetic), numbers (digit by digit) and interpunction.

TTS-MK engine is developed completely in C# and .NET framework version 2.0. To interface this engine with SAPI a special engine wrapper was developed in C++. Figure 2 shows the application of TTS-MK for visually impaired persons.

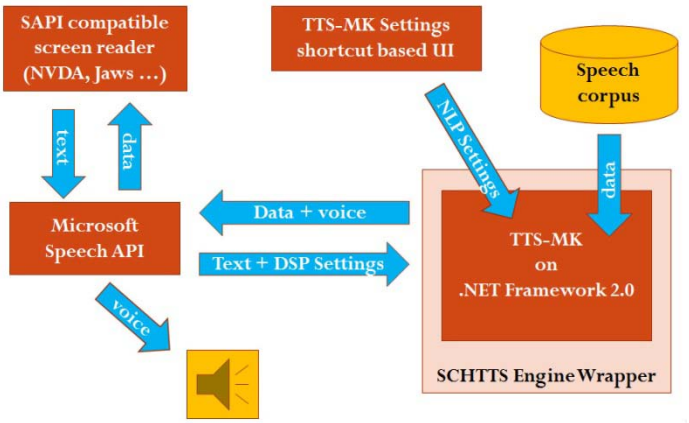


Fig. 2. Application of TTS-MK

4 Future work

Future work on TTS-MK can be separated in two parts. Improvements of NLP module of TTS-MK which will include improvements of orthoepy preprocessor, and improvements of DSP module of TTS-MK which will include unit-selection algorithms and upgrade of the corpus.

References

1. Chungurski, S., Kraljevski, I., Mihajlov, D., Arsenovski, S.: Concatenative Speech Synthesizers and Speech Corpus for Macedonian Language. 30th International Conference ITI Cavtat/Dubrovnik, Croatia. 669-674 (June 2008)
2. Chungurski, S., Kraljevski, I., Mihajlov, D., Arsenovski, S.: Evaluation of TTS-MK system for speech synthesis in Macedonian language. ETAI 2009, Ohrid. IE3-3. (September 2009)
3. Dutoit, T.: High Quality TTS Synthesis of the French Language. Ph. D. dissertation, Facultè Politechnique de Mons, France. (1993)
4. Koneski B.: Gramatika na makedonskiot literaturen jazik. Kultura, Skopje. (1987)
5. Chungurski, S., Kraljevski, I., Kakashevski, G., Mihajlov, D.: TTS approach to Macedonian Language. 16th Telecommunication forum TELFOR 2008, Belgrade. (November 2008)

Noise Robustness of Traditional Features for Macedonian Voice Dialing ASR

Branislav Gerazov and Zoran Ivanovski

Faculty of Electrical Engineering and Information Technologies, Skopje, Macedonia
{gerazov, mars}@feit.ukim.edu.mk

Abstract - Automatic Speech Recognition Systems of today are intensely deployed in real world application scenarios which are often characterized by suboptimal operating conditions. Thus their noise robustness has become a crucial parameter when assessing ASR in-field performance. The paper examines the noise robustness of traditional ASR feature sets as applied to a Voice Dialing Application built for Macedonian. The analysis focused on the following features: Linear Prediction Reflection Coefficients, Mel-Cepstral Cepstral Coefficients and Perceptual Linear Prediction Coefficients. The ASR system was trained with clean data, and in the evaluation phase the noise level in the test data was varied by adding white and babble noise. Results have been plotted for each feature type across varying SNR conditions.

Keywords. ASR, features, noise robust, noise, SNR

1 Introduction

The field of Automatic Speech Recognition (ASR) is an area of active scientific research for the last 70 years, [1]. The basic task of ASR systems is to transform the input speech audio signal into the corresponding sequence of words. To carry through this complex task ASR systems incorporate algorithms and methods from various fields such as digital signal processing, statistical modeling, machine learning, natural language processing etc., [2]. A generic ASR system consists of a signal processing frontend and a modeling and recognition backend. The frontend's task is to analyze the input audio and to extract from it the acoustic events relevant to the recognition task at hand. This information is to be represented in terms of a compact, efficient set of speech parameters, called features, [3]. The backend uses these features to analyze and recognize the phonetic content of the input speech signal through comparison with a trained set of models, outputting its semantic identity.

From the earliest days of ASR, it was clear that the features used to describe the acoustic realization of phones are of fundamental importance to the quality of the overall results of the speech recognition process, [1]. This is even more the case for the application of ASR systems in noisy environments. ASR systems on mobile intel-

S. Markovski, M. Gusev (Editors): ICT Innovations 2012, Web Proceedings, ISSN 1857-7288

ligent platforms must work well in application scenarios such as busy streets, supermarkets, train stations, in the car, at a cocktail party etc.

The paper gives an analysis of the robustness of traditional ASR features in noise environments when used in a Macedonian ASR system. The features taken under consideration are the Linear Prediction Cepstral Coefficients (LPCCs), Mel-Frequency Cepstral Coefficients (MFCCs), and Perceptual Linear Prediction Coefficients (PLPs). The ASR system is speaker based, with a small vocabulary of 36 words, [4]. It is designed as a prototype for a more advanced voice dialing ASR system to be built for Macedonian. The noise robustness of the considered features was tested using 5 types of noise: white, pink, babble, in-car and traffic noise. The accuracy of the ASR system was used to assess the noise robustness of the different features.

2 ASR System

The ASR system used for the analysis is a prototype system built for a voice dialing application in Macedonian. It is speaker-dependent, with a small-vocabulary of 36 words containing commands, names, and digits. The back-end of the system is based on context-dependent triphone HMMs with single-mixture GMMs. The recognition is carried through using a Viterbi based token-passing algorithm. The system was developed using Cambridge University's HMM toolkit – HTK, [5].

The database used for training the HMMs and used for evaluating the system's performance comprises 500 utterances with a total of 2010 words and a total recording length of more than 20 minutes. Table 1 gives the spread of the recorded material into training and test sets, at a 62:38 ratio. The test set, numbering 760 words allows for a measure of the ASR system's performance within 0.13%. The system uses a simple task grammar in which every user demand starts with a command word that can be optionally followed by a number or a name.

Table 1. Spread of the recorded material used in the ASR system's database

	Training set	Test set	Total
Utterances	380	120	500
Words per utterance	3	6	4
Words total	1250	760	2010
Duration	12:57	05:55	18:52

3 Features Used

Three types of features were taken into consideration for this analysis: Linear Prediction Cepstral Coefficients (LPCCs), Mel-Frequency Cepstral Coefficients (MFCCs), and Perceptual Linear Prediction Coefficients (PLPCs).

Developed in the 1960's in ASR research, the Linear Predictive Coding (LPC) analysis technique grew to be a common frontend choice for many ASR systems of the time. LPC analysis is based on the powerful source-filter model of the speech

production process, [3]. LPC provides good modeling of the spectral content of the speech signal, especially for quasi steady state voiced regions of speech, it gives a reasonable source – vocal tract separation, it is analytically tractable, with a low computational cost, and finally it has been shown to work well in ASR applications, [3]. Because of this LPC has been extensively used in ASR system frontends. In the analysis 12 LPCCs were extracted using recursion from the parameters of the LP filter of order 14.

The Mel-Frequency Cepstral Coefficients (MFCCs) are a prevalent choice in today's ASR systems. MFCCs are based on Mel frequency scale power cepstrum analysis, [6]. In the analysis the amplitude spectrum of the signal was extracted using the Fast Fourier Transform (FFT) and is processed through a bank of 26 triangular filters equidistant on the Mel-scale. 12 Cepstral Coefficients were then obtained using the Discrete Cosine Transform (DCT) from the log filterbank amplitudes.

One extension to the LPC model analysis, which brings it closer to the human auditory model, is the method of Perceptual Linear Predictive (PLP) analysis, [7]. PLP uses concepts from the psychophysics of hearing to derive an estimate of the auditory spectrum. It models well human hearing and gives superior results to both LPC and MFCC, especially in noisy environments, [8]. In the analysis the Mel filterbank coefficients were weighted with an equal-loudness curve and then compressed applying the cubic root. 12 Cepstral Coefficients were then extracted from LP parameters estimated from this auditory spectrum.

For all of the features extracted the speech was windowed with a Hamming window of 25 ms at 10 ms steps and frames were pre-emphasized using a first order difference with a coefficient of 0.97. Also re-scaling of the coefficients was carried through with a lifter of length 22. For all feature vectors energy coefficients and first and second order derivatives were calculated.

4 Adding Noise

The noise robustness of the presented features was assessed adding 5 types of noise to the test set recordings. These were white, pink, babble, in-car, and city traffic noise. The first four were provided by Institute for Perception-TNO, the Netherlands, and have duration of 4 minutes each. The last was obtained from SoundBible and has duration of 12 seconds. Fig. 1 shows the spectra of the five types of noise used in the analysis. Noise was added from these recordings by selecting a random part with the appropriate duration. The level of the noise was processed according to the level of the recording from the test set it's been added to before addition. Thus the 6 types of noise were added at 5 different signal-to-noise ratios (SNRs) that together with the clean recordings resulted in 31 different versions of the test set for assessing the noise robustness of the traditional feature sets.

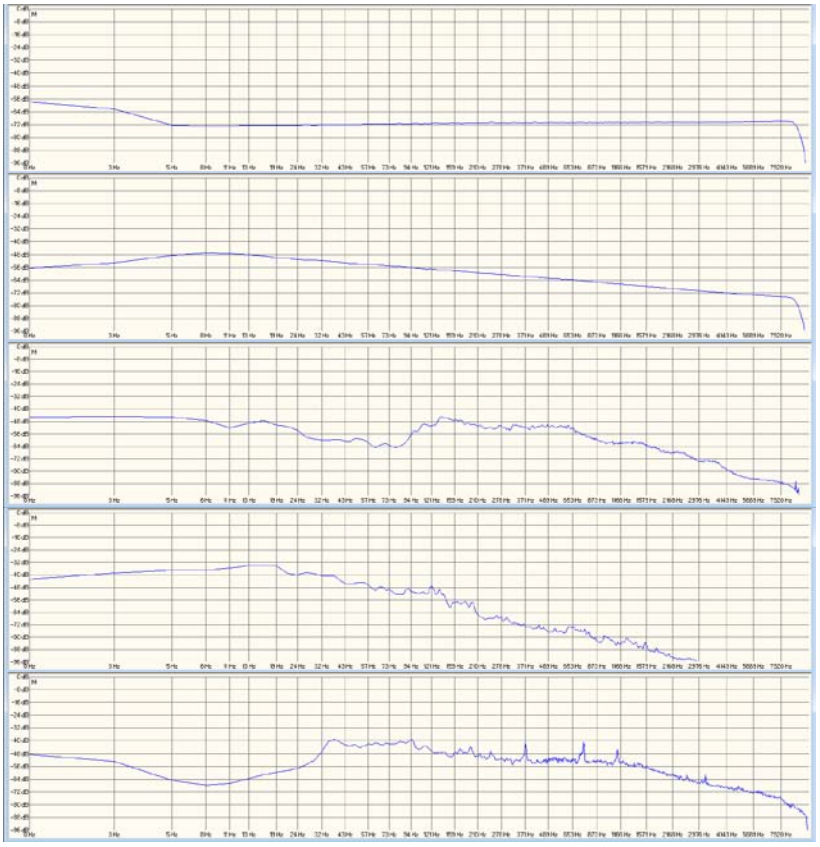


Fig. 1. Spectra of the five types of noise used in the analysis (*top to bottom*): white, pink, babble, in-car, city traffic

5 Results

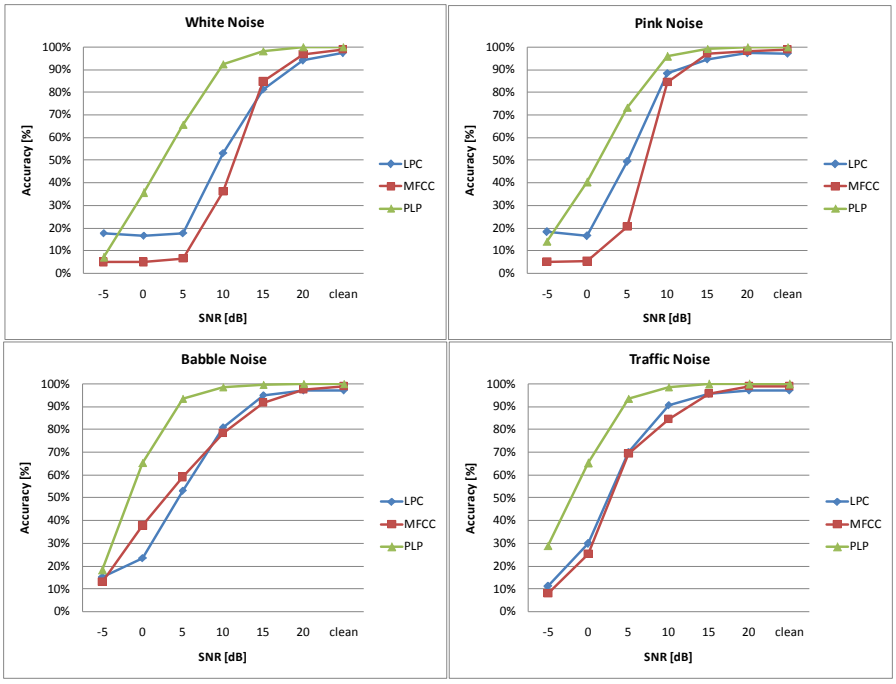
The three presented features were used in the front-end of the ASR system and its performance was assessed for each of the noise types for each of the SNRs, as well as the clean test set. The performance was taken as the ASR system's accuracy at the word level calculated using Eq. 1, where N is the total number of words, D is the number of deletions, S is the number of substitutions and I is the number of insertions.

$$Acc = \frac{N - D - S - I}{N} \cdot 100\% \tag{1}$$

Table 2 gives a spread of the results obtained for the system's accuracy for the various noise types and SNRs. Figs. 2 and 3 give plots of the system performance in respect to the noise type and features used. All of the plots show SNR on the x-axis and system accuracy on the y-axis.

Table 2. ASR system accuracy in respect to type of noise and SNR for the 3 different feature sets

SNR	LPC	MFCC	PLP	LPC	MFCC	PLP	LPC	MFCC	PLP
		white			**babble**			**in-car**	
-5	17,76%	4,87%	6,97%	15,26%	13,16%	18,29%	11,18%	8,16%	28,95%
0	16,71%	4,87%	35,53%	23,55%	37,76%	65,39%	29,87%	25,39%	65,26%
5	17,76%	6,58%	65,79%	53,16%	59,08%	93,68%	70,13%	69,47%	93,55%
10	53,03%	36,05%	92,50%	81,05%	78,42%	98,55%	90,66%	84,61%	98,68%
15	81,18%	84,61%	98,16%	95,13%	91,84%	99,74%	95,79%	95,79%	100,00%
20	94,08%	96,71%	100,00%	97,24%	97,50%	100,00%	97,24%	98,95%	100,00%
clean	97,24%	98,95%	100,00%	97,24%	98,95%	100,00%	97,24%	98,95%	100,00%
		pink			**traffic**				
-5	18,29%	5,13%	14,21%	47,11%	95,00%	100,00%			
0	16,45%	5,26%	40,26%	62,63%	98,68%	100,00%			
5	49,34%	20,66%	73,29%	75,79%	98,95%	100,00%			
10	88,42%	84,47%	95,92%	92,50%	99,08%	100,00%			
15	94,61%	97,24%	99,34%	96,45%	98,95%	100,00%			
20	97,37%	98,29%	100,00%	96,97%	98,95%	100,00%			
clean	97,24%	98,95%	100,00%	97,24%	98,95%	100,00%			



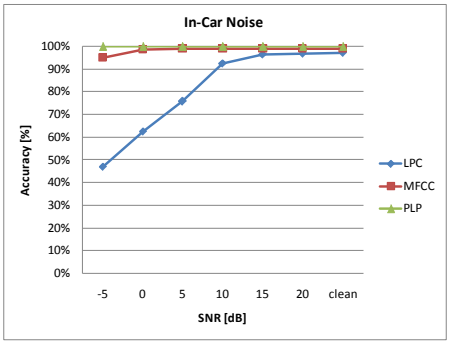


Fig. 2. ASR system accuracy in respect to noise type for the three different feature sets

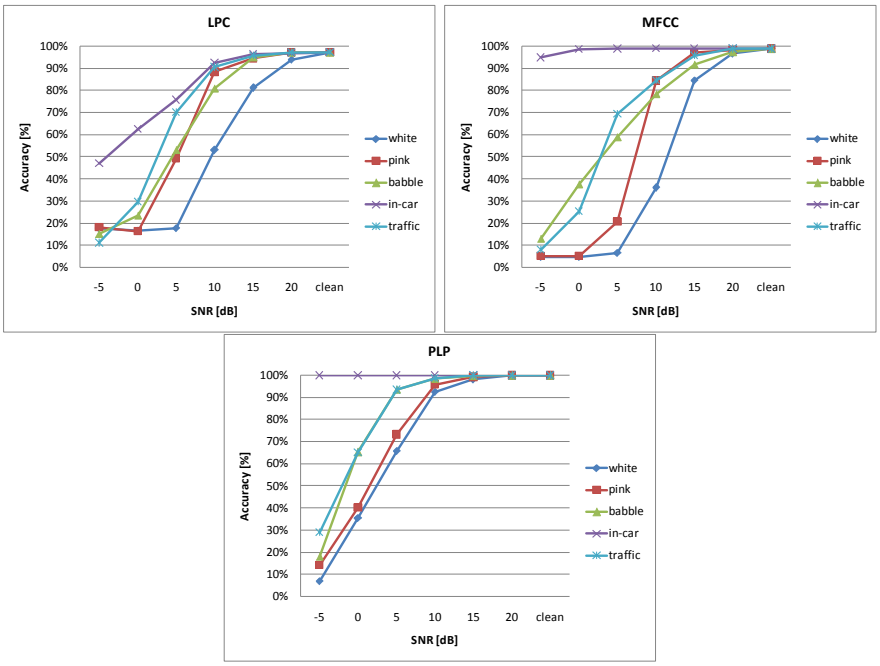


Fig. 3. ASR system accuracy in respect to the feature set for the 5 different types of noise

Table 2 and Figs. 2 -3 clearly show that PLPCs fare better in noise than MFCCs and LPCCs, granted by their added complexity. Also MFCCs surpass LPCCs in general, while LPCCs can be seen to outperform MFCCs in white and pink noise at lower SNRs. All of the features are more susceptible to white noise, than pink, than babble and traffic. On the other hand, in-car noise does little to hinder MFCC and PLPC performance.

6 Conclusion

From the presented results it can readily be concluded that from the three analyzed feature sets, PLPCs are more robust to noise than MFCCs and LPCCs. Thus if their added complexity can be handled by the computational power of the ASR system's target deployment hardware, then they are a clear choice to be made. They will be the features of choice for the final implementation of our ASR system.

References

1. Rabiner L., B. H. Juang: Historical Perspective of the Field of ASR/NLU, In: Springer Handbook of Speech Processing, Benesty J., Sondhi M. M., Huang Y. Eds., Springer-Verlag Berlin Heidelberg, 521 – 537 (2008)
2. Jurafsky Daniel and James H. Martin: Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition , 2nd Ed., Prentice Hall (2009)
3. Rabiner L.R., B. H. Juang: Fundamentals of Speech Recognition, Prentice Hall (1993)
4. Gerazov B., Z. Ivanovski: Prototype Automatic Speech Recognition System for a Voice Dialing Application for Macedonian, Summer Symposium on Electronics and Signal Processing LEOS 2012, Mavrovo, Macedonia, Sep 14 – 15 (2012) (in Macedonian)
5. The Hidden Markov Model Toolkit (HTK), `http://htk.eng.cam.ac.uk/`
6. Davis S.B., P. Mermelstein: Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences, In: IEEE Trans. Acoust. Speech Signal Process. 28(4), 357–366 (1980)
7. Hermansky H.: Perceptual linear predictive (PLP) analysis of speech, In: J. Acoust. Soc. Am. 87(4), 1738–1752 (1990)
8. Young S.: HMMs and Related Speech Recognition Technologies , In: Springer Handbook of Speech Processing, Benesty J., Sondhi M. M., Huang Y. Eds., Springer-Verlag Berlin Heidelberg, 539 - 557 (2008)

S. Markovski, M. Gusev (Editors): ICT Innovations 2012, Web Proceedings, ISSN 1857-7288

Towards Hierarchical Cognitive Systems for Intelligent Signal Processing

Rüdiger Hoffmann[1] and Matthias Wolff[2]

[1] Technische Universität Dresden,
Professur Systemtheorie und Sprachtechnologie, 01062 Dresden, Germany
ruediger.hoffmann@tu-dresden.de
http://www.ias.et.tu.dresden.de
[2] Brandenburgische Technische Universität Cottbus,
Lehrstuhl Kommunikationstechnik, 03046 Cottbus, Germany
matthias.wolff@tu-cottbus.de
http://www.tu-cottbus.de/kommunikationstechnik/

Abstract. Speech and many other acoustic signals show a hierarchical structure which has to be considered if systems for speech and audio processing are developed. Because systems for speech recognition and for speech synthesis follow the same hierarchy, a unified approach (UASR) was proposed in the year 2000, which was implemented during the following decade. The general application of Finite State Transducers (FST) results in a very efficient technology at all symbolic levels of the hierarchy. UASR proved to be successful not only for speech processing but also for many applications to other biological, technical, or musical signals, resp. At the same time, the idea of cognitive dynamic systems became popular mainly due to the work of S. Haykin. It is very promising to expand the UASR system to a hierarchical cognitive dynamic system, combining the hierarchical structure of UASR with the approach of cognitive systems, which is mainly elaborated for the so-called cognitive radio so far. The target system, the structure of which was defined now, will perform intelligent processing of speech and other signals.

Keywords: intelligent signal processing, hierarchical systems, cognitive systems, acoustic pattern recognition, audio processing

1 Introduction

Speech processing systems are basing on two basic principles which were formulated, e. g., in one of the fundamental papers on speech recognition [1]: "The first fundamental hypothesis is based on consistent evidence [...] indicating that the primary information-bearing attitude of the speech signal is the temporal variation of the short duration amplitude spectrum. [...] The second foundational rule, which is borne out by the results of [...] psychophysical experiments [...], asserts that speech is a composite signal, hierarchically organized so that simpler patterns at one level are combined in a well-defined manner to form more complex patterns at the succeeding level. [...] The structures at each level

S. Markovski, M. Gusev (Editors): ICT Innovations 2012, Web Proceedings, ISSN 1857-7288

of the hierarchy serve to constrain the ways in which the individual patterns associated with that level can be combined."

The growing success of statistical approaches in speech technology during the 1990-th resulted in a convergence of speech recognition and speech synthesis which had developed hitherto in separate ways. This was mainly due to the necessity of large databases or knowledge sources in both branches. This development had been predicted, e. g., in a classical textbook [2]: "Advanced systems both for synthesis and for recognition need the same speech knowledge, and there is considerable advantage for the two applications to be studied together. [. . .] I predict that the most significant progress in the more advanced forms of speech synthesis and recognition will in future come from research teams with a strong interest in both problems." The development of the so-called HMM synthesis was the most important result of this generalized sight.

Both approaches, which have been addressed by the citations above, the hierarchical as well as the analysis-synthesis concept, have been united in our UASR (Unified Approach for Speech Synthesis and Recognition) which was initiated as a long-term project in the year 2000 [3]. The following is an extended abstract of an overview of our work through the last decade, which is mainly intended to provide a commented collection of references to the original work.

2 The UASR System

The implementation of the aforementioned ideas resulted in a system which was composed from three components, each of which offering a hierarchical structure:

- the analysis (or recognition) branch with information flowing bottom-up,
- the synthesis branch with information flowing top-down,
- a set of databases, one for each hierarchy level, which are accessed by the analysis as well as the synthesis component of the respective level.

The first implementation included the feature, phonetic, lexical, and syntactic levels. It was described in [4], [5], and [6].

The algorithmic design of the processing elements at the "symbolic" hierarchy levels proved to be a challenge. UASR is a big software systems, and it was necessary to look for a method to design the components in an uniform and economic way. The application of finite state transducers which were introduced in the speech technology some years before by M Mohri [7], turned out as the most efficient way. After a general redesign, UASR is now implemented in FST technology completely [8].

We want to stress that UASR is not only a powerful experimental platform. Additionally, we implemented an embedded version in cooperation with the Fraunhofer society for speech control applications [9] [10]. The recent version works basing on an FPGA, but a special processor will be the final goal.

3 Useful Non-speech Applications

There are many applications of pattern recognition algorithms to acoustical and other signals from the real world. In the past, numerical pattern recognition methods proved to be sufficient for solving those tasks. Structural pattern recognition was essentially restricted to applications with speech signals. This situation changed around the year 2000 when the performance of classical approaches was not satisfactory for new tasks in non-destructive testing and other branches. Using UASR, we were able to demonstrate in may cases that structural methods are able to introduce much progress to different classes of signals. We mention some examples as follows:

- **Non-destructive testing.** This was the dominant application area, supported by a cooperation with the Fraunhofer society. The results are summarized in [11], [12], and the thesis [13].
- **State monitoring of machines.** This tasks is similar to the aforementioned, but does not require a sharp decision. Instead, the degree of membership to the one and only defined class has to be calculated. Different projects which were performed are summarized in [14].
- **Biological signals.** From different examples, we want to mention here the measurement of the blood pressure under real conditions (i. e., a moving test person), because this was a real challenging task [15].
- **Musical instruments.** This is a specific version of non-destructive testing which we are dealing with since many years in cooperation with the industry [16]. Among other results, we have shown that it is possible to identify specific instruments by structural methods [17] [18].
- **Musical signals.** The investigation of musical signals is useful for a number of applications in music retrieving, measuring similarities between musical works, etc. [19]. Additionally, speech research may benefit from this because basic findings about rhythm may be transferred to speech signals, which is an very actual problem in prosody research [20].

A more complete overview is given in [21] and [8].

4 Towards Cognitive Systems

Some years ago, S.Haykin coined the term "cognitive dynamic systems" for systems which show a purposeful behavior like human beings [22]. They are able to develop an internal model of their environment and, basing on this, to influence their environment actively. Surprisingly, elaborated applications of this theory are existing not only in the traditional fields of artificial intelligence (including speech technology), but mainly in "cognitive signal processing systems" like the cognitive radar and the cognitive radio [23].

Systems like cognitive radio are not really hierarchical (although the application of the well-known OSI model introduces a hierarchy). Therefore there is

much similarity between non-hierarchic cognitive systems and the classical theory of automatic control. Haykin points out, however, that cognitive systems in biology are organized in a hierarchical way, demonstrating it with a prominent example from [24]. The block diagram of an hierarchical perception-action system of this kind is very similar to that of UASR.

Special attention has to be directed to the interaction between the levels of the hierarchy. Although the information flow of the analysis branch is bottom-up, and of the synthesis branch top-down, the performance will be improved by bidirectional interaction. The cortical algorithm which was originally developed for modeling the visual cortex [25], seems to be very promising for the implementation in UASR [26] [27].

5 Conclusion

Basing on the ideas discussed above, UASR offers some potential directions for further development, as follows:

- The success of applying UASR to many non-speech problems resulted in the idea to generalize the hierarchical analysis-synthesis systems towards an intelligent system for hierarchical signal processing. It is discussed in detail in [8] and as an overview in [28].
- The step from the existing UASR towards a real "cognitive" system will be done by adding a cognitive backend at the top of the speech-related hierarchy. This could be performed in various ways, e. g. by adding a translation component like in the former Verbmobil system [29]. We decided to follow a solution which was proposed in [30], aiming to a dialogue controller. For a flowgraph of the extended UASR, cf. [8] or [31]. It is a challenge to demonstrate that the uniform application of FST technology (maybe generalized to a theory of Petri net transducers) can be extended to the semantic components also [32].
- The principle of "analysis by synthesis" (AbS), which plays an essential role in the development of speech technology, was successfully applied in introducing new approaches for the parametric speech synthesis (so-called HMM synthesis [33] [34]). We expect some progress in the personalization of speech technology by applying this principle to deeper studies of human prosody [35].

Acknowledgments. This work was partially funded by

- the Deutsche Forschungsgemeinschaft (DFG) under grants Ho 1674/3, Ho 1674/7, and Ho1684/8,
- the German Federal Ministry of Education and Research (BMBF), and
- the Arbeitsgemeinschaft industrieller Forschungsvereinigungen "Otto von Guericke" (AiF).

References

1. S. E. Levinson: Structural methods in automatic speech recognition. Proceedings of the IEEE 73 (1985) 11, 1625–1650.
2. J. N. Holmes: Speech Synthesis and Recognition. London: Van Norstrand Reinhold 1988.
3. M. Eichner; M. Wolff; R. Hoffmann: A unified approach for speech synthesis and speech recognition using stochastic Markov graphs. Proc. ICSLP, Beijing 2000, vol. 1, 701–704.
4. S. Werner, M. Eichner, M. Wolff, R. Hoffmann: Towards spontaneous speech synthesis – Utilizing language model information in TTS. IEEE Trans. on Speech and Audio Processing 12 (2004) 4, 436–445.
5. M. Eichner: Sprachsynthese und Spracherkennung mit gemeinsamen Datenbasen – Akustische Analyse und Modellierung. Dresden: TUDpress 2007 (Studientexte zur Sprachkommunikation, vol. 43).
6. S. Werner: Sprachsynthese und Spracherkennung mit gemeinsamen Datenbasen – Sprachmodell und Aussprachemodellierung. Dresden: TUDpress 2008 (Studientexte zur Sprachkommunikation, vol. 48).
7. M. Mohri: Finite-state transducers in language and speech processing. Computational Linguistics 23 (1997) 2, 269–311.
8. M. Wolff: Akustische Mustererkennung. Dresden: TUDpress 2011 (Studientexte zur Sprachkommunikation, vol. 57).
9. G. Strecha, M. Wolff, F. Duckhorn, S. Wittenberg, C. Tschöpe: The HMM synthesis algorithm of an embedded unified speech recognizer and synthesizer. Proc. Interspeech, Brighton 2009, 1763–1766.
10. G. Strecha, M. Wolff: Speech synthesis using HMM based diphone inventory encoding for low-resource devices. Proc. IEEE ICASSP, Prague 2011, 5380–5383.
11. C. Tschöpe, D. Hentschel, M. Wolff, M. Eichner, R. Hoffmann: Classification of non-speech acoustic signals using structure models. Proc. IEEE ICASSP, Montreal 2004, vol. 5, 653–656.
12. C. Tschöpe, M. Wolff: Statistical classifiers for structural health monitoring. IEEE Sensors Journal 9 (2009) 11, 1567–1576.
13. C. Tschöpe: Akustische zerstrungsfreie Prfung mit Hidden-Markov-Modellen. Dresden: TUDpress 2012 (Studientexte zur Sprachkommunikation, vol. 60).
14. S. Wittenberg: Statistische Ein-Klassen-Signalbewertung mit akustischen Datenbasen selbstbeschreibender Daten. Dresden: TUDpress 2012 (Studientexte zur Sprachkommunikation, vol. 63).
15. M. Wolff, U. Kordon, H. Hussein, M. Eichner, C. Tschöpe; R. Hoffmann: Auscultatory blood pressure measurement using HMMs. Proc. IEEE ICASSP, Honolulu 2007, vol. 1, 405–408.
16. G. Ziegenhals: Subjektive und objektive Beurteilung von Musikinstrumenten. Dresden: TUDpress 2010 (Studientexte zur Sprachkommunikation, vol. 51).
17. M. Eichner, M. Wolff, R. Hoffmann: Instrument classification using hidden Markov models. Int. Conf. on Music Information Retrieval (ISMIR), Victoria 2006, 349–350.
18. M. Eichner, M. Wolff, R. Hoffmann: An HMM based investigation of differences between musical instrumentsof the same type. Proc. Int. Congress of Acoustics (ICA), Madrid 2007, 5 pp.
19. S. Hübler, R. Hoffmann: Evaluation of onset detection algorithms in popular polyphonic music on a large scale database. Proc. of the 130th Audio Engineering Society Convention, London 2011, 1265–1270.

S. Markovski, M. Gusev (Editors): ICT Innovations 2012, Web Proceedings, ISSN 1857-7288

20. S. Hübler, R. Hoffmann: Comparing music and speech with a closer look on automatic music information retrieval. In: A. Esposito et al. (eds.): Towards Autonomous, Adaptive, and Context-aware Multimodal Interfaces. Berlin etc.: Springer 2011 (Lecture Notes in Computer Science, vol. 6456), 376–386.
21. R. Hoffmann, M. Eichner, M. Wolff: Analysis of verbal and nonverbal acoustic signals with the Dresden UASR system. In: A. Esposito et al. (eds.): Verbal and Nonverbal Communication Behaviours. Berlin etc.: Springer 2007 (Lecture Notes in Artificial Intelligence, vol. 4775), 200–218.
22. S. Haykin: Cognitive dynamic systems. Proc. IEEE ICASSP, Honolulu 2007, vol. 4, 1369–1372.
23. S. Haykin: Cognitive Dynamic Systems. Perception-action Cycle, Radar and Radio. Cambridge University Press 2012.
24. J. M. Fuster: Cortex and Mind – Unifying Cognition. New York: Oxford University Press 2003.
25. T. S. Lee, D. Mumford: Hierarchical Bayesian inference in the visual cortex. JOSA 20 (2003) No. 7.
26. R. Römer; T. Herbig: Konzeptionelle Beschreibung des cortikalen Algorithmus und seine Anwendung in der automatischen Sprachverarbeitung. In: R. Hoffmann (ed.): Elektronische Sprachsignalverarbeitung, Dresden 2009, Bd. 1 (Studientexte zur Sprachkommunikation, vol. 53), 33–40.
27. R. Römer: Untersuchungen zum cortikalen Algorithmus unter Verwendung von bidirektionalen HMMs. In: M. Wolff (ed.): Elektronische Sprachsignalverarbeitung, Cottbus 2012 (Studientexte zur Sprachkommunikation, vol. 64), 252–261.
28. M. Wolff, R. Hoffmann: An approach to intelligent signal processing. In: A. Esposito et al. (eds.): Cognitive Behavioural Systems. COST 2102 International Training School, Dresden 2011, Revised Selected Papers. Berlin etc.: Springer 2012 (Lecture Notes in Computer Science, vol. 7403).
29. W. Wahlster: Verbmobil – Foundations of Speech-to-Speech Translation. Berlin etc.: Springer 2000.
30. M. Huber, C. Kölbl, R. Lorenz, R. Römer, G. Wirsching: Semantische Dialogmodellierung mit gewichteten Merkmal-Werte-Relationen. In: R. Hoffmann (ed.): Elektronische Sprachsignalverarbeitung, Dresden 2009, vol. 1 (Studientexte zur Sprachkommunikation, vol. 53), 25–32.
31. M. Wolff, R. Römer, R. Hoffmann: Hierarchische kognitive dynamische Systeme zur Sprach- und Signalverarbeitung. In: M. Wolff (ed.): Elektronische Sprachsignalverarbeitung, Cottbus 2012 (Studientexte zur Sprachkommunikation, vol. 64), 159–178.
32. M. Huber, R. Lorenz: Petri net transducers in semantic dialogue modelling. In: M. Wolff (ed.): Elektronische Sprachsignalverarbeitung, Cottbus 2012 (Studientexte zur Sprachkommunikation, vol. 64), 286–297.
33. A. Falaschi, M. Giustiniani, M. Verola: A hidden Markov model approach to speech synthesis. Proc. Eurospeech, Paris 1989, 187–190.
34. K. Tokuda et al.: Speech parameter generation algorithms for HMM-based speech synthesis. Proc. IEEE ICASSP, Istanbul 2000, 1315–1318.
35. R. Hoffmann: Analysis-by-synthesis in prosody research. Keynote, Proc. 6th International Conference on Speech Prosody, Shanghai 2012, 1–6.

Phonetic and Prosodic Aspects in the Cross-lingual Pronunciation Tutoring

Oliver Jokisch

Chair for System Theory and Speech Technology
Dresden University of Technology, 01062 Dresden, Germany
oliver.jokisch@tu-dresden.de
http://www.ias.et.tu-dresden.de

Abstract. Computer-assisted pronunciation tutoring (CAPT) methods have been well-established in research and education. Common system approaches include the phonetic quality assessment, highlight problematic sections in the speech signal and usually rely on automatic speech recognition (ASR) regarding the target language L2. The contribution deals with the audiovisual CAPT system AzAR. An extensive feedback mechanism and several speech databases for Slavonic learners of German were developed. Currently, AzAR is adapted for Chinese Learners of German and for learners of the Basque language. This extended abstract summarizes some experiences of the multilingual experiments and system tests – focusing on phonetic and prosodic assessment aspects.

Keywords: Pronunciation tutoring (CAPT), Slavonic languages, Mandarin Chinese, German, Basque

1 Slavonic-German Transfers in Pronunciation Tutoring

The speech data collection and the adaptation of automatic speech recognition (ASR) algorithms are essential tasks for the development of pronunciation tutoring (CAPT) systems. The system "Automat for Accent Reduction (AzAR)" [1] was originally designed for the pronunciation tutoring of learners with a native language L1 from the Slavonic language group and for the target language L2 German. Within the cooperation project Euronounce [2], the concept was extended to the Slavonic target languages Polish, Slovak, Czech and Russian. The Euronounce database includes special lessons for phonetic peculiarities but also sentences to evaluate the prosodic aspects. It contains 130 speakers and about 200 hours of speech. In further steps, the Euronounce concept was also tested for Mandarin Chinese learners of German (as L2 and L3) and learners of the Basque language with different mother tongues. Preliminary results with focus on phonetic and prosodic aspects are summarized in the following sections.

2 Cross-lingual effects in Chinese learners of German

In this section, some experiences from the AzAR adaptation for Chinese Learners of German [3–5] are reported.

S. Markovski, M. Gusev (Editors): ICT Innovations 2012, Web Proceedings, ISSN 1857-7288

2.1 Analysis of phonetic and prosodic deviations

With acoustic and perceptual investigations, it was found that various segmental and intonational deviations contribute to difficulties in oral communication:

- Inaccurate production of those German vowels and consonants which are nonexistent in Chinese,
- Incorrect placement of tonal categories and wrong phonetic realization of a phonological category.

Chinese learners usually employ different strategies, such as epenthesis, deletion and modification to deal with unfamiliar sounds because Chinese syllables usually end with vowels, and the learners usually add a schwa /@/ after the consonant final. In a comparative analysis of German produced by Russian and Chinese learners, it was obvious that epenthesis occurs more frequently among Chinese learners than Russian learners [4].

With the visual-audio feedback information and after many times of trial and error, the learners became conscious of their pronunciation mistakes, and could make correspondent corrections. Chinese is a tone language, Chinese speakers thus raise or lower their pitches to express different lexical meanings instead of different linguistic purposes in intonation languages like German and English. The f0 deviations in Chinese speakers of German have been investigated in [3].

The visualization of intonation curve is proved to be particularly effective in the acquisition of L2 intonation. Automatic pitch tracking algorithm, however, usually displays many small pitch changes that make the learners confused about the sentence intonation, and moreover some small changes are linguistically unimportant.

2.2 Preliminary results

It proves possible for Chinese learners of German language to imitate standard pronunciations successfully. However, a faithful imitation of isolated words or sentences with visual aids can not guarantee a good pronunciation in ordinary speech. The articulatory constraints will still dominate for many students in normal speech without any audiovisual aids. The tutoring system should also guide the learners step by step from a successful imitation to an accurate production in free continuous speech. Therefore the next research interest will be focused on continuous and normal speech, which is the ultimate goal of language teaching.

3 Basque language as example of a new learning target

In the following section, the peculiarities of the target language L2 Basque are discussed – basing on the contributions [6, 7].

3.1 Language peculiarities

Basque is an isolated language which does not belong to the Indo-European language group, as one would expect from its geographical location [8]. It has two major languages as neighbors, Spanish and French, and the influence of these languages on Basque is noticeable, especially on people who are learning Basque as L2. Moreover, the fact that the Academy of the Basque Language has not made any decision about standard Basque intonation or prosody yet, may confuse new students regarding their reference in phrase accent and intonation.

Considering the importance of prosody acquisition in Basque, a suprasegmental analysis part has been added to database and CAPT system which will be useful for future developments. In the suprasegmental part, word level and sentence level intonation have been taken into consideration – concerning the segmental part, some phonetic and some phonological features:

- Phonetic features: phonemes that do not exist in the neighbor languages, as the /ts‘/ [9], the differentiation between the six sibilants of Basque and the vocalic system.
- Phonological features: the palatalization process in the context /ilV/ and /inV/ (V refers to a vowel), and the unvoicing process caused by the negative particle ez in the first phoneme of the next word.

A baseline Basque curriculum was designed, and a 16 kHz/16 bit PCM speech database was recorded from a Basque native speaker. The segmental part of the database consists of 125 sentences and 60 word pairs. The suprasegmental part contains 20 isolated words and 50 sentences.

3.2 Phonetic assessment

The ASR baseline system for the segmental part involves an HMM-based phoneme verification system in forced alignment mode, using GOP (Goodness Of Pronunciation) score as confidence measure. It was trained using a Basque studio database which contains recordings from native and non-native speakers, as well as dialectal and standard Basque data.

Measuring the general performance of the ASR system for the reference voice, it was concluded that the algorithm was not able to discern properly between sibilants. The explanation of this fact is that there are Basque speakers of different skill levels in the training database and so wrongly pronounced utterances were used to train the HMMs of phonemes that do not exist in Spanish. Later, a new data set was trained using only the native speakers, so that more robust decisions were obtained.

3.3 Prosodic assessment

For the suprasegmental part, a simple but efficient approach for the prosodic assessment was tested by directly calculating the RMSE between the realized and the reference f0 curve. For that purpose, a test corpus was recorded with

speakers whose mother tongue is different. They were asked to read some Basque sentences where prosodic segments were indicated, and the meaning of each part was translated to them – to force the application of their native intonation to the Basque sentence.

References

1. Jokisch, O., Koloska, U., Hirschfeld, D., Hoffmann, R.: Pronunciation learning and foreign accent reduction by an audiovisual feedback system. In Proc. 1st Intern. Conf. on Affective Computing and Intelligent Interaction (ACII), Beijing (China), pages 419-425, October 2005. Springer LNCS-3784.
2. Jokisch, O., Jaeckel, R., Rusko, M., Demenko, G., Cylwik, N., Ronzhin, A., Hirschfeld, D., Koloska, U., Hanisch, L., Hoffmann, R.: The EURONOUNCE project - An intelligent language tutoring system with multimodal feedback functions: Roadmap and specification. Proc. Konferenz Elektronische Sprachsignalverarbeitung (ESSV), pages 116-123, September 8-10, 2008. Frankfurt/M.
3. Ding, H., Jokisch, O., Hoffmann, R.: F0 analysis of Chinese accented German speech. Proc. 5th Intern. Symposium on Chinese Spoken Language Processing (ISCSLP), p. 49-56, Dec. 2006. Singapore.
4. Hilbert, A., Mixdorff, H., Ding, H., Pfitzinger, H., Jokisch, O.: Prosodic analysis of German produced by Russian and Chinese learners. Proc. 5th Intern. Conf. on Speech Prosody, May 11-14, 2010. Chicago.
5. Ding, H., Mixdorff, H., Jokisch, O.: Pronunciation of German syllable codas of Mandarin Chinese speakers. Proc. Konferenz Elektronische Sprachsignalverarbeitung (ESSV), 281-287, Sept. 2010. Berlin.
6. Odriozola, I., Jokisch, O., Hernaez, I., Hoffmann, R.: A Pronunciation Tutoring System for Basque - First Development Steps. Proc. Konferenz Elektronische Sprachsignalverarbeitung (ESSV), Aug. 2012. Cottbus.
7. Odriozola, I., Navas, E., Hernaez, I., Sainz, I., Saratxaga, I., Sanchez, J., Erro, D.: Using an ASR database to design a pronunciation evaluation system in Basque. In Proc. 8th Intern. Conf. on Language Resources and Evaluation (LREC), Istanbul, May 2012.
8. Hualde, J. I.: Basque Phonology, Taylor & Francis, 1991 (ISBN 9780415056557).
9. University of Bilbao, Aholab: SAMPA computer readable phonetic alphabet of Basque, retrieved 25 April 2012 from `http://aholab.ehu.es/sampa_basque.htm`

Automatic Music Classification into Genres

Gjorgji Madjarov[1], Goran Pesanski[2], Daniel Spasovski[2], and Dejan Gjorgjevikj[1]

[1] Faculty of Computer Science and Engineering
[2] Faculty of Electrical Engineering and Information Technologies
{gjorgji.madjarov,dejan.gjorgjevikj}@finki.ukim.mk
pesanski_goran@yahoo.com,sdaniel506@yahoo.com

Abstract. Musical genres are categorical labels created by humans to characterize pieces of music. Although music genres are inexact and can often be quite arbitrary and controversial, it is believed that certain song characteristics like instrumentation, rhythmic structure, and harmonic content of the music are related to the genre. In this paper, the task of automatic music genre classification is explored. Multiple features based on timbral texture, rhythmic content and pitch content are extracted from a single music piece and used to train different classifiers for genre prediction. The experiments were performed using features extracted from one or two 30 second segments from each song. For the classification, two different architectures flat and hierarchical classification and three different classifiers (kNN, MLP and SVM) were tried. The experiments were performed on the full feature set (316 features) and on a PCA reduced feature set. The testing speed of the classifiers was also measured.The experiments carried out on a large dataset containing more than 1700 music samples from ten different music genres have shown accuracy of 69.1% for the flat classification architecture (utilizing one against all SVM based classifiers). The accuracy obtained using the hierarchical classification architecture was slightly lower 68.8%, but four times faster than the flat architecture.

Keywords: music, genre, classification, flat, hierarchical

1 Introduction

A music genre is a conventional category that identifies pieces of music as parts of a set of rules and conventions.Although the artistic nature of music means that these classifications are often arbitrary and controversial, and some genres may overlap, this human made labels help people to better organize their music collections, based on their individual music perception and cognition, choose the music radio station to listen to, and plays important role in electronic music distribution.Music can be divided into different genres in several ways. There are many music genres,such as classical, rock, pop, disco, etc. The music genre, because of its flexibility, is constantly exposed to changes and as result many fused genres are produced that creates a genre classification hierarchy. It is determined that many of the elements that belong to an audio

S. Markovski, M. Gusev (Editors): ICT Innovations 2012, Web Proceedings, ISSN 1857-7288

signal can be used as features that are needed for music classification. These features include the spectral characteristics of the audio signal, timbre, pitch, tempo, energy distribution, rhythm or other content[1][2][3][4]. The majority of the research projects are focused exactly on providing better methods for feature extraction[4][5][6]. Though the concept of musical genre might not be well defined, recent approaches that use audio feature extraction combined with machine learning techniques have achieved promising results[1][2][6][7]. Nowadays, the digital music databases, mostly located on the Web, become more popular both for professional and private purposes. The need for automatic organization and classification increases every day. Manual or even semi-automated annotation of each music file is impractical, expensive and time consuming approach. This problem inspires the computer scientist and researches, together with the music workers, to work on a solution. The automatic music genre classification is a big challenge for many scientists and researchers. Different classification techniques for automatic music genre classification, such as Support Vector Machines (SVM)[1][2][7][8], Artificial Neural Networks (ANN)[6][9], Hidden Markov Models (HMMs)[10] and Gaussian Mixture Models (GMMs)[11] have been used by different researches. Marsyas (Music Analysis, Retrieval and Synthesis for Audio Signals)[12][13] as a framework for audio processing and speech analysis with specific emphasis on music information retrieval applications was used for the task of feature extraction in our research.

In this paper, we address the problem of automatic music genre classification in ten different genres. More specifically, three different sets of features represented by timbral texture, rhythmic content and pitch content are used for classifying ten different music genres.Two classification architectures (flat and hierarchical) are experimentally evaluated,using three different types of base classifiers: SVM [8], Multilayer Perceptron (MLP) [14] and k - Nearest Neighbours (kNN) [15].

Section 2 presents the feature extraction process and the features used in our experiments. Section 3 describes the datasets used in the experiments and the experimental setup. The experimental results are presented and discussed in Section 4. Finally, the conclusions are given in Section 5.

2 Feature Extraction and Datasets

2.1 Feature Extraction

The main focus of our research is the classification methods and techniques for music genre classification and not the feature extraction. As mentioned in the previous section, the feature extraction was performed using the MARSYAS [3] tool. Two different types of features were extracted. The first set consists of 31 timbral texture features extracted from each music sample including: time-domain Zero-Crossings (1), SpectralCentroid (1), Rolloff (1), Flux (1), Chroma (14) and Mel-FrequencyCepstralCoefficients MFCC (13) over a texture window

[3] http://marsyas.info/

of 1 sec. The second set consists of 48 Spectral features: SpectralFlatnessMeasure (24) and SpectralCrestFactor (24). Each feature extracted by MARSYAS is represented by 4 separate values,so the actual length of the feature vector is four times the number of features. Each music sample is represented by 79 features each represented by 4 values.

2.2 Datasets

In the paper we concentrate on analyzing the performance of different classification techniques for the problem of automatic music genre classification. Our goal was to classify music files in wav format according to their genre. We used the same ten music genres, used by Tzanetakiset al. [1][2]. The selected genres include: Blues, Classical, Country, Disco, Hip-Hop, Jazz, Metal, Pop, Reggae and Rock. 2760 music songs (23 hours of audio data) were collected and used in the experiments. All songs were stored as 22.5 kHz mono, 352 kbps wave files. For splitting the song wave files before the feature extraction WavSplit 1.2.1 for Linux [16] was used.

Three different datasets were used in this analysis. All three datasets are composed of 1000 instances for training and 1760 instances for testing. The exact distribution of the training and testing instances regarding the genres are shown in Table 1.

Table 1. The distribution of the training and testing instances regarding the genres

	Blues	Classical	Country	Disco	Hip-Hop	Jazz	Metal	Pop	Reggae	Rock
Training	100	100	100	100	100	100	100	100	100	100
Testing	167	94	101	159	240	139	150	179	131	400
Total	267	194	201	259	340	239	250	279	231	400

Each instance from the first dataset is represented by 30 second segment (between the 30th and 60th second) of the actual music song,referred as music sample. Each music sample was described by timbral texture features (124 features)and spectral features (192 features) mentioned in the previous section, which results with 316 features in total per sample.This data set is denoted as one sample features data set. Unlike the first dataset, in the second dataset,each song was represented by two music samples (both 30 seconds), the first beginning at the 30% of the duration of the song and the second at 60% of songs duration. Features (timbral texture and spectral) are extracted separately from each sample of the song and then the resulting features are concatenated in a single feature vector representing that instance. We denote the second data set as two sample features data set. In order to reduce the length of the feature vector obtained by concatenating the features from the two parts of the song as discussed before, in the third dataset we performed the PCA (Principal Component Analysis) feature selection method [17][18]. In this manner, the length of an instance in the third dataset was reduced from 612 features to 151 features.

3 Experimental Setup and Results

3.1 Experimental Setup

The comparison of the classification methods (SVMs Support Vector Machines, MLP Multilayer Perceptron and kNN k Nearest Neighbours) was performed using their implementations in Weka [19]. For training the SVMs, we used the SMO implementation. In particular, we used SVMs with a radial basis kernel. The kernel parameter gamma and the penalty C, for each combination of dataset and method, were determined by 10-fold cross validation using only the training set. The values 2^{-15}, 2^{-13}, , 2^{1}, 2^{3} were considered for gamma and 2^{-5}, 2^{-3}, ,2^{13}, 2^{15} for the penalty C. The number of neighbours in the kNN method for each dataset was determined from the values 1 to 9 with step 2. The Neural Networks are represented by MLP with 25 neurons in the hidden layer and value for the validation threshold of 10. After determining the best parameters values for each method on every dataset by 10-fold cross validation, the classifiers were trained using all available training examples and were evaluated by recognizing all test examples from the corresponding dataset. Two different architectures (flat and hierarchical) are considered and explored in our work. The following subsections include the brief description and the results obtained from each classifying architectures.

3.2 Experimental Results

Flat Classification The flat classification addresses the problems where the predefined classes are separately treated and there is no structure defining the relationships among them (or that structure is not considered even if it exists). According to this,we do not take into account the possible relationships between the classes for the purpose of the flat classification.

The 10-genre classification is performed by classifying the music samples in their appropriate genre, using the classifiers mentioned previously. One performance evaluation measure (accuracy)and the testing time measured in seconds were used to estimate the performance of the different classifiers.

Table 2 shows the results of the three classifiers on the first and the second dataset. One instance from the first dataset was represented by only one music sample per song, described by 316 features, while an instance from the second dataset was represented by two music samples per song, described by 316 features each.The first column of the table describes the classification genres. The other columns show the accuracy of the classifiers per genre. The first group of three columns shows the performance obtained on the first dataset, while the second group of three columns shows the performance obtained on the second dataset. The last two rows present the overall accuracy of the classifiers per dataset and the testing times accordingly.The best prediction results are achieved by SVM for both datasets.The MLP classifier showed similar performance results, but its testing time is about two times longer than the SVM classifier for the first dataset and 1.8 times longer for the second dataset. For some particular genres

MLP is even more accurate than the SVM. This is the case for the half of the genres for the first dataset and 4 out of 10 genres for the second dataset. Blues is the genre that decreases the overall accuracy of the MLP. The KNN classifier, compared to the SVM and MLP, provides lower accuracy for almost every genre.

Table 2. Classification accuracy comparison between different classifiers (accuracy in %)

	One sample features			Two sample features		
	SVM	KNN	MLP	SVM	KNN	MLP
Blues	50.90	33.53	36.53	43.11	32.34	35.93
Classical	91.49	80.85	88.30	91.49	82.98	86.17
Country	64.36	67.33	70.30	70.30	68.32	70.30
Disco	64.15	62.89	60.38	62.89	64.78	64.78
Hip-Hop	84.58	79.58	84.58	89.58	81.25	85.42
Jazz	70.50	48.92	72.66	80.58	49.64	71.22
Metal	82.00	75.33	84.67	87.33	82.00	88.67
Pop	32.40	22.35	36.31	32.40	24.02	29.05
Reggae	67.18	66.41	70.23	63.36	68.70	70.99
Rock	68.50	44.50	66.50	72.00	40.00	70.75
Accuracy(%)	67.16	55.51	66.19	69.09	55.91	67.05
Time(s)	34	29	63	79	62	151

Table 3 shows more detailed information about the musical genre classifier performance in the form of a confusion matrix. In a confusion matrix, the columns correspond to the actual genre and the rows to the predicted genre.The relative distribution of the values in the confusion matrix for the two other classifiers is very similar. These matrices show that Classical music instances are classified with the highest accuracy. On the other hand, Blues, Pop and Rock happen to be the genres that are most often confused with the others. For example, Blues is often mixed with Country, Rock and Pop music. Pop is mixed with Disco, Rock and Country music and etc.

Table 4 shows the confusion matrix for the second dataset. It can be noticed that the number of correctly classified instances increased, especially in Jazz, Rock and Hip-Hop genres. This led to improving the overall percentage of correctly classified instances using all classifiers, SVM being the best with classification accuracy of 69.1%.

For the third dataset (where PCA dimensionality reduction was performed), only the performance of the SVM classifier was measured. As we expected, the testing time of the classifier was shortened, while the accuracy slightly decreased. In particular, classifying the whole test set required 25 seconds, and the obtained accuracy was 65.75%.

Hierarchical Approach to Music Genre Classification Hierarchical classification refers to assigning samples to a suitable class from a hierarchical class

Table 3. Music genre confusion matrix using SVM classifier on test dataset

	Blues	Classical	Country	Disco	Hip-Hop	Jazz	Metal	Pop	Reggae	Rock
Blues	**85**	2	18	13	0	9	2	15	5	17
Classical	5	**86**	1	1	0	1	0	0	0	0
Country	25	1	**65**	1	0	2	0	1	2	4
Disco	4	0	9	**102**	5	3	0	27	0	9
Hip-Hop	1	0	4	12	**203**	3	1	15	0	1
Jazz	11	9	0	5	1	**98**	12	0	0	3
Metal	3	5	1	1	0	3	**123**	3	0	11
Pop	20	2	23	34	6	3	5	**58**	1	27
Reggae	11	0	6	14	1	3	1	2	**88**	5
Rock	12	9	23	49	1	3	19	10	1	**274**

Table 4. Music genre confusion matrix using concatenated features from two samples

	Blues	Classical	Country	Disco	Hip-Hop	Jazz	Metal	Pop	Reggae	Rock
Blues	**72**	0	23	11	0	17	3	16	2	22
Classical	1	**86**	1	2	0	2	0	0	0	2
Country	16	1	**71**	4	1	4	0	1	1	2
Disco	6	0	7	**100**	8	4	2	25	0	7
Hip-Hop	1	0	0	6	**215**	4	0	11	0	3
Jazz	11	1	1	2	1	**112**	9	0	1	1
Metal	3	3	1	1	0	2	**131**	1	0	8
Pop	23	3	34	27	9	3	2	**58**	1	19
Reggae	8	0	2	21	5	3	0	3	**83**	6
Rock	18	2	29	32	0	6	19	7	0	**288**

space [8]. By utilizing the previously defined hierarchical architecture,the classification problem can be decomposed into a smaller set of problems. In this approach the classification is accomplished with the cooperation of classifiers built at each level of the tree.One of the obvious problems with the top-down approach is that a misclassification at a parent class may force a sample to be misrouted before it can be classified into the correct child classes.

In many classification experiments, the hierarchical approach can lead to better results in the multi-class classification process. Fig. 1 shows the 2-level hierarchy that we considered in the experiments. Each test instance is passed through the hierarchical architecture of classifiers resulting in an instance classified in one of the 10 music genres.

The first level of the hierarchy consists of 4 nodes illustrating the most distinctive groups of music genres, based on the confusion matrices obtained from the flat classification. Classical music, as the most distinctive genre in flat classification, represents one node in the hierarchy. The other three nodes contain groups of genres that are similar to each other and often mutually misclassified by the classifiers. Hierarchical classification architectures of the three different classifiers discussed previously were trained and applied to the music genre classification problem.

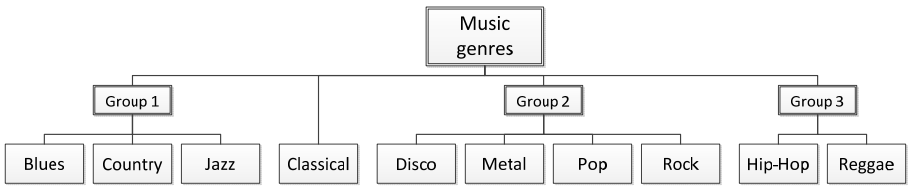


Fig. 1. 2-level music genre hierarchy

For the hierarchical architecture we present only the results of the best classifier. Table 5 shows the results at the 1^{st} and 2^{nd} level obtained by the hierarchy of SVM classifiers. It can be seen that the overall accuracy is slightly lower (only 0.2%) compared to the flat classification, but the testing time is significantly improved (more than four times). It can also be noted that for particular genres as Classical, Rock and Pop the accuracy is improved, especially for the experiments where concatenated features are used.

Table 5. Results from the hierarchy provided by the SVM classifier (accuracy in %)

	1st level		2nd level	
	One sample features	Two sample features	One sample features	Two sample features
Classical	87.23	92.55	87.23	92.55
Blues	76.17	75.18	43.11	41.32
Country			60.40	63.37
Jazz			64.75	74.10
Disco	86.26	87.73	60.38	61.64
Metal			83.33	88.67
Pop			31.84	35.75
Rock			71.00	73.50
Hip-Hop	81.67	82.21	86.25	88.33
Reggae			70.23	66.41
Accuracy(%)	83.01	83.92	66.25	68.81
Time(s)	7	15	9	19

4 Conclusions and Future Work

In this paper, we address the problem of automatic music genre classification.Three different sets of features represented by timbral texture, rhythmic content and pitch content were used for classifying ten different music genres. Two classification architectures (flat and hierarchical) were evaluated experimentally, using three different types of base classifiers: SVM, Multilayer Perceptron and k Nearest Neighbours.

SVM one-versus-all showed the best predictive performance comparing to the kNN and MLP classifiers. The SVM classifier showed predictive accuracy of 67.16% for the case where single 30 second music segment per song was used to build the model. The performance increased for additional 2% when features extracted from two 30 second segments from each song were used, but this also slowed down the prediction by more than 2 times.

On the other hand, the hierarchical approach, that was justified based on the similarities between the musical genres associated with the rhythm, harmony and pitch, showed significant improvements in the testing time comparing to the flat classification approach(more than four times), while showing only slightly lower (0.2%) predictive accuracy.

Future work will involve further analysis of the feature space, genre group dependent selective extraction and combination of different types of features on the second level of the classification hierarchy,examination of alternative classification schemes, and incorporation of more audio classes.We will also try to transform this problem into multi-label one and solve it by commonly used multi-label classification techniques.

References

1. George Tzanetakis, P.C.: Music genre classification of audio signals. IEEE Transactions on speech and audio processing **10** (2002) 293–302
2. George Tzanetakis, Georg Essl, P.C.: Automatic music genre classification of audio signals. In: 2nd Annual International Symposium on Music Information Retrieval. (2001)
3. Tao Li, Mitsunori Ogihara, Q.L.: A comparative study on content-based music genre classification. In: Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval. (2003) 282 289
4. Thomas Lidy, A.R.: Evaluation of feature extractors and psycho-acoustic transformations for music genre classification. In: Proceedings of the 6th International Conference on Music Information Retrieval. (2005) 34–41
5. Michael Scott Cuthbert, Christopher Ariza, L.F.: Feature extraction and machine learning on symbolic music using the music21 toolkit. In: Proceedings of the 12th International Society for Music Information Retrieval Conference. (2011) 387–392
6. Neumayer, R.: Musical genre classification using a multi-layer perceptron. In: Proceedings of the 5th Workshop on Data Analysis (WDA'04), Elfa Academic Press (2004) 51–66
7. Ran Tao, Zhenyang Li, Y.J.: Music genre classification using temporal information and support vector machine. In: ASCI Conference. (2010)
8. Chih-Wei Hsu, Chih-Chung Chang, C.J.L.: A Practical Guide to Support Vector Classication. Department of Computer ScienceNational Taiwan University, Taipei 106 (2010)
9. Aliaksandr Paradzinets, Hadi Harb, L.C.: Multiexpert system for automatic music genre classification. Technical report, Ecole Centrale de Lyon - Departement MathInfo (2009)
10. Karpov, I.: Hidden markov classification for musical genres. Technical report, Rice University (2002)

11. Andre Holzapfel, Y.S.: A statistical approach to musical genre classification using non-negative matrix factorization. IEEE Transactions on Audio, Speech, and Language Processing **16** (2008) 424–434
12. Tzanetakis, G.: Marsyas (2012) Software tool for Music Analysis, Retrieval and Synthesis for Audio Signals (Version 0.4.5).
13. Tzanetakis, G.: Marsyas User Manual. (2012)
14. Bishop, C.M.: Neural Networks for Pattern Recognition. 1 edn. Oxford University Press, USA (1996)
15. Bay, S.D.: Nearest Neighbor Classification from Multiple Feature Subsets. Intelligent Data Analysis **3** (1998) 191–209
16. Weihmann, T.: Wavsplit (2002) Software which splits large WAV files at given time positions (Version 1.2.1.).
17. Joakim Anden, S.M.: Multiscale scattering for audio classification. In: Proceedings of the 12th International Society for Music Information Retrieval Conference. (2011) 657–662
18. Philippe Hamel, Simon Lemieux, Y.B.D.E.: Temporal pooling and multiscale learning for automatic annotation and ranking of music audio. In: Proceedings of the 12th International Society for Music Information Retrieval Conference. (2011) 729–734
19. University of Waikato, N.Z.: Weka (1997) Machine learning software written in Java (Version 3.6.).

S. Markovski, M. Gusev (Editors): ICT Innovations 2012, Web Proceedings, ISSN 1857-7288

Automatic Fiber Placement (AFP) Technology, Actual State and Future Improvement through Using NDT (Ultrasonic) Equipment in On-line Processing

Prof. PhD Blagoja Samakoski[1], PhD Svetlana Risteska[1], Biljana Kostadinoska[1] and Ekaterina Sinadinova[2]

[1]Institute for Advanced Composites and Robotics (IACR), Prilep, Macedonia
{blagojas, svetlanar, biljanak}@iacr.edu.mk
[2]Mikrosam, Prilep, Macedonia
ekaterinas@mail.mikrosam.com

Abstract. In this paper are presented all of the technologies used for research at the Institute for Advanced Composites and Robotics (IACR). Focus is put on the automated fiber placement (AFP) technology and the equipment used to perform this technology. Analysis presented, refer to possible mistakes in the final product, composite part intended to be a primary structure of an airplane, due to the inability to know the influence of the compacting roller, or the characteristics of the prepreg on the occurrence of micro voids in the laminate of the part produced. As a response to this acknowledgement, preliminary research is done at IACR where ultrasonic sensor is used for nondestructive testing (NDT) in composite materials. This preliminary research has shown that it should be further inspected whether ultrasonic sensor following the compacting roller, could improve the impacting process and diminish the number of voids in the final product.

Keywords: IACR, Mikrosam, automated fiber placement (AFP), filament winding (FW), prepreg making, composites, voids(pores), NDT, ultrasonic sensor

1 Introduction

The research on which is based this paper is conducted at the Institute for Advanced Composites and Robotics (IACR), which is one of the very few institutes in the world, and only one in the region, that conducts research in this particular field. The institute is founded in 2009, by Mikrosam, Macedonian company which is a leader in the region, and among the best in the world in the composite machine industry. It is the only company that has developed the entire technology for producing composite products – prpereg making machine, filament winding machine and fiber and tape placement machine. As a result of this, IACR is equipped with the best composite technology which is an advantage that guarantees probably the most valid research in the field of composites. IACR offers comprehensive analysis and testing, done in

many of its laboratories for: software development, software engineering, motion control, chemical and mechanical composite research, data acquisition and process control, and for developing technological process of modern composites. This paper work is produced as a summary of the analysis made on a specific issue occurring in the AFP technology, which could result in a joint initiative to find an innovative solution for the voids' occurrence problem in AFP, and could further be readapted for other applications. The work explained in this paper is based on more than 30 years of extensive experience and knowledge in the composites' field and on analysis done on a daily basis by mechanical, electro-technical, software and chemical engineers who develop and constantly improve the upper mentioned technologies.

2 Composite Technologies

Composite materials due to their advantageous characteristics are becoming the materials of the future. These materials, although long ago discovered and widely used by high-tech industries (aerospace, automotive, marine, space), in other industries are yet to be discovered. Their specific characteristics strength, stiffness, and "weigh saving"[1]make them dominant over other materials and expand their application range, from leisure and sports to industrial complex components. Therefore, composites technologies are one of the most expensive and even for engineers yet unfamiliar technologies. Following this perception, short overview of the core composite technologies will precede the case study in this paper.

2.1 Prepreg Making

Prepreg by itself is a kind of a semi-product, "combination of a matrix (or resin) and fiber reinforcement"[2]. Due to its good mechanical characteristics, easy processing and lower cost, prepreg is used in aerospace, railway, marine, energy and other industries.

There are two main methods to make prepreg: hot melt – through which can be produced fabric and unidirectional (UD) prepreg, and solvent – through which can only be produced fabric prepreg.

Hot melt method starts by coating the resin on a silicon paper in a thin film, and then impregnating it onto a fiber on the prepreg machine, under roller pressure and heat [3].

In the solvent method, resin is dissolved in a solvent and dipped onto a fabric [4]. Later, impregnated prepreg is exposed to heating oven to decrease the solvent content [5].

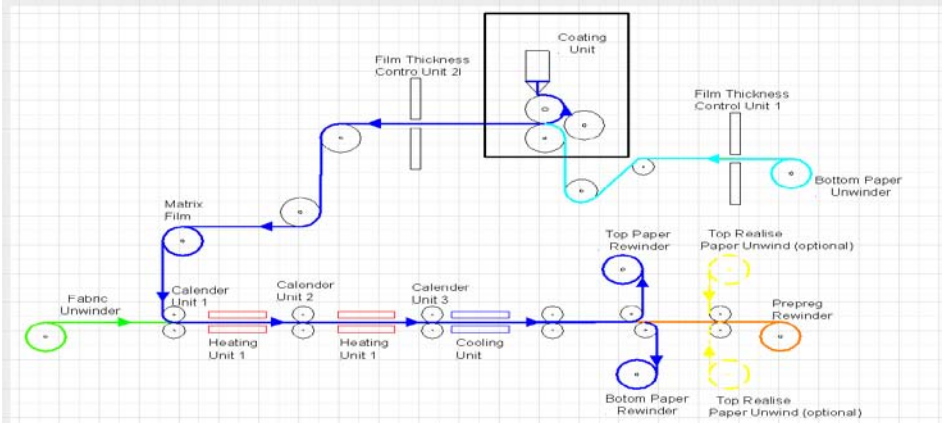


Fig. 1. Hot melt prepreg making method

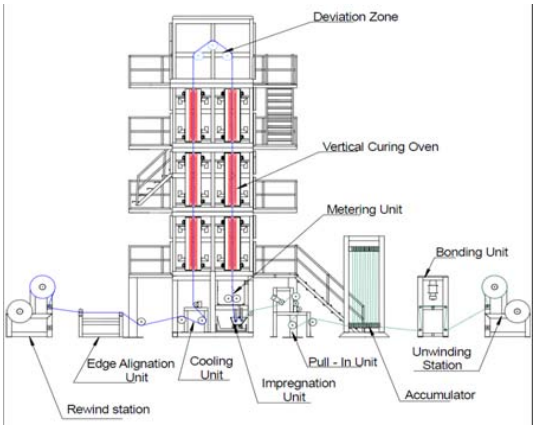


Fig. 2. Solvent prepreg making method

Fig. 3. Prepreg line -Type PLS-250-2F

Fig. 4. Vertical impregnating machine LI 300

(Source: Mikrosam, cases, prepreg production [6])

2.2 Filament Winding (FW)

Filament winding is a composite technology used to make high-pressure vessels (LPG, CNG), oxygen tanks, compressed natural gas cylinders, underwater pipes. It is a process in which "continuous reinforcements in a form of rovings or monofilaments are wound over a rotating mandrel" [7]. Winding angles and placement of reinforcements are controlled by specially designed machines, and so spherical, conical, and geodesic shapes can be made. Cylinders made out of composite materials surpass the quality and cost of heavy metal cylinders.

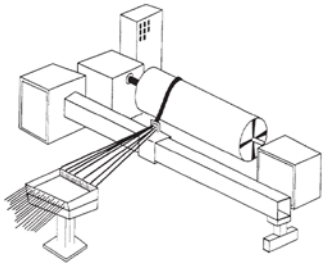


Fig. 5. Schematic of the FW process
(Source: Suong V. Hoa[8])

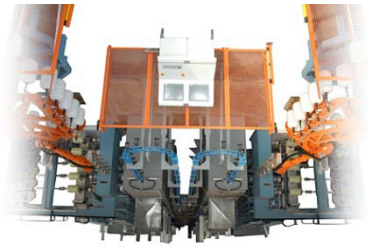


Fig. 6. Automated FW machine
(Source: Mikrosam [9])

2.3 Automated Fiber Placement (AFP)

Fiber placement is a process similar to filament winding [10]. This method is able to fully exploit special and an-isotropic characteristics of the composite materials as a benefit of the freedom this production process gives to the designer.

Automated fiber placement (AFP) is an automated composites manufacturing process of heating and compacting resin pre-impregnated non-metallic fibers on typically complex tooling (mandrels). The fiber usually comes in the form of what are referred to as "tows". A tow is typically a bundle of carbon fibers impregnated with epoxy resin and is approximately 3.2 mm wide by 0.1mm thick and comes on a spool.

AFP as such contributes to high productivity, improved quality, and low cost for large composite structures. Through automation, the process of FP has improved in lowering manufacturing cost (labor, material scrap, better control)[11].

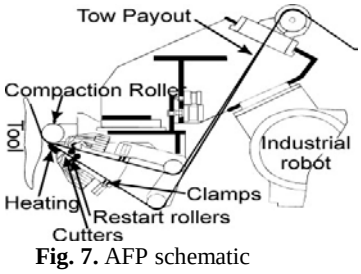


Fig. 7. AFP schematic
(Source: Salman Khan [12])

Fig. 8. AFP machine
(Source:Mikrosam [13])

3 Case study: Occurrence of Voids during AFP Process

3.1 NDT

In advanced technology applications such as aerospace, with industrial emphasis on economics and safety, it is critical to use and develop robust and practical composites' NDT methods. Composites' NDT encompasses a range of modified traditional and new tools including ultrasonic, x-ray, acoustic emission, thermal, optical, electrical and a variety of hybrid methods.

NDT stands for non-destructive testing. In other words, it is a way of testing without destroying. In today's world where new materials are being developed, older materials and bonding methods are being subjected to higher pressures and loads. NDT ensures that materials can continue to operate to their highest capacity without failing below the predetermined time limits. In addition, NDT can be used to ensure the quality right from raw material stage through fabrication and processing to pre-service and in-service inspection [14], [15].

The field of Nondestructive Testing (NDT) is very broad, interdisciplinary field that plays critical role in assuring that structural components and systems perform their function in a reliable and cost effective fashion. NDT technicians and engineers define and implement tests that locate and characterize material conditions and flaws that might otherwise cause planes to crash, reactors to fail, trains to derail, pipelines to burst, and a variety of less visible, but equally troubling events. These tests are performed in a manner that does not affect the future usefulness of the object or material. In other words, NDT allows parts and material to be inspected and measured without damaging them. Because it allows inspection without interfering with a product's final use, NDT provides an excellent balance between quality control and cost-effectiveness.

Nondestructive evaluation (NDE) is a term that is often used interchangeably with NDT. However, technically, NDE is used to describe measurements that are more quantitative in nature. For example, an NDE method would not only locate a defect, but it would also be used to measure something about that defect such as its size, shape, and orientation. NDE may be used to determine material properties, such as fracture toughness, formability, and other physical characteristics.

There are many NDT techniques/methods, which are used depending on four main criteria: type of material, type of defect, size of defect and location of defect. Therefore, it is important to choose the appropriate method[16].

Most Common NDT Methods are: Visual(optical and Laser), Liquid Penetrant, Laser techniques, Sherography, Holography, Radiography, Ultrasonic, Magnetic, Infrared thermography, Acustic Emission testing, Microwave technique

3.2 Ultrasonic Inspection

In ultrasonic testing, high-frequency sound waves are transmitted into a material to detect imperfections, or to locate changes in material properties. Most commonly used ultrasonic testing technique is pulse echo, whereby sound is introduced into a test

object and reflections (echoes) from internal imperfections, or part's geometrical surfaces, are returned to a receiver. Below is given an example of composite part. High frequency sound waves are introduced into a material and they are reflected back from surfaces or flaws. Ultrasound inspection methods are powerful tools for nondestructive testing and are widely used in the industry because high resolutions are possible depending on the chosen frequency (100 kHz to 40 MHz). In ultrasonic testing, stress waves are injected into the material or component to be examined and then, transmitted/reflected beams are monitored. Ultrasonic measurements can determine the location of a discontinuity in a part or structure by accurately measuring the time required for a short ultrasonic pulse generated by a transducer to travel through a thickness of a material,reflect from the back or the surface of the discontinuity, and be returned.

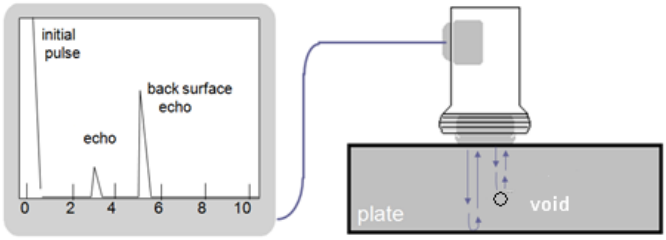


Fig. 9. (Example: use of noninvasive techniques to determine the integrity of a material, component or structure, or to quantitatively measure some specific characteristics of an object.)

Ultrasonic waves are used to evaluate the condition of a material, anomalies' absorption, or to deflect the sound waves, which are then detected as changes in the waves: holes, delaminations, voids; damage, debonds; resin-rich; poorareas

Our research suggests that we use these technologies to detect defects (pores) voids, in the process of placing the fibers through AFP technology.

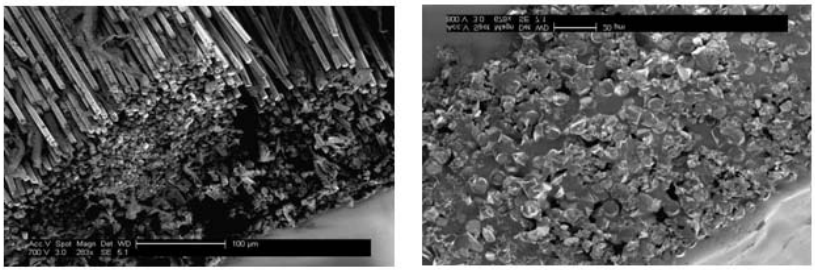
3.3 Voids (Pores)

Presence of pores in the final product is a significant mistake in the technological process of work. Due to the inability to know the influence of the compacting roller, or the characteristics of the prepreg as factors that influence the occurrence of pores in the laminate in the final product, today at IACR preliminary research is done on the use of ultrasonic sensor as one of the methods for NDT. The method used to detect defects without destructing the material (the prepreg in this case) is one of the advanced methods used in the world. By placing ultrasonic sensors above the compacting roller on the AFP machine, pores (compressed air) in the laminate might be detected while placing. With this early air detection (while placing the fibers), certain parameters (pressure of the compacting roller, temperature of placement, and speed of placement) that influence pores occurrence might be changed. After this change is made, pores occurrence will be observed again. During the on-line processing, the

quality of the final product will be improved and this will result in better mechanical characteristics of the final product material. With this practice, laminate voids and/ or improper track order might be observed, which will immediately signal to change the parameters in case there is a mistake that might result in a bad quality product.

Experiments and literature research show that the presence of more than 2-3% of pores in the final composite part has strong influence on the mechanical characteristics of a material [17], [18].

Below are given two SEM pictures with different % of pores in the sample examined. The 1st picture is an example of bad quality prepreg, and the 2nd is an example of a good quality prepreg, which needs to be used in the AFP process.



~ 10 % voids ~ 1,3 % voids

Fig. 10.

4 Solution – Improving the Characteristics of the Material

High number of voids in the material causes a problem on the forces that influence the final product. The aim of these methods is during the process of consolidation via sensors, defects (pores) to be early detected, impacting process to be improved, and the occurrence of voids in the final product to be decreased.

It is well known that void content of a composite may significantly affect its mechanical properties and therefore, it is an important indicator of composite's quality. Interlaminar shear strength and void content of the consolidated parts are considered to be key quality indicators.

AFP processing for the on-line consolidation system is determined by adjusting the three system processing parameters: roller pressure, speed, and temperature. Therefore, this study focused on these three parameters, and was done in two steps. At first, a two-factor central composite design of experiments was used to define the combination of processing parameters, and next, void content was calculated (SEM images).

The inspection of final composite part (placement prepreg) is a time consuming and expensive procedure, mainly due to necessity inspection using conventional NDT methods. Therefore, the objective of this study is to develop an ultrasonic technique, suitable for placement of preperg and voids determination, without interrupting the process. The most promising technique which enables inspection at a relatively long range and can be used for inspection prepreg from an outside perimeter is based on

ultrasonic guided waves. Guided waves can propagate long distances in planar and tubular structures, and have already been used for voids' inspection.

Our research anticipates that particularly in AFP, an ultrasonic sensor that will monitor the compacting roller, can have an impact on process improvement and voids' occurrence decrease in the final product.

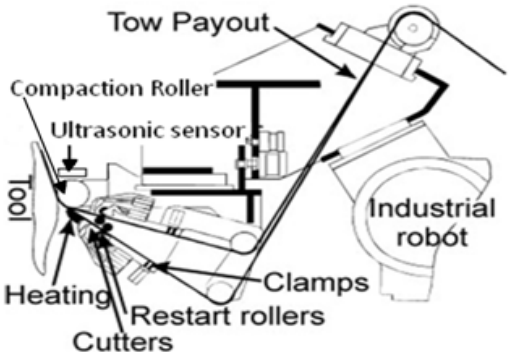


Fig. 11. (AFP with inserted ultrasonic sensor above the compaction roller)

5 Conclusion

Case study explained in this paper, and corresponding research made at IACR (in one of the most advanced composite materials laboratories), show that in order to acquire full advantage of the quality and cost saving of composite products, composite materials during the process of production have to be carefully inspected, examined and cured, to avoid troubles when they are transformed into final products. Since the fact that main cause for cracks, bursts, derails is the occurrence of voids in the laminate of the final product is not a novel issue, the idea of this paper is to initiate use of a different voids' inspection technology that will not require interruption of the production process (NDT). This technology is based on ultrasonic sensor which through guided waves would foresee possible voids, and simultaneously would give data on the parameters (pressure, temperature, speed) of the compacting roller that influence these cracks during the AFP process (as a process on which this research is focused).

Validity of this study lies in the expected results from the use of the proposed ultrasonic inspection technology, which suggests that being able to inspect and anticipate voids beforehand would mean high efficiency of the core composite technology, better quality of the composite final products, and competitive advantage for the producers and beneficiaries of these products.

References:

1. Pichai Rusmee.: High strength composites, Utah University, pp. 2, (2005) http://www.mech.utah.edu/~rusmeeha/labNotes/composites.html;

2. Hexel,: Prepreg technology, pp. 4, http://www.hexcel.com/Resources/DataSheets/Brochure DataSheets/Prepreg_Technology.pdf
3. Umeco,: An introduction to Advanced Composites and Prepreg Technology, pp 8-11, (2006)
4. Umeco, pp. 9-11 (2006);
5. Jiaching Liu, Wen-Yei Jang,: Development of a theoretical model for a solvent-type pre-preg manufacturing process, pp. 360, http://deepblue.lib.umich.edu/bitstream/2027.42/31924/1/0000877.pdf;
6. Mikrosam, Cases, Prepreg production, http://www.mikrosam.com/new/cases/en/22/;
7. Suong V. Hoa: chapter 5: Filament Winding and Fiber Placement, Principles of the Manufacturing of Composite Materials;
8. Suong V. Hoa;
9. Mikrosam, Cases, Filament Winding, http://www.mikrosam.com/new/article/en/automated-filament-winding-line-for-lpg-cng-hydrogen-and-other-types-of-high-pressure-vessels/;
10. Suong V. Hoa;
11. Salman Khan,:chapter 2:Automated Fiber Placement Process Overview, Thermal Control System Design for AFP Process, http://spectrum.library.concordia.ca/7393/1/Khan_MASc_F2011.pdf;
12. Dirk H.-J.A. Lukaszewicz, Carwyn Ward, Kevin D. Potter,: The engineering aspects of automated prepreg layup: History, present and future;
13. Mikrosam, Cases, AFP: Complete system, http://www.mikrosam.com/new/article/en/automated-fiber-placement-the-complete-system/ ;
14. Nicholas J.Carino, :Nondestructive Test Method" , Concrete Construction Engineering Handbook, Chapter 19, CRC Press Editor, 19/1-68pp, 1997;
15. Liudas MAŽEIKA, Rymantas KAŽYS, Renaldas RAIŠUTIS, Andriejus DEMČENKO, Reimondas ŠLITERIS: Long-range Ultrasonic Non-destructive Testing of Fuel Tanks, ECNDT 2006 - Fr.2.2.4;
16. Matthew D. Lansing, Michael W. Bullock "Endoscopic Shearography and Thermography Methodsfor Nondestructive Evaluation of Lined Pressure Vessels" Final Technical Report (October 1995 – September, 1996);
17. A. Bruce Hulcher and Joseph,: Dry Ribbon for Heated Head Automated Fiber Placement;
18. A. Bruce Hulcher, David M. McGowan and Brian W. Grimsley,: Processing and Testing of Thermoplastic Composite Cylindrical Shells Fabricated by Automated Fiber Placement;

Investigations on probabilistic analysis synthesis systems using bidirectional HMMs

Ronald Römer

TU- Cottbus, Chair of Communication Engineering,
Konrad Wachsmann Allee 1 , 03046 Cottbus, Germany
ronald.roemer@tu-cottbus.de

Abstract. This contribution deals with preliminary investigations on the behavior of the cortical algorithm in probabilistic hierarchic analysis synthesis systems. Such a subsystem is one the key components of Cognitive Dynamic Systems or Cognitive User Interfaces respectively. Both systems are typically characterized by the cybernetic circle that describes the perception of the environment along the sensory hierarchy, the selection of an optimal response and action articulation on the environment along the motor hierarchy. Further, a cognitive system should be able to predict the consequences of its own actions. For this purpose an inner model of the communication participant and its simulation is required. Based on this assumption, the bidirectional flow of information in analysis synthesis systems may be justified. Even though the cortical algorithm is drawn to cascaded bidirectional HMMs (CBHMMs), in this study the impact of the bidirectional information processing has been investigated just for simple single layer bidirectional HMMs. The proposed experiment is based on Shannon's channel model, at which synthetic source data are transmitted to the receiver - disturbed by Gaussian noise at different SNR. Finally, we compare the state recognition rate for all possible setups using single layer HMMs.

Keywords: Cognitive Dynamic Systems, Analysis-Synthesis-Systems and Cortical Algorithm.

1 Introduction

Recently S. Haykin proposed to combine model based signal processing techniques with new ideas from neurosciences. The resulting systems are called Cognitive Dynamic Systems [1]. Such systems are indicated by an intelligent feedback channel which is used for the target-oriented behaviour of the system. It may be described by the cybernetic cycle at which the system has to perceive it's environment, has to find a goal directed decision under uncertainty and executes an action at the environment.

In contrast to the classical control loop theory the analysis and synthesis stages comprise hierarchically organized structures. Both structures are indicated by a bidirectional flow of information and level specific working memories

S. Markovski, M. Gusev (Editors): ICT Innovations 2012, Web Proceedings, ISSN 1857-7288

which is justified by findings from neuroscience [2]. Moreover, it seems that the neocortex of the brain does not consist of a collection of specialized and dedicated cortical architectures, but instead possesses a fairly uniform, hierarchically organized structure. This uniformity implies that the same general computational processes are performed across the entire neocortex, even though different regions are known to play different functional roles [3]. First conceptual attempts to consider biologically motivated structures are made in [4] for instance. This scheme is based on Bayesian update procedures and is called Cortical algorithm. Both the analysis- and the synthesis stages are using predictive information that stem from neighboring hierarchical levels. Hence, a bidirectional flow of information takes place at the same time and is fused in each level according to the Bayesian procedure mentioned above [5]. A consistent justification of the simultaneous bidirectional flow of information may be sustained by the discovery of mirror neurons and the usage of an inner model of the communication participant [6]. The objective of this study is to determine whether these new ideas may help to improve technical transmission systems. Particularly, we want to investigate the impact of the bidirectional processing scheme on probabilistic analysis synthesis systems in noisy conditions.

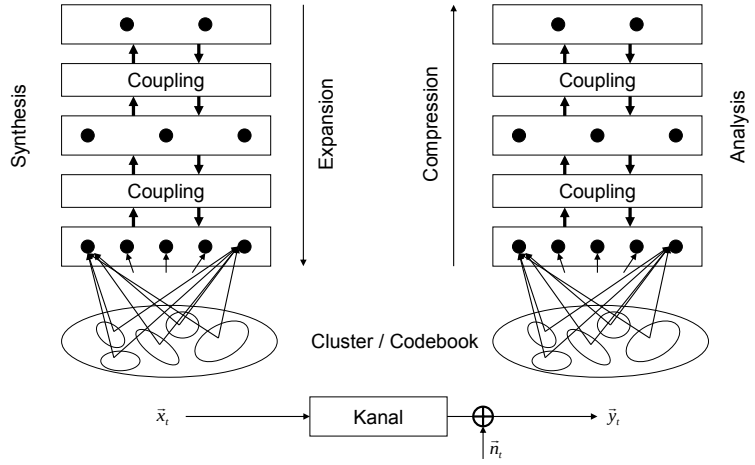


Fig. 1. Hierarchical analysis synthesis system used in a simple transmission model.

2 Cascaded Bidirectional HMMs

Hidden Markov Models may be separated into a static part, which is used to compute the observation probability and a second part, which captures the dynamics of the state transitions. Both parts are reflected by the state space representation of HMMs

$$\boldsymbol{p}\left[s(t+1)\right] = \boldsymbol{A} \cdot \boldsymbol{p}\left[s(t)\right] + \boldsymbol{\pi} \cdot \delta(t) \quad \text{and} \quad \boldsymbol{q}(t) = \boldsymbol{C} \cdot \boldsymbol{p}\left[s(t)\right]. \tag{1}$$

By the introduction of further additional HMM-layers $d = \{0...D-1\}$, the state probability distribution at each layer is computed by a temporal prediction distribution, a bottom-up distribution and a top-down distribution (see Figure 1). The figure depicts the structure of an analysis-synthesis system based on a CBHMMs. Transitions within the hierarchical levels are described by A-matrices. Transitions between hierarchical levels are modeled by coupling-matrices. Analysis as well as synthesis are characterized by an bidirectional information processing scheme and are embedded in a transmission model according to Shannon. In contrast to the case of synthesis where an additional bottom-up distribution is utilized, in the analysis case an additional top-down distribution is involved. CBHMM structures are extensively mathematically described in [5], information theoretic insights are published in [7]. Hence, at this place only the heart of the Cortical algorithm - the fusion equation - is given and explained in forward-backward notation.

$$\gamma_j^d(k) \propto \left[\sum_l \beta_l^{d-1}(k) \cdot b_{j,l}^d\right] \cdot \left[\sum_i \gamma_i^d(k-1) \cdot a_{i,j}^d\right] \cdot \left[\sum_r \alpha_r^{d+1}(k) \cdot b_{j,r}^{d+1}\right]. \tag{2}$$

In this equation three components are combined to get the a-posteriori state distribution at every level d and at each instant k. The temporal prediction based on $\gamma(k-1)$ from level d, the bottom-up component: $\alpha(k)$ from level $d+1$ and the additional third term, the top-down component: $\beta(k)$ from level $d-1$.

3 Experiments based on bidirectional HMMs

To demonstrate the feasibility of the bidirectional processing scheme we propose the following experiment which is based on Shannon's model. In this model the source selects symbols and the transmitter runs through a number of HMM states to generate feature vectors. These feature vectors are disturbed by Gaussian noise at different SNR and then received by the sink. Afterwards the state sequence is decoded by the Viterbi-algorithm. Finally, we compare the state recognition rate for all possible setups. When we apply the Cortical Algorithm the following improvements are expected: In case of analysis the observations should be interpreted more reliable if contextual knowledge from higher levels is

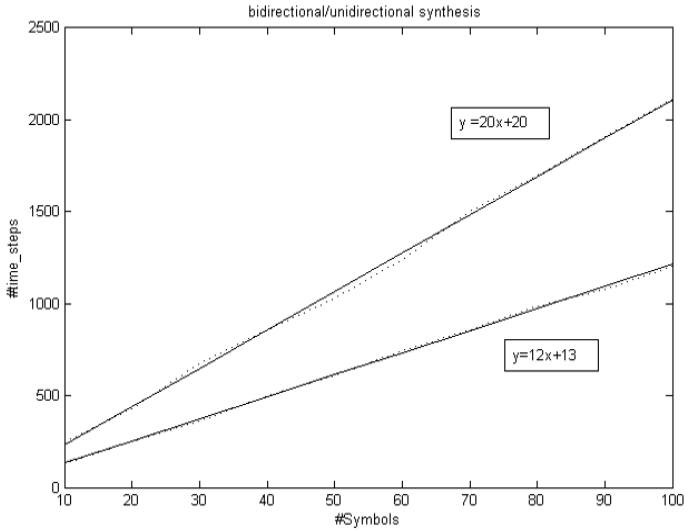


Fig. 2. Unidirectional/bidirectional synthesis: symbol duration approximated by linear regression analysis.

exploited (e.g. top-down-predictions). In case of synthesis the emission discontinuities are avoided when contextual knowledge from lower levels is considered (e.g. bottom-up-predictions). It is worth to mention, that the fusion equation may be realized by two methods (soft/hard). Because of the similar results for both methods the details are omitted here.

To avoid any side effects that stem from the hierarchy we started with a simple single layer HMM with 5 states and 5 classes. The HMM parameters were chosen identical for analysis and synthesis and were set manually, in doing so no training was necessary. For each SNR the source has emitted 20 symbols. The received disturbed features are used to decode the state sequence. Further, to investigate the noise influence on the receiver predictions, we have distinguished between signal and model based predictions.

4 Results

The results for the two kinds of prediction methods are given in figure 3 and 4 respectively and are partly taken from [8]. Both figures depict the results for all four processing modes according to the combination scheme: transmitter-mode/receiver-mode. The first finding was, that in case of bidirectional synthesis the transmitter needs about two times more to run through the HMM states as in the case of the unidirectional synthesis (see figure 2). This observation may be interpreted as insertion of redundancy, which could be exploited by the receiver. However when we used signal based predictions, the receiver was not able to

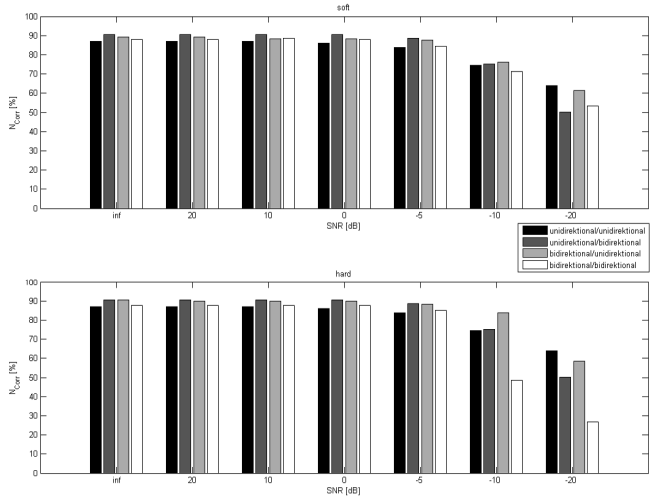


Fig. 3. SNR dependent state recognitions (signal based predictions).

exploit this redundancy. It may be explained by the fact, that with increasing noise level the accuracy of the signal based predictions decreases of course. Hence, the fusion of prediction and measurement cannot improve the state estimation or the recognition rate.

When we used model based predictions instead - which are not affected by the signal - we observed a significant improvement of the recognition rate over a wide range of SNR. We interpret this promising result as a strong indication for the need of inner model simulations.

5 Conclusion

Based on the results of this study we conclude that signal based predictions may deliver only noisy forecasts. Opposed to that finding, model based predictions may improve the state recognition rate over a wide range of SNR. Hence, we conclude that the bidirectional processing scheme may be applied gainfully in matched conditions (bidirectional/bidirectional). Future studies will have to show whether this findings can be generalized to CBHMMs. Furthermore we aim to study the impact of different kinds of model based predictions. Particularly the usage of Markov Decision Processes (MDP) and Partially Observable Markov Decision Processes (POMPD) we want to take into account. Furthermore, to build a flexible hierarchical speech recognition and synthesis system, we plan to realize the Cortical algorithm using weighted Finite State Transducers (FSTs). The transition from HMMs to weighted FSTs can be easily done

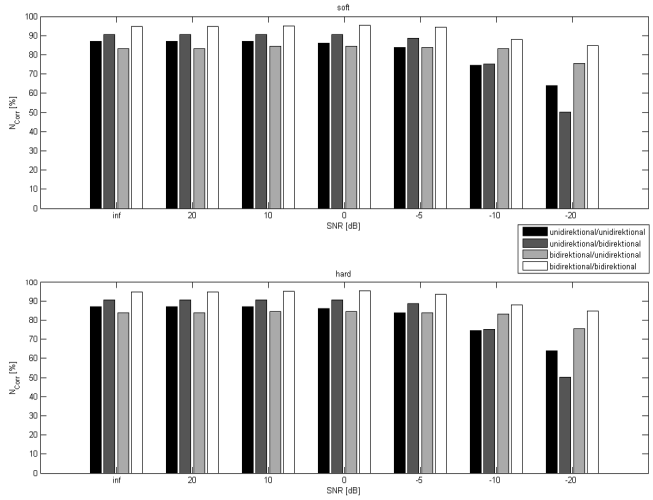


Fig. 4. SNR dependent state recognitions (model based predictions).

using arc-emission HMMs. In this case, both the transition probabilities and the emission probabilities are placed at the edges of the weighted FST.

References

1. Haykin, S.: Cognitive Dynamic Systems. IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP-2007, vol. 4, pp. 1369–1372, (2007)
2. Fuster, J.M.: Cortex and Mind, Unifying Cognition. Oxford Press (2003)
3. Mountcastle, V.: An Organizing Principle for Cerebral Function: The Unit Model and Distributed Systems. In: The Mindful Brain, Gerald M. Edelman and Vernon B. Mountcastle,eds.,(1978)
4. Mumford, D., Lee, T.S.: Hierarchical Bayesian inference in the visual cortex. Journal of the Optical Society of America, vol. 20, no. 7, (2003)
5. Roemer, R., Herbig, : Konzeptionelle Beschreibung des Corticalen Algorithmus und seine Verwendung in der automatischen Sprachverarbeitung. In: 20. Konferenz Elektronische Sprachsignalverarbeitung ESSV-2009, pp. 33–40. TUD Press, Dresden (2009), (in german)
6. Roemer, R.: Beschreibung von Analyse-Synthese-Systemen unter Verwendung von CBHMM's. In: 22. Konferenz Elektronische Sprachsignalverarbeitung ESSV-2012, pp. 67–76. TUD Press, Aachen (2011), (in german)
7. Roemer, R.: A Cortical Approach based on Cascaded Bidirectional Hidden Markov Models. In: Behavioural Cognitive Systems, Lecture Notes in Computer Science (2012)
8. Roemer, R.: Untersuchngen zum Cortikalen Algorithmus unter Verwendung con bidirektionalen HMMs. In: 23. Konferenz Elektronische Sprachsignalverarbeitung ESSV-2012, pp. 252–261. TUD Press, Cottbus (2012), (in german)