ICCGI 2011

The Sixth International Multi-Conference on Computing
in the Global Information Technology

June 19-24, 2011

Luxembourg City, Luxembourg

ICCGI 2011 Editors

Constantin Paleologu, University Politehnica of Bucharest, Romania

Constandinos Mavromoustakis, University of Nicosia, Cyprus

Marius Minea, University Politehnica of Bucharest, Romania

ICCGI 2011

Foreword

The Sixth International Multi-Conference on Computing in the Global Information Technology [ICCGI 2011], held between June 19 and 24, 2011, in Luxembourg, continued a series of international events covering a large spectrum of topics related to global knowledge concerning computation, technologies, mechanisms, cognitive patterns, thinking, communications, user-centric approaches, nanotechnologies, and advanced networking and systems. The conference topics focus on challenging aspects in the next generation of information technology and communications related to the computing paradigms (mobile computing, database computing, GRID computing, multi-agent computing, autonomic computing, evolutionary computation) and communication and networking and telecommunications technologies (mobility, networking, bio-technologies, autonomous systems, image processing, Internet and web technologies), towards secure, self-defendable, autonomous, privacy-safe, and context-aware scalable systems.

This conference intended to expose the scientists to the latest developments covering a variety of complementary topics, aiming to enhance one's understanding of the overall picture of computing in the global information technology.

The integration and adoption of IPv6, also known as the Next Generation of the Internet Protocol, is happening throughout the World at this very moment. To maintain global competitiveness, governments are mandating, encouraging or actively supporting the adoption of IPv6 to prepare their respective economies for the future communication infrastructures. Business organizations are increasingly mindful of the IPv4 address space depletion and see within IPv6 a way to solve pressing technical problems while IPv6 technology continues to evolve beyond IPv4 capabilities. Communications equipment manufacturers and applications developers are actively integrating IPv6 in their products based on market demands.

IPv6 continues to represent a fertile area of technology innovation and investigation. IPv6 is opening the way to new successful research projects. Leading edge Internet Service Providers are guiding the way to a new kind of Internet where any-to-any reachability is not a vivid dream but a notion of reality in production IPv6 networks that have been commercially deployed. National Research and Educational Networks together with internationally known hardware vendors, Service Providers and commercial enterprises have generated a great amount of expertise in designing, deploying and operating IPv6 networks and services. This knowledge can be leveraged to accelerate the deployment of the protocol worldwide.

We take here the opportunity to warmly thank all the members of the ICCGI 2011 Technical Program Committee, as well as the numerous reviewers. The creation of such a broad and high quality conference program would not have been possible without their involvement. We also kindly thank all the authors who dedicated much of their time and efforts to contribute to ICCGI 2011. We truly believe that, thanks to all these efforts, the final conference program consisted of top quality contributions.

Also, this event could not have been a reality without the support of many individuals, organizations, and sponsors. We are grateful to the members of the ICCGI 2011 organizing committee for their help in handling the logistics and for their work to make this professional meeting a success.

We hope that ICCGI 2011 was a successful international forum for the exchange of ideas and results between academia and industry and for the promotion of progress in the area of computing in the global information technology.

We are convinced that the participants found the event useful and communications very open. We also hope the attendees enjoyed the historic charm Luxembourg.

ICCGI 2011 Chairs:

Ani Calinescu, Oxford University, UK
Emmanuel Chaput, ENSEEIHT / IRIT-CNRS, France
Tibor Gyires, Illinois State University, USA
Jean Johnson, The Inclusion Trust - Takeley, UK
Hermann Kaindl, TU-Wien, Austria
Yasushi Kambayashi, Nippon Institute of Technology, Japan
Constandinos Mavromoustakis, University of Nicosia, Cyprus
Mary Luz Mouronte López, Ericsson S.A., Spain
Krishna Murthy, HCL America, USA
Marius Minea, University Politehnica of Bucharest, Romania
Constantin Paleologu, University Politehnica of Bucharest, Romania
Liviu Panait, Google Inc., USA
Amir Razavi, University of Surrey - Guildford, UK
José Rouillard, Université Lille Nord, France
John Terzakis, Intel, USA
Luc Vouligny, Institut de Recherche d'Hydro-Québec - Varennes, Canada

ICCGI 2011

Committee

ICCGI Advisory Chairs

Emmanuel Chaput, ENSEEIHT / IRIT-CNRS, France
Constantin Paleologu, University Politehnica of Bucharest, Romania
Tibor Gyires, Illinois State University, USA
Luc Vouligny, Institut de Recherche d'Hydro-Québec - Varennes, Canada
Hermann Kaindl, TU-Wien, Austria

ICCGI Industry/Research Chairs

Krishna Murthy, HCL America, USA
Jean Johnson, The Inclusion Trust - Takeley, UK
John Terzakis, Intel, USA
Yasushi Kambayashi, Nippon Institute of Technology, Japan
Mary Luz Mouronte López, Ericsson S.A., Spain

ICCGI Special Area Chairs

Evolutionary Computation
Liviu Panait, Google Inc., USA

Autonomic/Autonomous Systems
Ani Calinescu, Oxford University, UK

Knowledge/Cognition
Constandinos Mavromoustakis, University of Nicosia, Cyprus

e-Learning/Mobility
José Rouillard, Université Lille Nord, France

Industrial systems
Marius Minea, University Politehnica of Bucharest, Romania

Digital eco-systems
Amir r Razavi, University of Surrey - Guildford, UK

ICCGI 2011 Technical Program Committee

Werner Aigner, FAW, Austria
Areej Al-Wabil, King Saud University - Riyadh, Saudi Arabia
Mohammed Aldasht, Palestine Polytechnic University - Hebron, Palestine
Panos Alexopoulos, IMC Technologies SA - Athens, Greece
Ali Alharbi, The University of Newcastle, Australia
Fernando Almeida, University of Porto, Portugal
Matjaz B. Juric, University of Ljubljana, Slovenia

Mark Perry, University of Western Ontario - London, Canada
Dana Petcu, Western University of Timisoara, Romania
Willy Picard, Poznan University of Economics, Poland
Dorin Popescu, University of Craiova, Romania
Radu Prodan, University of Innsbruck, Austria
Amir r Razavi, University of Surrey - Guildford, UK
Kornelije Rabuzin , University of Zagreb - Varazdin, Croatia
Ramutis Rindzevicius, Kaunas University of Technology, Lithuania
José Rouillard, Université Lille Nord de France
Pawel Rózycki, University of Information Technology and Management - Rzeszow, Poland
Glen Sagers, School of Information Technology, Illinois State University, USA
Ana Sasa, University of Ljubljana, Slovenija
Ashok Sharma, Mahindra Satyam, India
Marcin Solarski, IBM Software Services - Cracow, Poland
Cosmin Stoica Spahiu, University of Craiova, Romania
Yongning Tang, School of Information Technology, Illinois State University, USA
Paulius Tervydis, Kaunas University of Technology, Lithuania
John Terzakis, Intel, USA
Guglielmo Trentin, National Research Council - Genoa & University of Turin, Italy
Patravadee Vongsumedh, Bangkok University, Thailand
Luc Vouligny, Institut de Recherche d'Hydro-Québec - Varennes, Canada
Mihaela Vranic, University of Zagreb, Croatia
Chieh-Yih Wan, Intel Corporation, USA
Zhengchuan Xu, Fudan University - Shanghai, P. R. China
Wei Zhang, Microsoft, USA
Zhengxu Zhao, Shijiazhuang Tiedao University, China

Copyright Information

For your reference, this is the text governing the copyright release for material published by IARIA.

The copyright release is a transfer of publication rights, which allows IARIA and its partners to drive the dissemination of the published material. This allows IARIA to give articles increased visibility via distribution, inclusion in libraries, and arrangements for submission to indexes.

I, the undersigned, declare that the article is original, and that I represent the authors of this article in the copyright release matters. If this work has been done as work-for-hire, I have obtained all necessary clearances to execute a copyright release. I hereby irrevocably transfer exclusive copyright for this material to IARIA. I give IARIA permission or reproduce the work in any media format such as, but not limited to, print, digital, or electronic. I give IARIA permission to distribute the materials without restriction to any institutions or individuals. I give IARIA permission to submit the work for inclusion in article repositories as IARIA sees fit.

I, the undersigned, declare that to the best of my knowledge, the article is does not contain libelous or otherwise unlawful contents or invading the right of privacy or infringing on a proprietary right.

Following the copyright release, any circulated version of the article must bear the copyright notice and any header and footer information that IARIA applies to the published article.

IARIA grants royalty-free permission to the authors to disseminate the work, under the above provisions, for any academic, commercial, or industrial use. IARIA grants royalty-free permission to any individuals or institutions to make the article available electronically, online, or in print.

IARIA acknowledges that rights to any algorithm, process, procedure, apparatus, or articles of manufacture remain with the authors and their employers.

I, the undersigned, understand that IARIA will not be liable, in contract, tort (including, without limitation, negligence), pre-contract or other representations (other than fraudulent misrepresentations) or otherwise in connection with the publication of my work.

Exception to the above is made for work-for-hire performed while employed by the government. In that case, copyright to the material remains with the said government. The rightful owners (authors and government entity) grant unlimited and unrestricted permission to IARIA, IARIA's contractors, and IARIA's partners to further distribute the work.

Table of Contents

Data Mining Classification: *The Potential of Genetic Programming*

Nabil H. El Kadhi
Computer Engineering Department Chairperson
AHLIA University - Manama, Bahrein
EPITECH- Kremlin Bicêtre FRANCE
nelkadhi@ahliauniversity.edu.bh

Fatima A. Habib
MITCS Program Student
AHLIA University, Manama, Bahrein
ARAMCO Saudi Arabia
fatima.habib@aramco.com

***Abstract*—Data Mining (DM) is one of the techniques used for the process of searching through a huge volume of data in a database to uncover useful and interesting information. Classification is a commonly studied issue in data mining. It involves predicting the categorical attribute (class) value on the basis of other attributes (predicting attributes) values. One of the classification approaches to data mining is gene expression programming (GEP), which is a development of Genetic algorithm (GA) and Genetic Programming (GP). In this paper, we investigate the potential of genetic programming for data mining classification. It is therefore important to investigate the advantages and challenges associated with using tree-based Genetic Programming algorithms for data mining classification. This study determines how better data mining classification performance can be achieved using genetic programming. It demonstrates how the search scope can be refined through heuristics and machine learning methods to reduce and change the search space for our Genetic Programming classifiers. A specific design and application of GP to two classes data analysis is presented as well as a set of experimental results showing the efficiency of the suggested application.**

Keywords - Data Mining; Genetic Algorithm; Genetic Programming; Classification.

I. INTRODUCTION

In this paper, we are interested in analyzing a large volume of data. With huge data volume we have a huge potential for discovering knowledge through prediction and description. In the same time it creates a problem because of the difficulty of knowledge extraction and data classification. Therefore, there is a need to develop intelligent and powerful tools to derive interesting, specific, and useful information out large databases. Well, a number of techniques have been developed to aid in data mining [11] and have even proved to be practical. It is important, however, to note that data mining involves a number of tasks and not all these tasks are suitable for data analysis. In light of this, there has been growing interest on classification as a suitable approach to data mining leading to an emergence of a number of techniques that can be used in the classification task. Some researchers have continuously argued that techniques that use genetic algorithm (GA) are superior to conventional approaches in data mining classification in regard to performance, search space and time. Note that there has also been an emergence of several GA-based approaches and system for data mining classification. Genetic programming has evolved out of GA concepts, and has increasingly been accepted as a suitable classification technique. Most researchers have not explored extensively variant GP approaches to generally validate GP as a suitable method to data mining classification. This paper does not negate the usefulness of other approaches but investigates the potential of GP as a classification technique. Therefore, it uses the SQL-based GP to validate it as the most suitable technique for purposes of data mining classifications. GP is emerging as a powerful technique in deriving knowledge from large databases. The objectives of the study are:

- To show the potential of GP in induction of classifiers and validate the technique as suitable in data mining.
- To show that the search scope can be reduced through heuristics and GP techniques to facilitate data mining classification in large databases.
- To illustrate the process and implementation of data mining classification through a GP-based technique.
- To investigate the challenges and advantages of using the tree-based GP algorithms in data mining classification.

The paper is organized as following. Section II review some related works. Section III come through the proposed system design including all the Genetic algorithm parameters and organization. Section IV deals with implementation aspects and test results before concluding in Section V.

II. RELATED WORKS

Saraee and Sadjay [9] presented three approaches for classification via GP and illustrated that GP is superior to the traditional techniques for classification in terms of performance on space and time required for processing. Thus, with GP it is possible to create optimized classification rules in a manner that the traditional techniques cannot accomplish at low cost. Furthermore, the method allows the comparison of attributes of same type within a data set by generating specific set of rules. Ten et al. [10] have introduced a concept mapping method for evaluating the fitness of individuals through improved representation of booleanized attributes and token

 ISBN: 978-1-61208-139-7

competition. The approach applies a covering algorithm that uses a memory vector similar to an artificial immune system to generate several rules while eliminating the redundant ones. This GP classifier is authenticated upon a number of benchmark datasets chosen from the UCI Machine Learning Repository. Thus, the GP approach with a basic tree structure of only 'AND' and 'NOT' functions is effective in representing solutions in classifications. Furthermore, this approach is capable of generating understandable rules from datasets even when the background information is lacking on the particular dataset. On the other hand, the approach indicated low performances on some datasets, which nonetheless, does not disqualify the GP classifier as a potential method in data mining.

A new variant of linear genetic programming where genetic programs are characterized by the linear sequences of a C programming language is introduced in [1]. This system uses an algorithm that eliminates instructions that are not effective (the introns) prior to the execution of a program during fitness processing. This allows a substantial acceleration in the implementation speed, which is significant with composite data sets since they are being used in real-time.

Folino et al. [5] introduced a variant of cellular genetic programming that uses the boosting and bagging techniques to induce a group of predictors in data classification. Bagging is based on bootstrap samples or replicates of same size of the training set. Boosting enhances the performance of the weak learning algorithm. Carreno et al. [3] conclude further that REC has obtained high performance in building understandable classifiers with impressive predictive quality. Muharram and Smith [7] showed that genetic programming is efficient in generating highly predictive non-linear features using the novel attribute group. Espejo et al. [4] presented a number of solutions related to the use of GP in classification purpose have been evaluated. It revealed that the flexibility of this method makes it suitable in formulation of classifiers as well as in some pre-processing and post processing tasks. However, these researchers identify high training time as a drawback of GP. Furthermore, this becomes worse when large amounts of data are involved. Sakprasat et al. [8] have demonstrated the usefulness of strongly-typed GP in automatic approval of credit. GP can be applied in data mining for a problem regarding automatic approval of credit, and it is effective even with data that lacks some values. GP has been shown to be a superior method of classification in terms of reduced amount of space and time required for processing. The approach has also been credited with reduced number of rules, which is achieved through comparison of same type attributes. GP has also the capacity to generate comprehensible rules even when some information is absent. So many GP techniques have been used in the market using classification.

III. PROPOSED SYSTEM DESIGN

This paper focus on using Genetic Programming for classification task in data mining. This section presents the design of the proposed GP model. It shows how various GP techniques would be incorporated into an algorithm that would be used in data mining classification. The various components of the GP algorithms are highlighted.

A. GP Model and Abstract Tree

The application of GP techniques, including the natural selection should allow generation of better data mining solutions. Therefore, this paper proposes a GP algorithm that has immense benefits in induction of classifiers. The process of induction will be characterized by random selection of a given population of computer programs that will be used to produce subsequent generation by applying genetic operators, as it is with biological evolution. A fitness function, which is a binary function, will be used to direct the process of the GP algorithm in selecting the computer programs or rather the individuals. A grammar-based abstract or derivation trees would be used in the representation of computer programs in this GP model using abstract trees [6]. The abstract trees have the advantage of restricting the destructive aspect of the genetic operators on computer programs such syntactically erroneous code is not created. A context-free grammar would be used as the basis for constructing the abstract trees. The trees would be characterized by a terminal and a non-terminal node. The terminal nodes comprise independent values, while the non-terminal nodes depend on its components evaluation. The generation of a suitable computer program from a specific grammar would require a random selection of non-terminals. Regarding the GP Program, an initial population will be randomly generated before running the GP algorithm in a continuous loop. The loop will be characterized by two activities:

- Evaluation of individual program by the specifically designed fitness function.
- Generation of a new population of computer programs through selection of these individuals depending on their fitness and the use of the genetic operators like reproduction, mutation and crossover. The general program model is presented Fig. 1.

The initial generation of programs uses the genetic algorithm (GA) shown Fig. 2. Each of the iterations represents a generation. Basically, the GA is generic but various components would be incorporated to allow it to perform specific tasks. These additional aspects of the GA include the representation of chromosomes, selection strategy, and genetic operators (crossover and mutation). Thus, arrays of bits are used to represent chromosomes (computer programs), while simulation evolution – the

 ISBN: 978-1-61208-139-7

selection strategy – is used to process the individuals or rather the computer programs.

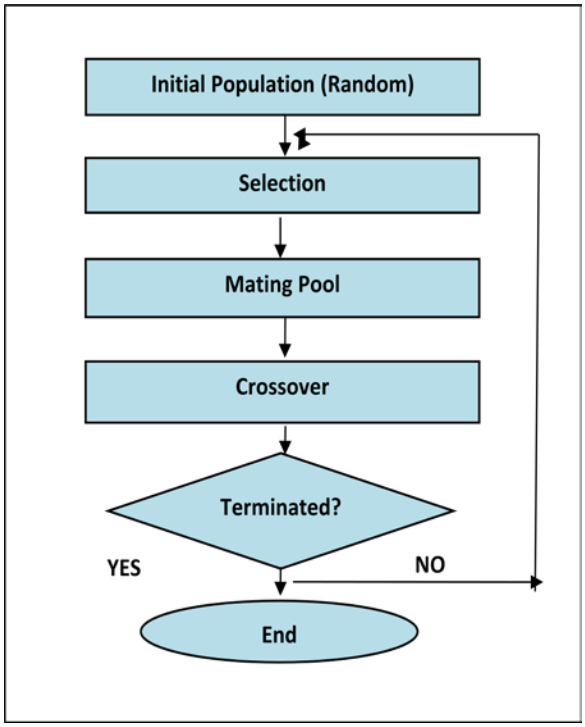


Figure 1. Our Genetic Program Model

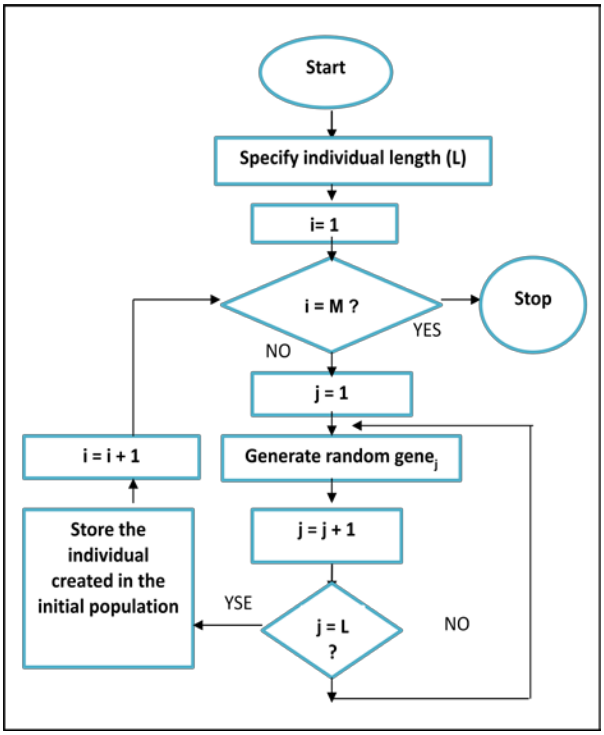


Figure 2. GP Selection Model (Algorithm)

B. Selection Strategy

Our GP model adopts a tournament selection approach so that various individuals are chosen randomly from a population for evaluation and the best individual in this group is designated as parents. This process is then repeated as often as there are individuals to choose. The selected parents are used in generation of new offsprings through uniform and random operation. The selection size will depend on the problem, population size, and so on. The parameter for selection is based on the population size. The tournament size ranges from two to a number equivalent to the number of individuals in the population. The selection would be done by selecting random indices - index1 and index2, where the model for both indices is characterized by classes 1 and 0 and having values that range from 1 to 100. This kind of selection is applied to minimize the selection pressure.

C. Fitness Functions

The fitness function is designed to direct the GP algorithm in selecting individuals (computer programs). The fitness functions do not change; they are non-mutable. A mutable fitness function will give different results and make comparison of the programs fitness difficult. The model demonstrates that through the use of various heuristic approaches with GP, the classification task in data mining can be simplified without compromising the integrity of the results. The systems is also useful in information retrieval (IR) applications such as document clustering; document matching and ranking, document indexing, query induction, optimization and representation. Fig. 3. shows the suggested fitness measurement function.

$$\text{Fitness} = (0.5*\text{acc}(0) + (.05*\text{acc}(1))$$

$$\text{Where accuracy for class } 0 = (f_{10}+f_{00})/(f_{10}+f_{00}+f_{01}+f_{11})$$

$$\text{And the accuracy for class } 1 = (f_{01}+f_{11})/(f_{10}+f_{00}+f_{01}+f_{11})$$

Figure 3. Used Fitness Function

The Fitness function is in fact a simple linear formula which designed specifically for the GP giving a result that will be either 0 or 1. Please notice that our reproduction selection act by copying the individual to next generation without making any changes to it. As shown Fig. 3, the different classes are measured by specific functions (f). The fitness function, is one of the most essential component in GP that measures how good an individual is as a solution to the problem. It could take several criteria into account and guide GP to seek the most preferable part of solution space

 ISBN: 978-1-61208-139-7

by probably tuning the weight of the criteria involved. Before setting up the fitness function for GP, some measures involved have been introduced. this is because the study mainly focuses on binary classification. Fig. 3 functions show some of measures and illustrate the number of instances belonging to four possible categories after classifying each of examples in a given dataset f_{10} indicates the number of the False negative. Negative (summation of instances that were predicted 1 but they were actually 0). f_{00} indicates the number of the True Negative (summation of instances that were predicted 0 and they were actually 0). f_{01} indicates to the False Positive. (summation of instances that were predicted 0 but they were actually 1). f_{11} indicates the True positive. (summation of instances that were predicted 1 and they were actually 1). This is applied when a program operates efficiently, or when its fitness function is very high [9].

D. Crossover

The crossover involves one or two programs (parents) being used to produce two fresh offspring for the subsequent generation. The first offspring are selected through the tournament method, whereby segments of these offsprings are selected randomly and exchanged to give two offspring for the next generation. The crossover operator permits the exchange of segments that have been produced using the same rule of evolution. The crossover operation would be executed 30 times; this number represents the maximum number of generations l. A total of 50 crossovers would be performed between the two models until the best fitness is obtained. This implies that the solution performs crossover between indexes of the models 1, 2, 3, 4, 5, 6 until 50 child models are created. The crossover for class 0s would also be created from the two models in order to determine the new class 0 by calculating the number of rules in both sets of the models. If 4 or 3 rules are obtained then random numbers of rules are chosen and used in the predictive class. The rules should be between 1 and 3 if the sum of the rules is equal to 3. The results of the 3 rules are mixed together in order to choose the best rule to be applied/used in the predictive class. The rules were selected based on fraction.

The initial population is first considered as a whole. Covering algorithm are used here to ensure a high level of accuracy. The basic idea about covering algorithm is to maximize instances of the desired class and minimize the instances of other classes. In our model, we start with a general rule of: *IF X then Y=0*. The idea here is to replace X with an exact condition. After that, the rule having the maximum accuracy will be selected. Then we delete instances that do not satisfy the condition of the selected rule. The same technique will be repeated on the remaining data until we reach a rule with almost 100% of accuracy.

E. Elitism

Elitism involves copying the best chromosomes to the initial population. This is beneficial in preventing the loss of chromosomes with good quality during the genetic operation. Thus, it ensures that the best programs of present generation survive the subsequent generation depending on the rate of elitism. It helped in maintaining subpopulations close to local and global optima. Through elitism the best chromosomes would be maintained. We ensure elitism by storing the best fitness in the next generation until achieving a better fitness. The proposed GP model will culminate into developing a data mining tool that will use the techniques discussed above to induce classifiers. The tool or program incorporates genetic operators, SQL grammar and a fitness function.

F. Analysis and Advantages

The proposed system has a number of advantages over various existing data mining systems including those which use GP. To start with, the advantages of GP over other techniques of knowledge discovery in databases are well documented. It is important to note that there are a number of approaches to knowledge discovery in electronic databases. These include description, clustering, association, regression, and classification. Recent studies have shown that classification is a suitable approach to knowledge discovery in large databases [6]. The development of genetic programming has been hailed as a milestone towards creation of superior techniques that are able to mine useful information from databases ([2] and [9]). Thus, genetic programming can be used to solve a number of real world business problems.

The suitability of the genetic programming and classification as data mining techniques is a clear advantage of the proposed system over other systems that do not use GP in knowledge discovery. The proposed GP model has a number of advantages. The GP design, for instance, can handle both numerical and string data types. The use of SQL makes the tool suitable in mining databases since programs evolved are represented in the SQL language. The GP approach also gives high accuracy of results. Furthermore, the proposed tool can include optimization metric such as area under roc curve (AUC) and Brier score (BRI) [6]. Although the system is general purpose and may be able to handle specific problems. The use of the SQL to represent the computer programs has its advantages. The SQL is a declarative language, this makes it easier to use tree structures. Furthermore, in comparison to systems that use linear GP, the system is preferable. In Linear Genetic Programming, individuals are represented as sequence of instructions which makes execution of the instructions time consuming.

 ISBN: 978-1-61208-139-7

IV IMPLEMENTATION- EXPERIMENTAL RESULTS

The experimental prototype has been developed using VB.NET 2008 and MS Access. The records were stored in a single table in MS access. The dataset contained information collected by the US Census Service concerning housing in the area of Boston Massachusetts. It was obtained from the StatLib archive: (http://lib.stat.cmu.edu/datasets/boston). The dataset has 506 cases. There are 14 attributes in each case of the dataset. Table 1 shows the used attributes CATMEDDEV is an attribute which has been created by categorizing median value (MEDV) into two categories. The median housing price in housing tracts in the Boston area falls into the "high" or "low" category. Using '1' for high values of MEDV and 0 for low values. As already mentioned the tool was created in VB.NET 2008 to read the data, generate the initial population, perform mutation and determines the best fit values. Fig. 5. shows the system interface. For seek of Cross Validation the Dataset have been divided into two subsets in order to train data and to test data, therefore, cross validation technique is used to evaluate the GP results. In the project, dataset was divided to 70 % for training and 30% for testing and evaluation. After generating the initial population, the execution is done 50 times (cycles). Each execution was done twice, one for 0s models and for the 1s models. The initial population was created in the following order:

- Randomly choose the number of conditions.
- Randomly operate between each condition (<><=>=).
- Randomly obtain a number between the minimum and maximum of the attributes

The initial population model is between 0 and 1. Each gene in the initial population is generated as a random number, distributed uniformly over the range [0,n] (the productions are numbered from 0 to n, so that any production can be represented by a number in the range [0,n]). In this method, all individuals (genotypes) in the initial population have a fixed length. Therefore the process of generating each individual is done using the following steps:

1 Each gene is generated randomly.
2 Repeat step 1 until the number of genes of that individual reaches the pre specified length.
3 If all individuals in the initial population are created, then continue to perform other GP operations, otherwise, repeat steps 1-3.

One iteration of the algorithm is referred to as generation. The basic GA is generic and there are several aspects that can be implemented differently according to the problem. This can be achieved by representation of solution or chromosomes, adopting appropriate selection strategy, applying crossover and mutation operators.

TABLE 1. DATABASE ATTRIBUTES

Attribute	Description
CRIM	Crime rate
ZIN	Proportion of residential land zoned for lots over 25,000 sq.ft.
INDUS	Proportion of non-retail business acres per town.
CHAS	Charles River dummy variable (1 if tract bounds river; 0 otherwise)
NOX	Nitric oxides concentration (parts per 10 million)
RM	Average number of rooms per dwelling
AGE	Proportion of owner-occupied units built prior to 1940
DIS	Weighted distances to five Boston employment centres
RAD	Index of accessibility to radial highways
TAX	Full-value property-tax rate per $10,000
PTRATIO	Pupil-teacher ratio by town
B	1000(Bk - 0.63)^2 where Bk is the proportion of blacks by town
LSTAT	% lower status of the population
MEDV	Median value of owner-occupied homes in $1000
CATMED DEV	The predicted classes

GA is implemented by having arrays of bits that represent chromosomes. The individuals in the population are then processed by simulation evolution. Simple bit manipulation operators are then used to allow the implementation of crossover, mutation and other operations. For example, to generate the initial population for a random number 3 for 0 models. In the first round, this is chosen randomly between the operations of each condition. The result will be: Med<.07 and max<=8.9 and Dix<8.88 Where the numbers 0.07, 8.9 and 8.88 are between the minimum and maximum values of the attributes. The values for model 1 are generated in the same way. The results of the experiments involving the GP-SQL Mining Tool and the database containing the Boston housing information reveal the effectiveness of the system. To start with, the system

 ISBN: 978-1-61208-139-7

was able to apply the genetic operators effectively to evolve programs and select the most suitable rule for classification. Thus, it was able to reproduce programs without any changes for the next generation, perform crossover mutation on programs. Furthermore, it allowed diversity in the population of programs to prevent an early derivation of algorithm solution. These genetic operators are also applied in LOGENPRO, Neural Net, and GPSQL Miner, which showed impressive results or accuracy in the experiments. In this regard, it is substantial, therefore, to comment that the GP is a suitable approach to classification or induction of classifiers in data mining.

The effectiveness of the system was tested against a large database of Boston housing data. The results showed that the system is promising in regards to its application in classification of data in the data mining process. It generated reliable results that are highly accurate when a large database that has a lot of noise was used. The system was further validated against other common GP-based approaches for data mining classification including GP-Knn and GPSQL Miner. The level of accuracy for the system is higher than that achieved by both the GPSQL Miner and GP-Knn. Furthermore, the literature revealed that the various methods integrated into the system are beneficial in maintain the best chromosomes, reducing the search space, and generating the most suitable program for classification. The system is advantageous in that it uses an interface that is user friendly, and at the same time maintains the advantages of using the SQL grammar in manipulating the database. However, the use of a high-level programming language increased the execution time needed compared to a system that is based on a low-level language. The accuracy of classifiers is shown in Table 2. Our system (labeled GP-SQL Mining Tool) shows higher accuracy than the other two systems; it is better than the least accurate of the three algorithms (LOGENPRO) by more than 5%. Thus, the GP-SQL Mining Tool gives better results, and therefore, is reliable and suitable for induction of classifiers in large databases.

TABLE .2 THE ACCURACY OF CLASSIFICATION BY VARIOUS GP SYSTEMS

System	Accuracy
LOGENPRO	91.04%
GPSQL Miner	95.04%
Neural Net	96.7%
GP-SQL Mining Tool	**96.8%**

The experimental set of rules results in one run through GP-SQL Miner tool in this dataset is showed Fig. 4. The results of the data mining in the Boston housing database reveal the effectiveness of the proposed GP system. To start with, the system was able to apply the genetic operators; thus it was able to replicate programs without changes for the next generation (reproduction), select any two parent programs to generate new offspring (crossover) and randomly select non-terminal points from parent programs. Furthermore, it allowed diversity in the population of programs to prevent an early derivation of algorithm solution. The results of the experiment show that the tool provided excellent results. In fact, the fitness of 96 % is highly accurate.

V CONCLUSION AND FUTURE WORK

In this paper, we have explored the idea of combining data mining and genetic programming. Thus, data mining is a process of retrieving useful information from electronic databases, or rather; it is a process of knowledge discovery. This study was focused on highlighting the benefits that genetic programming can add in data mining. It did not examine the GP application in classification in a real work situation, which is vital if the tool needs to be applied in commercial settings. The achieved tests showed interesting and encouraging results in for data classification with a high efficiency rate. A subset of the obtained results is shown Fig. 4. On the other hand, GP is a technique that employs the evolutionary concepts to generate computer programs for use in classification task. Data mining involves a number of tasks, but more emphasis has been put on the classification task. Genetic programming is a based on genetic algorithms that use evolutionary concepts in the data mining process. It involves genetics operators such reproduction, mutation, and crossover. The technique has developed from the simple linear-based GP to cellular type of GP. We proposed a GP-based system that uses a heuristic approach in the induction of classifiers.

Figure 4. Set of results (predicted Models)

A number of literatures were examined and studied. It is shown that GP techniques are more superior to traditional classification techniques in data mining. It is possible to compare same type attributes within a dataset to generate a reduced number of effective rules. Several GP-based techniques have been proposed. These range from a concept mapping technique that evaluates individual fitness through booleanized attributes and token completion, cellular GP, linear GP, to GPSQL Miner tool that uses the SQL grammar for classification. A variant GP approach uses the C-programming language in representation of genetic programs. Our system utilizes the SQL grammar; this is significant as it makes it easy to evaluate data stored in the database. In addition, the design employs heuristic measures such as cross validation, fitness function, elitism, and genetic operators like reproduction, and cross over to produce more favorable results.

In the reproduction process, individuals are copied exactly as they appear, while in mutation and crossover there changes in the offspring program. However, crossover involves exchange of properties between different programs, while mutation exchange can happen within the program. The fitness function was important in determining the accuracy of the system. Elitism prevents the loss of chromosomes with good quality. The proposed system validated the assumption that GP is an effective technique for data mining classification. It was designed to reduce the search space considerably making it easy to induce classifiers and classify the data. A tool was developed through VB.NET programming language, and an Ms Access relational database was used to store data. The results showed that the system is promising in regards to its application in classification of data in the data mining process It generated reliable results that are highly accurate when a large database having a lot of noise was used. The system was further validated against other common GP-based approaches for data mining classification including GP-Knn and GPSQL Miner. The level of accuracy for the system is higher than that achieved by both the GPSQL Miner and GP-Knn. Furthermore, the literature revealed that the various methods integrated into the system are beneficial in maintain the best chromosomes, reducing the search space, and generating the most suitable program for classification. The system maintains the advantages of using the SQL grammar in manipulating the database. However, the use of a high-level programming language increased the execution time when compared to a system that is based on a low-level language. It is recommended to conduct further tests in a commercial setup. This study was focused on highlighting the benefits that genetic programming can add in data mining. Therefore, future works will involve investigating the effectiveness of the system and such a GP approach on data mining classification on more databases, both small and large, and which comprise real data.

REFERENCES

[1] W. Banzhaf, P. Nordin, R. Keller, and F. Francone, "Genetic Programming: An Introduction: On the Automatic Evolution of Computer Programs and Its Applications", Morgan Kaufmann, 1998.

[2] S. Bhattachariya, O. Pictet and G. Zumbach, "Knowledge-intensive genetic discovery in foreign exchange markets". IEEE Transaction on evolutionary Computation, Vol 6 (2), April 2002, pp. 169-181.

[3] E. Carreno, G. Leguizamon, and N. Wagner, "Evolution of classification rules for comprehensible knowledge discovery." In proceedings of Evolutionary Computation Congress CEC 2007, pp. 1261-1268.

[4] P. Espejo, S. Ventura and F. Herrera, "A survey on the application of genetic programming to classification." IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews, Vol 40(2), March 2010. pp. 121-144.

[5] G. Folino, C. Pizzuti and G. Spezzano, "GP ensembles for large-scale data classification." IEEE Transaction on Evolutionary Computation, Vol 10(5), Oct 2006, pp. 604–616.

[6] Y. Ishida, and A. Pozo, "GPSQL Miner: SQL-grammar genetic programming in data mining." CEC'02 Proceedings of the Evolutionary Computation Vol 02, May 2002, pp. 1226–1231.

[7] S. Muharram and G. Smith, "Evolutionary constructive induction". IEEE Transactions on Knowledge and Data Engineering, Vol 17(11), Nov 2005, pp. 1518-1528.

[8] S. Sakprasat, and M. Sinclair, "Classification rule mining for automatic credit approval using genetic programming." In proceedings of Evolutionary Computation Congress CEC 2007, Sept 2007, pp. 548-555.

[9] M. Saraee, and R. Sadjay, "Optimizing classification techniques using genetic programming approach." In proceedings of INMIC 2008, Dec 2008, pp. 345-348.

[10] P. Tan, M. Steinbach, and V. Kumar, "Introduction to Data Mining." Addison Wesley Editions, 2006.

[11] H. Tan and E. Frank, "Data mining: Practical Machine Learning Tools and Techniques" Second Edition, Morgan Kaufman, 2005.

 ISBN: 978-1-61208-139-7

Significance of Low Frequent Words in Patent Classification

Akmal Saeed Khattak
Natural Language Processing
Department of Computer Science
University of Leipzig
Leipzig, Germany
akhattak@ informatik.uni-leipzig.de

Gerhard Heyer
Natural Language Processing
Department of Computer Science
University of Leipzig
Leipzig, Germany
heyer@ informatik.uni-leipzig.de

***Abstract*—Low frequent terms are often considered noise but in case of patent documents it might refer to technical terms. This paper shows the significance of low frequent terms in patent classification. Our experiments show that low frequent terms cannot be ignored in patents as it give better performance in terms of f-measure and accuracy than high frequent terms. Experiments are shown to prove that set of low frequent terms outperforms set of high terms in classifying patent documents.**

Keywords - patent classification; text classification; taxonomy; International Patent Classification (IPC)

I. INTRODUCTION

The process of assignment of one or more predefined classes to text documents automatically is called text classification or categorization. There are many applications of text classification like organizing web pages into hierarchical categories, indexing journal articles by subject categories (e.g., the Library of Congress, MEDLINE, etc.), responding to Census Bureau occupations, filtering email messages, tracking news events and filtering by topics, archiving patents using International Patent Classification (IPC). Patent Classification or Categorization is one of the application area of text classification. Text classification approaches for patent classification problems have to manage simultaneously very large size of hierarchy, large documents, huge feature set and multi-labeled documents [1]. IPC is a standard taxonomy developed and maintained by World Intellectual Property Organization (WIPO). The IPC consists of about 80,000 categories that cover the whole range of industrial technologies [1]. There are 8 sections at the highest level of the hierarchy, then 128 classes, 648 subclasses, about 7200 main groups, and about 72000 subgroups at lower levels [1]. The top four levels from the 80000 classes are mostly used in automatic patent classification systems [1]. The IPC is a complex hierarchical system, with layers of increasing detail. For example, Section: G Physics, Class: G02 Optics, Subclass: G02C Spectacles, sunglasses or goggles ..., Main group: G02C5 Construction of non-optical parts.

Patent classification is a kind of knowledge management where documents are assigned predefined categories. Patent collections consist of huge vocabulary and this large vocabulary reduces the classification performance in terms of accuracy. The reason for low accuracy of classifier is due to inclusion of noisy words that is needed to be differentiate from dominant words. We reduce the vocabulary size by considering only frequent terms that have frequency above than a threshold based on some document frequency of that those terms in the entire collection. In experiments, it was found that low frequent terms can efficiently figure out dominant terms and due to inclusion of low frequent terms the classification accuracy is increased.

The remainder of this paper is structured as follows. Section II discusses related work in the field of text classification and its application patent classification. Section III gives a methodology consisting of previous algorithms to classify patents. Section IV gives analysis and experiment results. In experiment section, we discuss results on two datasets. Finally Section V is about the key lessons learned and some direction for future work for further exploration.

II. RELATED WORK

Sebastiani [2][3] has written an excellent survey on machine learning methods for text categorization and various challenges in it. Ceci and Malerba [4] investigated the issues regarding representation of documents and also the learning process. Dumais and Chen [5] explores the use of hierarchies to classify a large collection of web content. A number of statistical classifications and machine learning techniques have been applied to text categorization, including nearest

 ISBN: 978-1-61208-139-7

neighbor classifiers [6][7], Centroid-Based Classifier [8], Naive Bayes (NB) [9], Decision Trees [10] and Support Vector Machines (SVM) [11]. These machine learning techniques can be applied to patents as patent is a text document. Larkey [13] developed a classification tool based on a k-Nearest Neighbors (k-NN) approach. Chakrabarti, Dom and Indyk [14] developed a hierarchical patent classification system using 12 subclasses organized in three levels. Krier and Zaccà [15] discussed a comprehensive set of patent classification experiments. These authors organized a comparative study of various classifiers but the detailed results are not disclosed [12]. Fall, Torcsvari, Benzineb and Karetka [12] showed that instead of using full texts, the first 300 words from the abstract, claims, and description sections gives better performance regardless of classifiers.

III. METHODOLOGY

The documents are stored in many kinds of machine readable form such as PDF, DOC, Post Script, HTML, XML. The content of documents is transformed into a compact representation. Representation of text influence the classifier in achieving better performance. Text classification consists of 3 phases: text representation, building classifier model, testing classifier (evaluation). Vector Space Model (VSM) is a common way to represent document in a vector of terms [17]. Once documents are represented as a vector of terms, terms are weighted across the document collection using weighting schemes. Table 1 shows three weighting schemes TFIDF (Term Frequency Inverse Document Frequency), BM25 (Best Match) and SMART (System for Manipulating and Retrieving Text) formulas. The formulas for these weighting schemes are given in Table 1. After the assignment of weights to terms, classifiers are build on training set and using this model data is tested from the testing set. The four classifiers used are NB, SVM, Decision Trees, KNN (for k=1 and 3). The naïve Bayesian classifier is a statistical classifier [22]. Bayes' theorem is the basis for Bayesian classification [22]. The basic idea in a Naive Bayesian Classifier is the assumption that the effect of an attribute value on a given class is independent of the other values of other attributes [22]. SVM is a state-of-the-art machine learning method developed by V.Vapnik et. al. [21] is well suited for text classification [11]. The reason that SVMs work well for text classification is the huge dimensional input space, and document vectors sparsity [11]. Decision tree does not require any knowledge [23]. Given a training data a decision tree can be induced. From decision tree rules are created about the data and using these rules documents in testing set are classified [23]. Another type of classifier is an instance based classifier called K-nearest neighbor or KNN. KNN can be applied to many fields of data mining. KNN is a supervised learning algorithm. The similarity between all documents of testing and training set is computed. For each document in testing set K nearest neighbors in training documents are considered and the class is assigned based on the majority of K nearest neighbors [24]. The last step in text classification is evaluation. Using the contingency table 2, the classifiers are evaluated using the measures shown in Table 3.

TABLE 1 Different Term Weighting Schemes

Term Weighting	Formula
$w_{ij}=tf_{ij}.\log\frac{N}{n_i}$	TFIDF [18]
$w_{ij}=\frac{(k+1).tf_{ij}}{k(1-b)+b.\frac{doc_{len}}{avg_{doclen}}+tf_{ij}}.log\frac{N-df_j+0.5}{df_j+0.5}$	BM25 [19]
$w_{ij}=\frac{(1+\log(tf_{ij}))}{(1+\log(avgtf_{ij}))}.\frac{1}{\left(0.8+0.2.\frac{doc_{len}}{avg_{doclen}}\right)}.\log\frac{N}{df_j}$	SMART [20]

TABLE 2 Contingency Table

		Predicted	
		negative	positive
Actual examples	negative	a	b
	positive	c	d

TABLE 3 Evaluation Measure

Evaluation	Formula
Accuracy [2]	$A=\frac{(a+d)}{(a+b+c+d)}$
Precision [2]	$P=\frac{d}{(b+d)}$
Recall [2]	$R=\frac{d}{(c+d)}$
F-measure [2]	$F=\frac{(2.P.R)}{(P+R)}$

IV. DATASET AND EXPERIMENTAL RESULTS

Dataset-1: The main focus of these experiments are to explore the impact of low frequent terms on patent

classification in comparison with frequent terms irrespective of the hierarchical structure of patents. Only main group label of documents are considered ignoring rest of the labels (subclass, class, section). First dataset was downloaded from http://www.freepatentsonline.com [27]. For labels of documents only main group classes are considered. Documents were in HTML form. Documents contain several sections like Title, Document Type and Number, Abstract, Inventors, Application Number, Publication Date, International Classes, Claims, Description and some others. First of all, the set of patent documents (both training and testing) are preprocessed. All HTML tags are removed and hence converted to plain text. Only text under claim section of patent documents are considered here. The plain text is then preprocessed. Preprocessing extract content word. All case words are treated as small. An algorithm for suffix stripping is applied in order to perform stemming [29]. In literature, it can be found that stemming is not useful in terms of accuracy but it is useful in reducing the dimensions of text. All words that are less than or equal to 4 characters are also removed. All stop words are removed. After preprocessing a set of unique 4351 terms (word type) is obtained. Now the preprocessed text is represented in a representation model. Experiments are performed on 1484 documents. The train / test split is 66 / 34 %. Experiments are made on 4 classifiers (naïve Bayesian, support vector machine, j48, k nearest neighbor) using four weighting schemes (tfidf, bm25, smart). Following are some threshold on terms selection to investigate the effect of terms (both low and high frequent) on patent classification:

I. that occur in more than 10 document and less than 101 documents (low frequent terms)
II. that occur in more than 100 documents and less than 201 documents (frequent terms)
III. that occur in more than 200 documents (high frequent terms)

All the experiments on this dataset was carried out using WEKA [25]. The main focus was to investigate the significance of low frequent terms in comparison with frequent terms. Low frequent terms contribute more in getting better classification accuracy than frequent terms. It can be seen from Table 4 and Figure 1. Table 4 shows that in 11 out of 15 cases f-measure of classification using low frequent terms terms give better f-measure than frequent terms. This fact can be seen in Figure. 1. NB, SMO, J48 and KNN (for K=1 and K=3) classifiers when used with TFIDF weighting scheme gives better performance in terms of f-measure when terms that occur in more than 10 and less than 101 (terms that satisfies threshold I) documents are considered than terms comes under the criteria of II (terms that occur in more than 100 and less than 201 documents) and III (terms that occur in more than 200 documents) are considered.

NB, SMO and J48 classifiers when used with BM25 weighting schemes give better results for low frequent terms (I) than for high frequent terms (both II and III). The exception where frequent terms perform better than low frequent terms is when KNN (for both K=1 and 3) with BM25 is used. Similarly NB, SMO and J48 classifiers used with SMART weighting performs better in case of low frequent terms than high frequent terms. The only exception where frequent terms perform better than low frequent terms is when KNN (for K =1 and 3) is used with SMART as shown in Table 4. Only short documents in the downloaded documents are considered to make a dataset of 1484 documents. The reason behind this is WEKA is not much scalable. It just cannot classify a dataset more than 2000 documents. It got stuck in it. A patent collection can be made of both large and short documents. Thats why the LIBSVM [26] library was used in octave [28] to classify 4238 documents consisting both long and short documents

TABLE 4 F-measure on different classifiers using TFIDF, BM25 and SMART weighting schemes

	Classifier + WS	**I**	**II**	**III**
1	NB+TFIDF	**0,3730**	0,2500	0,2110
2	SMO+TFIDF	**0,3180**	0,2540	0,2540
3	J48+TFIDF	**0,3940**	0,3120	0,2640
4	KNN-1+TFIDF	**0,2640**	0,2540	0,2420
5	KNN-3+TFIDF	**0,2280**	0,2260	0,2270
6	NB+BM25	**0,3960**	0,2850	0,2610
7	SMO+BM25	**0,4510**	0,3790	0,3840
8	J48+BM25	**0,3930**	0,2730	0,2640
9	KNN-1+BM25	0,2730	0,2730	**0,2940**
10	KNN-3+BM25	0,2030	0,2550	**0,2770**
11	NB+SMART	**0,4040**	0,2850	0,2720
12	SMO+SMART	**0,4630**	0,3930	0,3250
13	J48+SMART	**0,3750**	0,3120	0,2470
14	KNN-1+SMART	0,2010	**0,2750**	0,2550
15	KNN-3+SMART	0,1690	0,2460	**0,2610**
Number of Terms		847	110	85

 ISBN: 978-1-61208-139-7

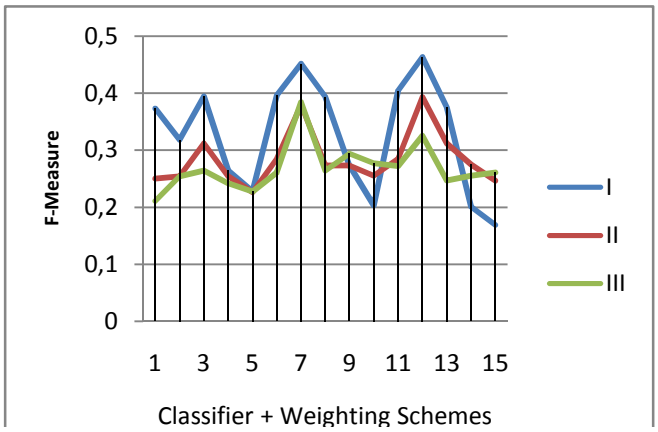


Figure 1. F-measure of different sets using classifiers and weighting schemes shown in Table 4.

Dataset-2: The other dataset is extracted from a benchmark dataset of TREC chemical patents. The total number of documents on which experimentations are made is 4238. Text were extracted for 21 main group classes. Each class have either 201 or 202 documents. Different datasets beside complete patent document consisting of various fields (title, abstract, claims, background summary, description) of patents were created. Table 5. shows word tokens and word types in each field of patent document collection on which a set of experiments were carried out. All words that are less than or equal to 4 characters are removed. All stop words are removed and stemming is performed.

LIBSVM library is used in octave to classify patent documents. 10 fold cross validation is used. The kernel type used in experiments here is linear. There are four variations of kernel in LIBSVM named as linear, polynomial, radial and sigmoid. It can be proved by experimenting that linear kernel type give better results. Two term sets or feature sets (Low Frequent Term Set LFTS and High Frequent Term Set HFTS) based on a document frequency threshold given below are created for each field and the complete patent text. The threshold criteria is as follows:

I. Terms that occur in more than 10 and less than 101 documents are considered as LFTS because it occurs between 0.24% and 2.4% documents in the entire collection.
II. Terms that occur in more than 500 and less than 1001 documents are considered as HFTS because it occurs between 12% and 24% documents in the entire collection.

The focus was to investigate the performance of low and high frequent terms and see which one gives better accuracy. It can be seen from Figure 2 and Table 6. that LFTS show better results for each field text using TFIDF, BM25 and SMART weighting scheme. Table 6. shows the performance of classifier in terms of accuracy using low and high frequent terms set on different fields of patents using TFIDF, BM25 and SMART. In Fig. 2, the blue line represents LFTS and the red line represents HFTS. Clearly it can be seen that LFTS outperforms HFTS in all cases listed in Table 6. The classification in case of title, abstract, claims, background summary, description and complete patent performs better for low frequent terms as compared to high frequent terms and gives 4.55%, 5.68%, 6.44%, 6.98%, 11.42% and 12.71% respectively better results when used with TFIDF. Similarly using LFTS with BM25 weighting scheme gives 9.49, 17.97, 4.74, 0.71, 1.35 and 3.21 percent better accuracy on all fields of title, abstract, claims, background summary, description and complete patent respectively than HFTS with BM25. Same is the case when SMART weighting scheme is used with all these fields. LFTS combined with SMART gives 9.31, 3.33, 5.19, 3.8, 3.13 and 2.48 percent better results as compared to HFTS.

TABLE 5 Word Tokens and Types in different section of patents

Field of Patent Document	Word Tokens	Word Types
Title	19717	4027
Abstract	156035	9700
Claims	761773	18488
Background Summary	2185892 (around 2.2 million)	45709
Description	5151686(around 5.2 million)	83738
All Patent Document	8283579 (around 8.3 million)	106045

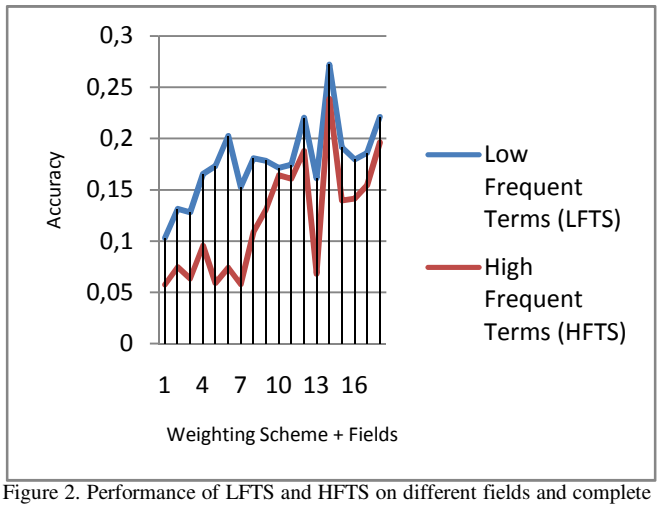


Figure 2. Performance of LFTS and HFTS on different fields and complete patents.

TABLE 6 Accuracy of Classifier on different fields (Title, Abstract, Claims, Background Summary and Description) and complete patents using TFIDF,BM25 and SMART

	Field + Weighting Scheme	Low Frequent Terms (LFTS)	High Frequent Terms (HFTS)
1	title + TFIDF	**0,1031**	0,0576
2	abs + TFIDF	**0,1314**	0,0746
3	claims + TFIDF	**0,1279**	0,0635
4	background summary + TFIDF	**0,1654**	0,0956
5	description + TFIDF	**0,1734**	0,0592
6	all + TFIDF	**0,2025**	0,0741
7	title + BM25	**0,1527**	0,0578
8	abs + BM25	**0,1808**	0,1090
9	claims + BM25	**0,1784**	0,1310
10	background summary + BM25	**0,1713**	0,1642
11	description + BM25	**0,1744**	0,1609
12	all + BM25	**0,2202**	0,1881
13	title + SMART	**0,1612**	0,0681
14	abs + SMART	**0,2721**	0,2388
15	claims + SMART	**0,1914**	0,1395
16	background summary + SMART	**0,1796**	0,1416
17	description + SMART	**0,1859**	0,1546
18	all + SMART	**0,2211**	0,1963

V. CONCLUSION AND FUTURE WORK

The main focus was to investigate the significance of low frequent terms in patent classification. Experiments above show that low frequent terms gives better performance in terms of f-measure and accuracy as compared to high frequent terms. Low frequent terms are potential discriminant terms and in patents it might refer to technical terms and might be very specific term. By selecting specific terms the classification of patents can be improved. So low frequent terms cannot be ignored as noise. In future, other threshold method like information gain and chi-square will be used for term selection and compare with the threshold based on document frequency used in this paper. The future work is to marginalize noise in patent documents to improve patent classification at different levels of IPC hierarchy specially at the main group level (higher level of details) where specific terms can improve patent classification. We also plan to investigate consider term proximity (closeness) within a document that might increase the performance of patent classification.

REFERENCES

[1] D. Tikk, G. Biró, and A. Törcsvári, "Experiment with a hierarchical text categorization method on WIPO patent collections", Applied Research in Uncertainty Modelling and Analysis, International Series in Intelligent Technologies, Volume 20, pp. 283-302, 2005.

[2] F. Sebastiani, "Machine learning in automated text categorization", in ACM Computing Surveys archive Volume 34 , Issue 1, pp. 1 – 47, 2002.

[3] F. Sebastiani, "Text Categorization", in A. Zanasi (ed.), Text Mining and its Applications to Intelligence, CRM and Knowledge Management, pp. 109-129, WIT Press, Southampton, UK, 2005.

[4] M. Ceci and D. Malerba, "Classifying web documents in a hierarchy of categories: a comprehensive study", Journal of Intelligent Information Systems Volume 28, Issue 1, pp. 37 – 78, 2007.

[5] S. Dumais and Chen, " Hierarchical classification of web content", in Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval, pp. 256– 263, New York: ACM, 2000.

[6] Y. Yand, "An evaluation of statistical approaches to text categorization", Information Retrieval, 1(1-2), 69-90, 1999.

[7] Y. Yang and X. Lin, "A re-examination of text categorization methods", In The 22nd Annual International ACM SIGIR Conference on Research and Development in the Information Retrieval, New York: ACM Press, 1999.

[8] E. Han and G. Karypis, "Centroid-based document classification analysis and experimental resultshttp://www.cs.umn.edu/wkarypis", 2000.

[9] D. Lewis, "Naive (Bayes) at forty: the independence assumption in information retrieval", In The 10th European Conference on Machine Learning, pp. 4–15, New York: Springer, (1998).

[10] D. Lewis and M. Ringuette, "Comparison of two learning algorithms for text categorization", In Proceedings of the Third Annual Symposium on Document Analysis and Information Retrieval (SDAIR'94), 1994.

[11] T. Joachims, "Text categorization with support vector machines: learning with many relevant features", In The 10th European Conference on Machine Learning, pp. 137–142, New York: Springer, 1998.

 ISBN: 978-1-61208-139-7

[12] C. J. Fall, A. Torcsvari, K. Benzineb, and G. Karetka, "Automated categorization in the international patent classification. ACM SIGIR Forum, 37(1), pp. 10–25, 2003.

[13] L. S. Larkey, "A patent search and classification system", In Proceedings the 4th ACM conference on digital libraries, pp. 179–187, 1999.

[14] S. Chakrabarti, B. Dom, and P. Indyk, "Enhanced hypertext categorization using hyperlinks", Proc. SIGMOD98, ACM International Conference on Management of Data, ACM Press, New York, pp. 307- 318, 1998.

[15] M. Krier and F. Zaccà, "Automatic categorization applications at the European Patent Office", World Patent Information 24, pp. 187-196, 2002.

[16] C. J. Fall , K. Benzineb , J. Guyot , A. Törcsvári, and P. Fiévet , "Computer-Assisted Categorization of Patent Documents in the International Patent Classification", In Proceedings of the International Chemical Information Conference, Nîmes, October 2003 (ICIC'03).

[17] G. Salton, A. Wong, and C. S. Yang, "A vector space model for information retrieval", Communications of the ACM, 18(11), pp. 613–620, November 1975.

[18] G. Salton, and C. Buckley, "Term-weighting approaches in automatic text retrieval", Inform. Process. Man. 24, 5, 513–523, Also reprinted in Sparck Jones and Willett [1997], pp. 323–328, 1988.

[19] S. E. Robertson, S. Walker, M. Hancock-Beaulieu, A.Gull, and M. Lau, "Okapi at TREC-3", In Harman, D. K. (ed.) The Third Text Retrieval Conference (TREC-3) NIST, 1995.

[20] Y. H. Tseng , C. J. Lin, and Y. Lin, "Text mining techniques for patent analysis", Information Processing and Management 43, pp. 1216–1247, 2007.

[21] V. Vapnik, "The Nature of Statistical Learning Theory", Springer, New York, 1995.

[22] J. Han and M. Kamber, "Data Mining: Concepts and Techniques", Elsevier, 2006.

[23] Teknomo and Kardi, "Tutorial on Decision Tree, http://people.revoledu.com/kardi/tutorial/DecisionTree, last accessed on 20.04.2011.

[24] Teknomo and Kardi, "K-Nearest Neighbors Tutorial. http://people.revoledu.com/kardi/tutorial/KNN", last accessed on 20.04.2011.

[25] I. H. Witten and E. Frank, "Data Mining: Practical machine learning tools and techniques", 2nd Edition, Morgan Kaufmann, San Francisco, 2005.

[26] C. C. Chang and C. J. Lin, LIBSVM : a library for support vector machines, 2001. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm, last accessed on 20.04.2011.

[27] http://www.freepatentsonline.com, last accessed on 20.04.2011.

[28] http://www.octave.org, last accessed on 20.04.2011.

[29] http://snowball.tartarus.org/download.php, last accessed on 20.04.2011.

 ISBN: 978-1-61208-139-7

Mining Association Rules Inside a Relational Database – A Case Study

Mirela Danubianu, Stefan Gheorghe Pentiuc
Faculty of Electrical Engineering and Computer Science
"Stefan cel Mare" University of Suceava
Suceava, Romania
mdanub@eed.usv.ro, pentiuc@eed.usv.ro

Iolanda Tobolcea
Faculty of Psychology and Educational Sciences
"A.I.Cuza" University of Iasi
Iasi, Romania
itobolcea@yahoo.com

Abstract - **In the context of the necessity to find new knowledge in data, last decade, data mining has become an area of great interest. Although most data mining systems work with data stored in flat files, sometimes it is beneficial to implement data mining algorithms within a DBMS, in order to use SQL or other facilities provided, to discover patterns in data. In this paper we consider a way to discover association rules from data stored into a relational database. We make also a comparative study of performances obtained by applying the following methods: stored procedures in database or candidate and frequent itemsets generated in SQL using a k-way join and a subquery-based algorithm. This study is used to choose the best solution to implement in the particular case of building a dedicated data mining system for personalized therapy of speech disorders optimization.**

Keywords-data mining method, association rules, relational database, SQL , stored-procedures

I. INTRODUCTION

The most common way to store data collected in various areas is in relational databases. Information and Communication Technology development has lead to a huge volume of data stored and to the inability to extract useful information and knowledge from this data by using the traditional methods. For this reason, data mining has developed as a specific field. Mining association rules is one of the commonly used methods in data mining. Association rules model dependencies between items in transactional data. Most data mining systems work with data stored in flat files. However, it has been shown it is beneficial to implement data mining algorithms within a DBMS, and using of SQL to discover patterns in data can bring certain advantages.

There is some research focused on issues regarding the integration of data mining with databases. There have been proposed language extensions of SQL to support mining operations. For instance, in [1], DMQL extends SQL with a series of operators for generation of characteristic rules, discriminant rules and classification rules.

This paper aims to present some aspects of coupling data mining algorithms with database management systems.

Section II shows the reasons to try to mine data directly in databases and a possible architecture for such data mining system. In Section III, there are defined association rules, presents the methods and some algorithms that find association rules in data. In Section IV, we focus on the possibility to use SQL in order to generate the frequent itemsets. Section V contains a case study.

II. PERFORMING DATA MINING INTO A DATABASE MANAGEMENT SYSTEM

As we have noted above, sometimes it is useful to implement data mining algorithms in database management systems (DBMS).

First, one can use the database indexing capability and query processing and optimization facilities provided by a DBMS. Second, for long running mining applications it can be useful the support for checkpointing and last, but not least one can exploit the possibility of SQL parallelization, especially for a SMP environment.

This involves the development of data mining applications tightly–coupled with a relational database management system. In [2], a methodology to achieve this goal is presented.

Accordingly to this methodology, the records of a database are not fetched into the application, but modules of the application program that perform various operations on the retrieved records, are pushed in the database system.

Although in this case the core of data mining is found in the database, no changes were made on database management software, and appropriate functionalities were provided by user-defined procedures and functions stored in the database.

A possible architecture for such data mining approach is presented in Figure 1.

The graphical interface allows users to formulate the data mining problem and to establish parameters, such as: minimum thresholds for support and confidence for association rules, or minimum value for accuracy for classification.

Preprocessing module aims to translate the mining problem in the corresponding SQL instruction set. We consider the Oracle 9i SQL dialect because it contains object relational capabilities and it allows user-defined function and functions table.

Finally, the processing results are converted and presented to the user in an intelligible form through the graphical interface.

 ISBN: 978-1-61208-139-7

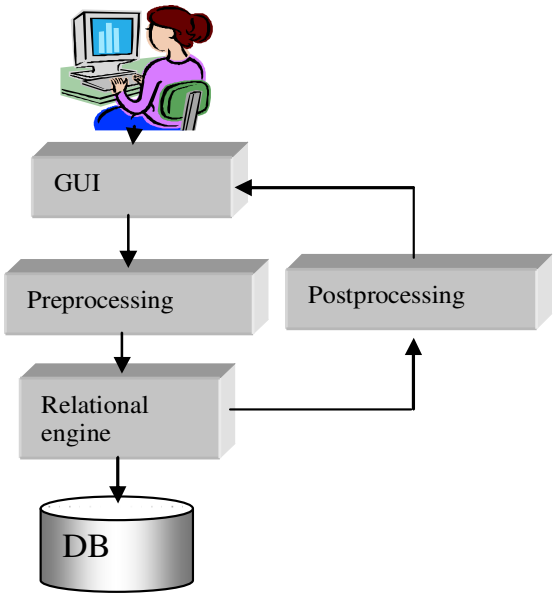


Figure 1. A proposed architecture for data mining in a DBMS

III. ASSOCIATION RULES: DEFINITION, METHODOLOGY, ALGORITHMS

Association rules aims to discover the dependencies between the items in transactional databases.

Briefly, one can define association rules as follows: if we have a set of transactions, where each transaction is a set of items, an association rule [3] is an implication A->B, where A and B are disjoint sets of items. It means that if a transaction contains the items in A, it tends also to contain the items in B. A is called the antecedent of rule and B is called its consequent.

In terms of association rules, a set of k items is called *k-itemset*. In a transactional database, for an itemset A we can define a measure called support, that represents the fraction of transaction that contains A. If we note the database with D, the expression for the support for A is:

$$sup(A) = |A| / |D| \quad (1)$$

where |A| is the number of transaction containing A, and |D| is the cardinality of the database.

The support for the rule A-> B is defined as:

$$sup(A \rightarrow B) = sup(A \cup B) \quad (2)$$

and represent the percentage of all transactions that contain both A and B.

The rule holds in the transaction database with confidence calculated with the following expression:

$$conf(A \rightarrow B) = \frac{sup(A \cup B)}{sup(A)} \quad (3)$$

The problem of finding association rules is to generate all association rules that have support and confidence greater than two user–specified thresholds called *minsupp* and *minconf*.

The mining process for association rules can be decomposed in two phases: first find all combination of items whose support is greater than *minsupp*, called frequent itemsets and second, use these frequent itemsets to generate the rules.

Since the generation of frequent itemsets is the most expensive part in terms of resources and time consuming, a lot of algorithms for this task were developed. Most algorithms use a method that build candidate itemsets, which are sets of potential frequent itemsets, and then test them. Support for these candidates is determined by taking into account the whole database D. The process of generating candidate itemsets considers the information regarding the frequence of all candidates already checked. So, the procedure is the following: the closure of frequent itemsets assumes that all subsets of a frequent itemset are also frequent. This allows remove those sets that contain at least one set of items that is not frequent, from candidate itemsets. After generating, the appearance of each candidate in the database is counted, in order to retain only those having the support greater than *minsup*.

Then we can move to the next iteration. The whole process ends when there are no potential frequent itemsets.

The most known algorithm, which uses the above mechanism, is Apriori [4]. On this basis some variants such as Apriori Tid, Apriori All, Apriori Some or Apriori Hibrid were developed. Figure 2 presents the Apriori algorithm. We use the following notation:

D- the transaction database (transaction table)
t- tuples in *D*
k-itemset-set of *k* items
F_k- frequents k-itemsets
C_k- k-itemsets candidates (potential frequents)
c.count – number of transactions containing each c candidate set of items

```
1 F1 = [frequent 1-itemsets]
2  k=2
3  while Fk-1≠Φ do
4   Ck=gen_apriori(Fk-1)    &&generating new candidates
5   for each t∈ D do
6      Ct={Ck|Ck⊂t)          &&candidates in t
7      for each c ∈ Ct do
8        c.count=c.count+1
9    end
10  Fk = {c ∈ Ck | c.count ≥ minsup}
11  k=k+1
12 end
13 end
```

Figure 2. The Apriori algorithm

Once found all frequent k-itemsets one can generate rules having minimum confidence *minconf*. In order to do that, we consider all non-empty subsets of every frequent itemset *f*. Then, for each subset *x*, we find the confidence of the rule x-> *(f-x)*, and if it is equal or greater than *minconf*, we output the rule.

Figure 3 presents in pseudo code the procedure for finding the association rules. We use the following notation:
F_k- frequents k-itemsets
A_i– a subset of j items from F_i (i<k,j<i)
B–the remaining subset of (i-j) items from F_i

```
1  F2 = [frequents 2-itemsets]
2   k>2
3  for i=2,k do
4    for j=1,i-1 do
5            Ai-1={aj|aj∈Fi}
6            B={b|b∈Fi ^ b≠aj}
7      conf(Ai-1→B)=supp(Ai-1∪B)/suppAi-1
8      Rij={Ai-1→B|conf(Ai-1→B)>minconf}
9    end
10 end
```

Figure 3. The algorithm for finding association rules

As we can see above, the rules are generated in an iterative way. In each iteration *j* we generate rules with consequent of size *j*. Then we consider the following property: for a frequent itemset, if a rule with consequent *b* holds, then rules with consequents that are subsets of *b* holds also. We use this property to generate rules in iteration j based on rules with consequents of length (j-1) found in the previous iteration.

IV. USING SQL FOR FREQUENT K-ITEMSETS GENERATION

In most implementation Apriori works with data stored in flat files but, data is ordinary stored in databases, so, for moving it in a flat file there are necessary some additional preprocessing operations.

Here we describe a way to find the frequent k-itemsets using SQL and manipulating data directly in a database. First of all we assume that the transaction table D has two columns: transaction identifier and item. As we don't know the number of items per transaction, this structure is more practical that alternatives presented in [5] where each item of a transaction is placed in a different column.

A. *Stored procedure approach*

In the first step of the algorithm one have to generate the frequent 1-itemset, by finding the support for each item and by removing those items that have support lesser that *minsup*.

The SQL query that performs this task, corresponding to the first line in the pseudo-code described above is:

```
insert into F1
    select item, count(*)                                   (4)
      from D
        group by item
        having count(*)>minsup;
```

Further, each step k of the algorithm, first generate a candidate k-itemset C_k from which we will find the frequent k-itemsets F_k. In order to do that, we have to execute the following:

- to generate C_2, which will be stored in a table with two columns (one for each item in the combination) we use:

```
insert into C2
    select a.item1, b.item2                                 (5)
     from F1 a, F1 b
    where a.item<b.item
    order by a.item;
```

- to find F_2 we must count the support of all 2-itemsets from C_2, and insert into F_2 only those who have the support greater than minsup. F_2 is a table with three columns (two columns for the two items and one column for support). We use a stored procedure that contain a sequence such as:

```
        ….
select count(id) into vsup
        from D
        where D.item=vc1 and exists
            (select tid
              from D d1                                     (6)
              where d1.item=vc2 and D.tid=d1.tid);
if vsup>vminsup then
        insert into F2
        values(vc1, vc2, vsup);
end if;
                  …….
```

To generalize, in order to find a frequent k-itemset F_k we will use the following SQL statements:

```
insert into Ck
  select a.item1,a.item2,…, a.item (k-1), b.item(k-1)
    from Fk-1 a, Fk-1 b
    where a.item1=b.item1 and                               (7)
        a.item2=b.item2  and
              ……
        a.item(k-2)=b.item(k-2)and
        a.item(k-1)<b.item(k-1)
    order by a.item1, a.item2,…, a.item(k-1);
```

for generating C_k and a stored procedure including:

```
    select count(id) into vsup
      from D
      where D.item=vc1 and exists                           (8)
```

```
(select tid
 from D d1
 where d1.item=vc2 and D.tid=D1.tid and exists
       (……
          (select tid
           from D dk-1
          where dk-1.item=vck
          and dk-2.tid=dk-1.tid)
                                        …);
```

in order to find the support of candidate k-itemsets.

One can see that for counting support for a k-itemset we need a select SQL statement that have k-1 subqueries, so, the number of nested selects is direct related with the size of the itemset to be analyzed.

B. The SQL Approach

To take advantage of the query optimizer, which is present in every relational DBMS, it is possible to implement Apriori algorithm only in SQL. In this case the generation of candidates Ck is made in the same manner as above.

In fact, the expression (7) generates an extensive set of candidate k-itemsets. If we consider that to be frequent, a k-itemset must contain only subsets of frequent (k-1) itemsets, it could be necessary to eliminate candidates that contain subsets that are not frequent from those generated by (7). This operation is the so-called pruning step, and is very useful because, after that, the remaining candidates could fit into memory, and the counting for support could be made pipeline, without materializing the candidates.

In [6] it is shown that we can perform the prune step in the same time with the join of F_{k-1} and F_{k-1}. In order to combine these tasks we could use the following SQL statement:

```
insert into Ck
 select I1.item1,I1.item2,…,I1.item(k-1), I2.item(k-1)
from Fk-1 I1, Fk-1 I2, Fk-1 I3,…,Fk-1 Ik
where
      I1.item1=I2.item1 and                                    (9)
      I1.item2=I2.item2 and
      ……
      I1.item(k-2)=I2.item(k-2)and
      I1.item(k-1)<I2.item(k-1)and
          I1.item2=I3.item1  and              && skip item1
                          ……
          I1.item(k-1)=I3.item(k-2)and
          I2.item(k-1)=I3.item(k-1)and
                 ……
          I1.item1=Ik.item1 and          && skip item(k-2)
                       ……
          I1.item(k-1)=Ik.item(k-2)and
          I2.item(k-1)=Ik.item(k-1)
order by I1.item1, I1.item2,…, I1.item(k-1);
```

The expression (9) is a **k-way join**. This method is used since for any k-itemset there are k subsets of size k-1, which must be member of the frequent (k-1)-itemsets (F_{k-1}).

In the above expression after joining I_1 and I_2, we obtain the following k-itemset ($I_1.item_1$, $I_1.item_2$,…,$I_1.item_{(k-1)}$, $I_2.item_{(k-1)}$). It contains two (k-1)-itemsets, which are frequent since they are member of F_{k-1}. These are ($I_1.item_1$,$I_1.item_2$, …,$I_1.item_{(k-1)}$) and ($I_1.item_1$, $I_1.item_2$,…,$I_1.item_{(k-2)}$, $I_2.item_{(k-1)}$). The rest of (k-2) subsets must be checked, and in order to do that we use additional joins whose selection predicates are build by skipping one item at a time from the k-itemset.

Concrete, first we skip $item_1$ and check if the (k-1) itemset ($I_1.item_2$, …, $I_1.item_{(k-1)}$, $I_2.item_{(k-1)}$) belong to F_{k-1}. This is done by the join with I_3. Second we skip $item_2$ to see if ($I_1.item_1$, $I_1.item_3$, …, $I_1.item_{(k-1)}$, $I_2.item_{(k-1)}$) belong also to F_{k-1}.

In general we perform the join with I_n and, in the predicate we skip the item (n-2) from the k-itemset to check if the subset build by deleting the $(n-2)^{th}$ item from the original k-itemset, belong to F_{k-1}.

Further it is necessary to count candidates' support to find frequent itemsets. In order to do that we use the candidate itemsets C_k and the transaction table D. We consider the two approaches, which were found to be the best ones in [5]. There are: the K-way joins and the subquery-based implementation.

In k-way joins the candidate itemset C_k is joined with k transaction tables D and the support is counted using a group by clause on all k items. The general expression for finding a frequent k-itemset F_k is:

```
insert into Fk
    select item1, item2, …, itemk, count(*)
      from Ck, D d1, …, D dk
      where    d1.item=Ck.item1 and
               d2.item= Ck.item2 and                          (10)
               ……..
               dk.item=Ck.itemk and
               d1.tid=d2.tid and
               d2.tid = d3.tid and
               ….
               dk-1.tid=dk.tid
      group by item1, item2, ….,itemk
      having count(*)>:minsup;
```

The subquery-based approach tries to reduce the amount of work during the support counting by using the common prefixes between the items in C_k. To do this, the support counting phase is split into a cascade of k subqueries.

The n^{th} (n=1..k) subquery is build based on the result of $(n-1)^{th}$ subquery, which is joined with the transaction table D, and the distinct itemsets consisting of the first n+1 columns of C_k. To obtain the final output we make a group-by on the k items to calculate the support and we remove those rows that have the calculated support less than minsup.

The statement for generating F_k is:

```
insert into Fk
   select item1, item2, …, itemk, count(*)
   from (SQk)                                                 (11)
   group by item1, item2, ….,itemk
   having count(*)>:minsup;
```

where SQ_k is the k^{th} subquery . Subquery Q_n (n=1..k) is

Select $item_1$, $item_2$, …, $item_n$, tid
From D d_n,(Subquery Q_{n-1}) SQ_{n-1},(select distinct $item_1$,…, $item_n$ from C_k) c_n
where $SQ_{n-1}.item_1 = c_n.item_1$ and (12)
… and
$SQ_{n-1}.item_{n-1} = c_n.item_{n-1}$ and
$SQ_{n-1}.tid = d_n.tid$ and
$d_n.item = c_n.item_n$;

V. CASE STUDY

The Center for Computer Research in the University "Stefan cel Mare" of Suceava has implemented the TERAPERS project [7]. TERAPERS is a system, which is able to assist teachers in their speech therapy of dislalya and to follow how the patients respond to various personalized therapy programs. This system is currently used by the therapists from Regional Speech Therapy Center of Suceava.

In the context of the need for more efficient activities, it was showed that data mining methods, applied to data collected in TERAPERS, can provide useful knowledge for personalized therapy optimization. So, the idea about Logo-DM system has started. This is a data mining system, which aims to use the data from TERAPERS database in order to answer the questions such as: what is the predicted final state for a child or what will be his/her state at the end of various stages of therapy, which the best exercises are for each case and how they can focus their effort to effectively solve these exercises or how the family receptivity, which is an important factor in the success of the therapy - is associated with other aspects of family and personal anamnesis [8][9].

Specifically, association rules could reveal interesting relationships between patients' anamnesis, their diagnosis and the results obtained at the end of various stages of therapy.

To find the best way to implement the association rules we have used a real dataset from TERAPERS, on which we have applied the three methods of determining the frequent itemsets discussed above.

Data considered is stored in a relational table with 96 columns (attributes) and 300 cases. A step in the preparation process of this table for applying Apriori algorithm is to change the structure of the initial table in order to obtain the following structure: *id, item.*

After this transformation we have obtained a data set that can be assimilated to a transactional one. As each patient has associated a set of features, we can consider the patient identifier as a transaction identifier and the set of features as a set of items. The transactional dataset obtained as result of pre-processing operations contains 300 cases (transactions), but the average of items per transaction is 65.

We have applied on this data the three methods for finding the frequent itemsets discussed above. All of them have provided as output the same number of frequent itemsets and the same maxim length of the itemsets, but major differences were recorded on response times. To enable the query optimizer to choose the best execution plan, we have constructed some indexes. The source table for data used by Apriori algorithm was indexed on both columns (id, item), and C_k and F_k were indexed on columns ($item_1$, $item_2$, …, $item_k$).

The results obtained are presented in figures below.

Figure 4 and Figure 5 show the number of itemsets obtained and the maximum length of these itemsets. We note that, in this case, taking into account the data characteristics, it should be imposed a minimum value for the threshold *minsup* equal to 0,5. For this value is obtained a considerable number of itemsets (approximately 4000000) with a maximum length of 32 items.

Figure 6 presents a summary of execution times obtained when, in order to finding frequent itemsets, we use the following methods: stored procedures, k-way joins and subquery-based joins. It may be noted that for the last two ways, execution times are relatively closed and they are lower than for stored procedures.

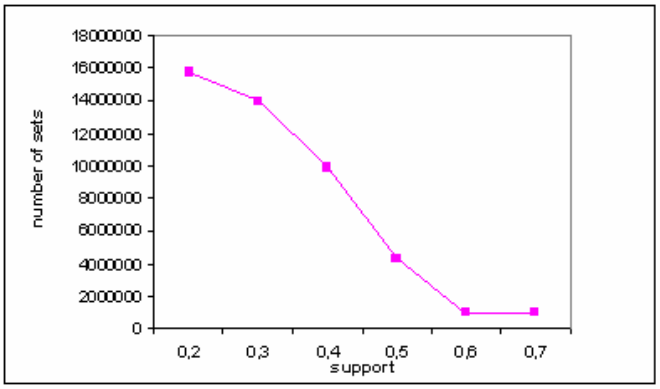


Figure 4. Frequent itemsets found for different values for *minsup*

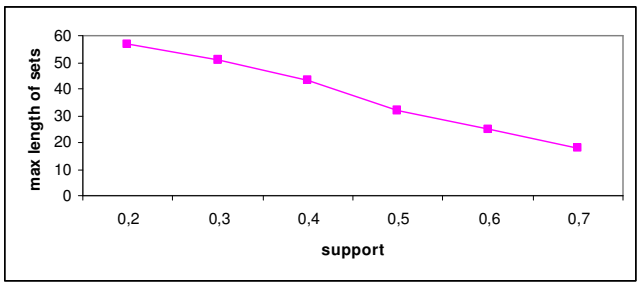


Figure 5. Maximum length of itemsets for different values for *minsup*

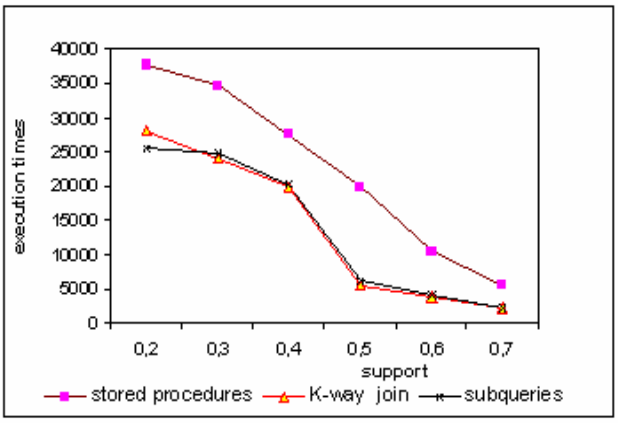


Figure 6. A comparation of execution times

VI. CONCLUSION

Last time, data mining has become an area of special importance. Based on the results obtained using the patterns provided by data mining in fields such as business or medicine, we have tried to extend the implementation of these methods in other areas. This paper presents a part of the efforts made in order to achieve and to early evaluate the performances of finding the association rules for a system that aims to optimize the personalized therapy of speech disorder.

To take advantage of the facilities offered by the query optimizer, it is proposed to incorporate data mining algorithms in the database that stores data to be analyzed.

First, it was noted that in this context the data pre-processing operations are relatively simple to be achieved. Then it was observed that, due to the data set properties, the largest number of candidate itemsets was not obtained in the second step, and for about half of iterations the number of candidates and of frequent itemsets increases, then this number decreases. The data source contains a lot of items that are found in a large percentage of transactions. As a result it is indicated for the support's threshold to be established to a value of least 0,5. Even in this case the number and the maxim length of frequent itemsets are high. As a result it will get a large number of association rules that should be filtered in order to preserve only the most interesting ones.

ACKNOWLEDGMENT

This paper was supported by the project "Progress and development through post-doctoral research and innovation in engineering and applied sciences– PRiDE - Contract no. POSDRU/89/1.5/S/57083", project co-funded from European Social Fund through Sectorial Operational Program Human Resources 2007-2013.

REFERENCES

[1] J. Han, Y. Fu, K. Koperski, W. Wang, and O. Zaiane. DMQL: A data mining query language for relational datbases. In Proc. of the 1996 SIGMOD workshop on research issues on data mining and knowledge discovery, Montreal, Canada, May 1996.

[2] R. Agrawal and K. Shim. Developing tightly-coupled data mining applications on a relational database system. In Proc. of the 2nd Int'l Conference on Knowledge Discovery in Databases and Data Mining, Portland, Oregon, August 1996.

[3] R. Agrawal, T. Imielinski, and A. Swami. Mining association rules between sets of items in large databases. In Proceedings of the ACM SIGMOD International Conference on Management of Data (ACM SIGMOD '93), 1993

[4] R. Agrawal and R. Srikant. Fast algorithms pentru mining association rules. In Proceedings of the 20th International Conference on Very Large Databases (VLDB '94), Santiago, Chile, June, 1994

[5] K. Rajamani, B. Iyer, and A. Chaddha. Using DB/2's object relational extensions for mining associations rules. Technical Report TR 03,690., Santa Teresa Laboratory, IBM Corporation, sept 1997.

[6] S. Sarawagi, S. Thomas, and R. Agrawal. Integrating Association Rule Mining with Relational Database Systems: Alternatives and Implications. In Proc. of the ACM SIGMOD Conference, Seattle, Washington, June 1998.

[7] M. Danubianu, S.G. Pentiuc, O. Schipor, M. Nestor, I. , Ungurean. Distributed Intelligent System for Personalized Therapy of Speech Disorders, Proceedings of ICCGI08, 2008, Atena

[8] M. Danubianu, S.G. Pentiuc, I. Tobolcea, O.A. Schipor. Advanced Information Technology - Support of Improved Personalized Therapy of Speech Disorders, International Journal of Computers Communications & Control, ISSN 1841-9836, 5(5): 684-692, 2010.

[9] S. G. Pentiuc, I. Tobolcea, O. A. Schipor, M. Danubianu, D. M. Schipor, "Translation of the Speech Therapy Programs in the Logomon Assisted Therapy System," Advances in Electrical and Computer Engineering, vol. 10, no. 2, pp. 48-52, 2010, http://dx.doi.org/10.4316/AECE.2010.02008

 ISBN: 978-1-61208-139-7

Statistical Machine Translation as a Grammar Checker for Persian Language

Nava Ehsan, Heshaam Faili
Department of Electrical and Computer Engineering, University of Tehran
Tehran, Iran
{n.ehsan@ece.ut.ac.ir, hfaili@ut.ac.ir}

***Abstract*—Existence of automatic writing assistance tools such as spell and grammar checker/corrector can help in increasing electronic texts with higher quality by removing noises and cleaning the sentences. Different kinds of errors in a text can be categorized into spelling, grammatical and real-word errors. In this article, the concepts of an automatic grammar checker for Persian (Farsi) language, is explained. A statistical grammar checker based on phrasal statistical machine translation (SMT) framework is proposed and a hybrid model is suggested by merging it with an existing rule-based grammar checker. The results indicate that these two approaches are complimentary in detecting and correcting syntactic errors, although statistical approach is able to correct more probable errors. The state-of-the-art results on Persian grammar checking are achieved by using the hybrid model. The obtained recall is about 0.5 for correction and about 0.57 for detection with precision about 0.63.**

***Index Terms*—Natural Language Processing, Syntactic Error, Statistical Machine Translation, Grammar Checker, Persian Language**

I. Introduction

Proofreading tools for automatic detection and correction of erroneous sentences are one of the most widely used tools within natural language applications such as text editing, optical character recognition (OCR), machine translation (MT) and question answering systems [1]. The editorial assistance tools are useful in helping second language learners not only in writing but also in learning a language by providing valuable feedbacks [2]. Kukich [3] has categorized the errors of a text into five groups, 1. Isolated, 2. Non isolated or syntactic errors, 3. Real-word errors, 4. Discourse structure, and 5. Pragmatic errors. The first category refers to spelling errors. Detecting errors of second and third categories need syntactic as well as semantic analysis. The last two hierarchies cannot be considered as spelling or grammatical error. In this article we just focus on correcting syntactic errors and presuppose that the text is spell checked correctly. This paper is going to describe a statistical grammar checker approach within the framework of phrasal statistical machine translation. SMT has the potential to solve some kind of errors occurring in the sentences [4]. We will show that training statistical model would be helpful in detecting and correcting grammatical errors which were not addressed in the rule-based grammar checker [5] especially those errors which need contextual cues for recognition. We will also introduce a hybrid of statistical and rule-based approaches for grammar checking and achieved the state-of-the-art results on Persian grammar checking. Grammar checkers cannot check the whole syntactic structure of the text [6]. In the proposed model, frequently occurred error types have been identified for evaluating both error detection and correction of the system in terms of precision and recall metrics.

The remainder of the paper is organized as follows: Section 2 outlines related works of grammar checking. In Section 3, the limitation of previous Persian grammar checker is discussed. Section 4 describes the use of SMT framework for grammar checking followed by preparing training and test data set. Finally, the evaluation results for each approach individually and the hybrid model are reported in Sections 5 and 6, respectively.

II. Related Work

Grammar checkers deal with syntactic errors in the text such as subject-verb disagreement and word order errors. Grammar checking entails several techniques from the NLP research area such as tokenization, part-of-speech tagging, determining the dependency between words or phrases and defining and matching grammatical rules. Grammar checking techniques are categorized into three groups: syntax-based, statistical or corpus-based and rule-based [2]. In syntax-based approach the text is parsed and if parsing does not succeed the text is considered as incorrect. It requires a complete grammar or mal-rules or relaxing constraints which are obviously difficult to obtain due to complex nature of natural languages. Mal-rules allow the parsing of specific errors in the input and relaxing constraints redefine unification so that the parse does not fail when two elements do not unify [2]. The existing grammar checkers [7][8] fall into rule-based category in which a collection of rules describe the errors of the text, while [9][10][11] use statistical analysis for grammar checking. Although, rule-based grammar checkers have been shown to be effective in detecting some class of grammatical errors, manual design and refinement of rules are difficult and time-consuming tasks. Deep understanding of the linguistics is required to write non-conflicting rules which cover a suitable variety of grammatical errors. Although there have been some prior works on Persian spell checking, [12][13] from the best of our knowledge, the only work on Persian grammar checking is a rule-based system which is introduced in [5]. The limitations of this grammar checker are described in detail in next section.The ALEK system

 ISBN: 978-1-61208-139-7

developed by [9] uses an unsupervised method for detecting English grammatical errors by using negative evidence from edited textual corpora. It uses TOEFL essays as its resource. Integrating pattern discovery with supervised learning model is proposed by [11]. A generation-based approach for grammar correction is introduced by [10] which checks the fluency of sentences produced by second language learners. The N-best candidates are generated using n-gram language model which are reranked by parsing using stochastic context-free grammar. A pilot study of [4] presents the use of phrasal SMT for identifying and correcting writing errors made by learners of English as a second language and the focus was on countability errors associated with mass nouns. The statistical phase of grammar checking procedure introduced in this paper also relies on phrasal SMT framework for detecting and correcting syntactic errors. To overcome the negative impact of some types of errors on recall metric, the system is augmented with the rule-based procedure.

III. Limitations of Persian Rule-based Grammar Checker

The proposed rule-based grammar checker [5] faces some limitations. It is based on regular expression patterns and detects errors which can be matched by regular expressions, thus it cannot detect those patterns which are difficult or impossible to be modeled by regular expressions. The other problem is having pre-defined pattern and suggestion for each type of error. For example whenever it detects two repeated words, it shows an error although not all two repeated words are incorrect and one of them is deleted due to pre-defined suggestion. Our method is an SMT-based approach which does not follow any specific pre-defined rule or suggestion. For example by detecting repeated words, in some cases one of the words may be eliminated or sometimes a preposition is added between duplicate words and in some cases it does not recognize any error. In addition regular expressions cannot detect any recursive pattern. The errors which need context free grammar or statistical or semantic analysis or disambiguation are also undetectable by regular expressions. Existing techniques for Persian, based on hand-crafted rules or statistical POS tag sequences [5] are not strong enough to tackle the common incorrect preposition or conjunction omission errors due to lack of information about language model. In our experiments not only all the syntactic errors described in the rule-based grammar checker are included but also, the following errors are added. The errors are illustrated with an example.

(1) Omission of prepositions: Some words need special prepositions to complete their meanings. Prepositions depend on nouns and can complement the other words. Since the lexical information is important to correct omission of prepositions, we need to define large number of rules which include all the phrases containing prepositions and the words around it. Thus, it is not feasible to define the patterns by regular expressions. For example, بحث سر تفاوتها است/ *bahs sar tafaavotha ast* (the discussion is differences) should be corrected as بحث بر سر تفاوتها است/ *bahs bar sar tafaavotha ast* (the discussion is about differences).

(2) Omission of را/ *ra* (definite object sign): Object is a mandatory argument of transitive verbs. The meaning of transitive verb is incomplete and unclear without the object. The direct object should be addressed in the sentence by preposition را/ *ra*. Finding the object of the sentence requires semantic analysis and it cannot be detected by regular expressions. Since this is an important preposition, the rule has been considered separately. For example, کار شروع کردید/ *kar shooroo kardid* (you started work) should be corrected as کار را شروع کردید/ *kar ra shooroo kardid* (you started the work).

(3) Omission of conjunctions: The omission of conjunctions is not always incorrect, but there are some cases that the omission makes the sentence grammatically wrong. This usually happens when a clause appears in the middle of the sentence. Lexical information is also important in this case. For example, تعریفی خود شما دارید چیست/ *tarifi khod shoma darid chist* (what is the description you have) should be converted to تعریفی که خود شما دارید چیست/ *tarifi ke khod shoma darid chist* (what is your description).

(4) Using indefinite noun when a demonstrative pronoun is used: Demonstrative pronouns are independent words that precede the noun. After demonstrative pronouns such as این/ *in* (this) and آن/ *an* (that) a definite noun should be used, unless a description is given within a phrase, like آن کتابی که خواندم/ *an ketaabi ke khaandam* (the book that I have read). Since regular expressions cannot identify to which word of the sentence the descriptive phrase belongs, defining this rule with regular expression may result in many false alarms. For example,آن کتابی را خواندم/ *an ketaabi ra khaandam* (I read that a book) should be changed to آن کتاب را خواندم/ *an ketaab ra khaandam* (I read that book).

(5) Connecting indefinite postfix to the first noun in possessive nouns (ezafe construction [14]): Persian is a dependent-marking language [15] and tends to mark the relation on the non-head. In case of having possessive nouns indefinite postfix ی/ *i* should be connected to the last word. The postfix ی/ *i*, is not only used as indefinite sign, but also can be used as a copula for the second singular person like آزادی/ *azaadi* (you are free) and it may also belong to its own word like آزادی/ *azaadi* (freedom). The morpheme ی/ *i* is used in forming various lexical elements in derivational morphology and will cause ambiguities [14]. Since regular expressions deal with the surface of the word without semantic analysis these ambiguities are not distinguishable by regular expressions. For example, کتابی داستان/ *ketaabi daastaan* should be corrected as کتاب داستانی/ *ketaab*

daastaani (a story book).

(6) Using adjective before noun: In Persian, adjectives usually follow the nouns. This rule was omitted from the rule-based grammar checker [5] according to its low precision. Adjectives can sometimes stand as the adverb of the sentence and they can precede the noun. The part-of-speech tagger used in the rule-based grammar checker could make mistakes in recognizing adjectives. Also there is no chunking process before applying regular expressions, thus the rule-based system cannot understand if the adjective belongs to the same phrase as nouns or not. For example, جالب کتاب/ *jaaleb kettab* (book interesting) should be converted to کتاب جالب/ *ketaab jaleb* (interesting book) but دانشمند مرد را دید/ *daaneshmand mard ra did* (scientist saw the man) is correct.

(7) Using wrong plural morpheme: Morphologically, Persian falls into polysynthetic languages in which a single word may have many morphemes and also several morphemes exist for marking plurality in Persian, like ان/ *an*, ها/ *ha* , یون/ *yun* and ات/ *at* [16]. Some words like درخت / *derakht* (tree) can become plural with ان/ *an* and ها/ *ha*. Both words درختان/ *derakhtan* (trees) and درختها/ *derakhtha* (trees) are correct but, some words like انسان/ *ensaan* (human) or میز/ *miz* (table) can become plural with ها/ *ha* but not with ان/ *an* and unfortunately there is no special rule for that and it cannot be defined by regular expressions. For example, انسانان) ان + انسان) *ensaanan* should be converted to انسانها/ *ensaanha* (humans).

In brief, the mentioned errors can be detected either by using probabilistic context free grammar (PCFG) as a modeling formalism, or by using statistical or semantic analysis or by defining too many lexical rules. Due to the discussed limitations of regular expressions, we used a statistical approach based on SMT framework for grammar checking to overcome the problems of the existing rule-based grammar checker.

IV. SMT Framework for Grammar Checking

Machine translation refers to usage of computer to automate some or all of the process of translating from one language into another [17]. Automatic grammar checking is modeled as a machine translation where the erroneous sentence is translated to the correct sentence. Machine translation is considered as a hard task [15] in general due to differences between the languages which are referred to as translation divergences [15]. Translation divergence could be structurally, like differences in morphology, argument structure, ordering, referential density and linking of predicates with their arguments or it could be lexically like homonymous, polysemy, many-to-many translation mappings and lexical gaps. The less divergence between source and target languages leads to better translation. Unlike machine translators that the input sentence belongs to a language other than output sentence and could have many differences structurally and lexically, in the proposed model for grammar checking both input and output sentences belong to the same language except the input sentence has some syntactic errors. As it will be described later the syntactic errors considered in this paper could cause lexical gap or divergences in morphology, argument structure or ordering, between source and target sentences. Stylistic and cultural differences which are another source of difficulty for translators do not appear in this model. The noisy channel model is the foundation of statistical machine translation [18]. In this article we explore its application to grammar checking. The noisy channel is used whenever the received signal does not identify the sent message. Grammar checking could be modeled as a noisy channel where the intended message is the correct sentence while the received signal is the erroneous sentence. We assume grammar checking from an incorrect sentence to correct sentence. The suggested correct sentence is the one whose probability is the highest:

$$\hat{C} = \arg\max_c P(C|E) = \arg\max_c \frac{P(E|C)P(C)}{P(E)} \quad (1)$$

The probability in the denominator of equation 1 is ignored since we are choosing the correct sentence for a fixed erroneous sentence, thus is a constant. Equation 1 shows that we need to compute $P(E|C)$ and the language model $P(C)$. We assume that the noisy sentence is the result of applying syntactic errors on the correct sentence. The syntactic error rules considered in this paper are classified in Table I and we refer to them as $R = r_1, r_2, ..., r_n$. The relationship can be expressed as follows:

$$P(E|C) = \sum_{i=1}^{n} P(E, r_i|C) \quad (2)$$

The conditional probability $P(E|C)$ is computed as follows:

$$P(E|C) = \sum_{i=1}^{n} P(E|C, r_i) * P(r_i|C) \quad (3)$$

Two assumptions have been made, although we will later show in our experiments that these assumptions will not really affect the system's accuracy regarding to precision and recall metrics.

(1) Each sentence could just have one syntactic error. In other words, just one of the error rules could be applied on the sentence.

(2) The condition probability $P(r_i|C)$ has uniform distribution. It means that each rule is equally likely to be applied on a correct sentence. That is, the probability of appearing each error mentioned in Table I, is the same.

Thus, equation 1 is defined as follows:

$$\hat{C} = \arg\max_c P(C|E) = \arg\max_c P(E|C, r_i)P(C) \quad (4)$$

where r_i is the error rule which was applied on the correct sentence, (C).

 ISBN: 978-1-61208-139-7

TABLE I
PERSIAN SYNTACTIC ERROR RULES

ID	Error description
1	Omission of preposition
2	Omission of conjunction
3	Using plural noun after cardinal numbers
4	Using a verb or preposition after a genitive noun ending with sign ی/ *i*
5	Using a verb before copulative verbs
6	Using a superlative adjective before preposition از/ *az* (than)
7	Using a preposition or conjunction at the end of the sentence
8	Omission of را/ *ra* (definite object sign)
9	Using را/ *ra* (definite object sign) after verb or preposition or in the beginning of the sentence
10	Using two consecutive adverbs of question or pronouns without و/ *va* (and)
11	Double plural noun
12	Using adjective before noun
13	Disagreement between the subject and the verb
14	Repeating a word
15	Using wrong plural morpheme
16	Using indefinite noun when a demonstrative pronoun is used
17	Connecting indefinite postfix to the first noun in ezafe constructions

In order to use the phrase-based translation model, we need a training data. Training data construction is described in detail later in this paper. Here, we just mention that we have parallel corpora of correct and erroneous sentences in which correct sentences are infected by one of the error rules to produce the erroneous sentences. If more than one rule are applicable on a sentence, separate sentences are produced each containing only one error. Phrasal translation model uses phrases as well as single words as the fundamental units. Phrase translation probability, $\phi(\overline{e_i}, \overline{c_i})$, is defined as the probability of generating phrase $\overline{c_i}$ from incorrect phrase $\overline{e_i}$. Distortion refers to a word having a different position in the input and output sentences. The more reordering the more expensive is the translation. The distortion is parameterized by $d(a_i - b_{i-1})$, where a_i is the start position of the erroneous phrase generated by the i-th correct phrase, and b_{i-1} is the end position generated by the $(i-1)$-th correct phrase. The phrase translation probability and distortion probability are computed as follows [15]:

$$\phi(\overline{e}, \overline{c}) = \frac{count(\overline{e}, \overline{c})}{\sum_{\overline{e}} count(\overline{e}, \overline{c})} \tag{5}$$

$$d(a_i - b_{i-1}) = \alpha^{|a_i - b_{i-1} - 1|} \tag{6}$$

where, α is a small constant.

Similar to the translation model for statistical phrasal MT the conditional probability in equation 1 is decomposed into:

$$P(E|C) = \prod_{i=1}^{I} \phi(\overline{e_i}, \overline{c_i}) d(a_i - b_{i-1}) \tag{7}$$

Moses [17] is a statistical machine translation system for automatically training translation models and decoding for any language pair in which GIZA++ [19] is used for word-alignments and SRILM [20] is used as the language model toolkit. As in phrase-based models, factored translation model which is used in Moses can be seen as the combination of several features $h_i(C|E)$. These features are combined in a log-linear model. If there are N features, then the log-linear translation model is:

$$P(C|E) = \frac{1}{Z} \prod_{i=1}^{N} \alpha_i^{h_i(C|E)} \tag{8}$$

Here, Z is a normalizing constant and α_i is the weight assigned to feature In practice, the noisy channel model factors (the language model $P(C)$ and translation model $P(E|C)$), are still the most important feature functions in the log-linear model, but the architecture has the advantage of allowing for arbitrary other features as well [15]. If two weights and features are used, and set them relative to the language model $P(C)$ and the conditional probability $P(E|C)$ as follows:

$$h_1(C|E) = \log_{\alpha_1}^{P(C)} \tag{9}$$

$$h_2(C|E) = \log_{\alpha_2}^{P(E|C)} \tag{10}$$

We can see by replacing equations 9 and 10 in equation 8 that the fundamental equation 1 is the special case of the log-linear model [21]. Language model is used to score the fluency of the output. Phrase translation table saves the extracted phrases. The construction of phrase table and learning phrases are described in [22]. Distortion model allows reordering of the input sentence, but at a cost. The main features of Moses are distortion model, language model, translation model and word penalty [22].

A. Preparing Training and Test Data

Training data is a collection of aligned sentences in two files, one for ungrammatical sentences and one for correct sentences. In order to prepare the erroneous corpus, the set of specific error types mentioned in Table I are used. These errors include all the patterns defined in [5] and those defined in Section 3. A Persian part-of-speech tagged corpus named Peykareh [23] containing collection of formal news and common well-formed texts is used. A set of example sentences for each of various error types should be collected, thus, the error rules are injected to the sentences of this corpus automatically. The underlying assumption that each sentence has one syntactic error has to be held. If more than one type of error are possible in a sentence, erroneous sentences are made each containing only one type of error and the corresponding correct sentence is placed for each sentence in another file. The sentences larger than 25 words are pruned to make the training phase more practical. The prepared training set contains about 340,000 erroneous sentences; each corresponds to a correct sentence in another file. The language model is also created from the correct corpus by SRILM toolkit [20]. We used some parts of Peykareh, other than those used in training phase, as test set, to evaluate the result of the system. For each type of

 ISBN: 978-1-61208-139-7

TABLE II
CATEGORIES OF DEFINED ERROR RULES

Category	Rule number
An unnecessary word	7,9,14
A missing word	1,2,4,7,8,9,10
A word or phrase that needs replacing	9,12
A word used in the wrong form	3,4,5,6,11,13,15,16,17

error mentioned in Table I, a set of 20 samples are injected in the test set. These sentences are considered as the input of the system. The results are illustrated in Figure 1. In each case if the created error leaves the sentence meaningful, that sentence is not evaluated. Thus, the assessment set are those sentences that contain real grammatical errors which contains 321 erroneous sentences. Dealing with null subject and pro-drop feature of Persian language, error number 13 the subject-verb disagreement error, is examined when the subject is a pronoun and it is not dropped.

V. EXPERIMENTAL RESULTS

Error rules are classified and results are evaluated for each rule individually. Correction is referred to those sentences which the output is a correct sentence and detection is referred to those sentences which system made changes to the input sentence, in the scope of the error, but the suggested change may not be correct. The error rules are those shown in Table I. Nicholls [24] identifies four error types: an unnecessary word, a missing word, a word or phrase that needs replacing and a word used in the wrong form. Table II shows that to which categories our defined error rules belong, the error rules are referred by their number defined in Table I. As shown in Table II some error rules could belong to more than one category. The first and second category cause lexical gap between the erroneous and correct sentences, the third category causes ordering differences and the category fourth causes morphological differences while error number 17 will cause difference in argument structure which emphasizes on dependant marking feature of the language. A similar classification is also introduced by [25] which does not contain the third category and includes a separate category for agreement errors. The recall results of 20 samples of each error rule are demonstrated in Figure 1. The horizontal numbers in Figure 1 indicate the error numbers described in Table I.

The recall of the system was 0.44 for correction and 0.48 for detection with precision 0.61. The results of Figure 1 indicate that this approach is successful in detecting some rules which were not detectable by regular expressions as discussed before, rules number 1,2,8,12,15,16 and 17, on the other hand it cannot detect the rules number 4,7 and 9 which are detectable by regular expressions. These rules are those that could be belonged to different categories. Since there could be different ways for correcting these errors, it seems that the translation probabilities of possible solutions are distributed in the training set. The rule-based grammar checker proposed in [5] is tested on this test set which the recall results are demonstrated in Figure 2. Here, detection means flagging the error with or without suggestion. The recall was 0.23 for correction and 0.35 for detection on the defined rules with precision 0.94.

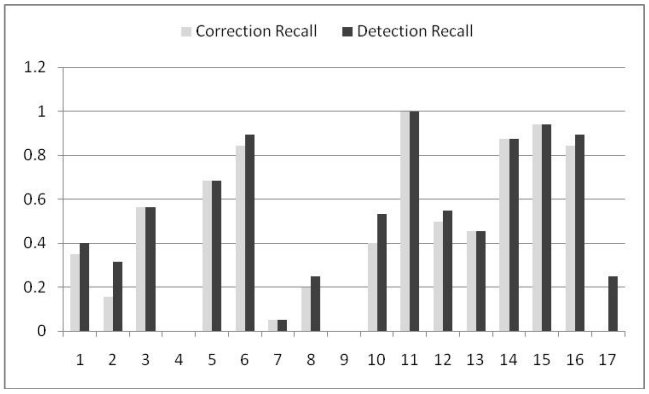


Fig. 1. Results for SMT-based Grammar Checker

VI. COMBINING SMT AND RULE-BASED GRAMMAR CHECKER

The proposed SMT procedure performed as an error corrector. It does not contain error detection in first stage and all sentences are regarded as erroneous for correction. The error detection defined in previous section actually refers to incorrect correction. Some rules have negative impact on recall of the SMT-based technique which are detectable by rule-based approach. We would expect to see a greater improvement by combining these two techniques. In this case, errors are detected either by SMT-based or rule-based grammar checker or both. If the correction differs in two systems, both could be shown to the user as suggestions. If one of the suggestions is correct we assume that the system corrected the sentence. The results are illustrated in Figures 3 and 4 for detection and correction recall respectively. The recall improved to 0.53 for error correction and 0.66 for error detection while the precision was 0.67.

VII. DISCUSSION

In the previous experiments, we relied on the assumption that each error is equally likely to be applied on a correct sentence and in the test set the errors were uniformly distributed. The likelihood occurrences of the errors were not considered.

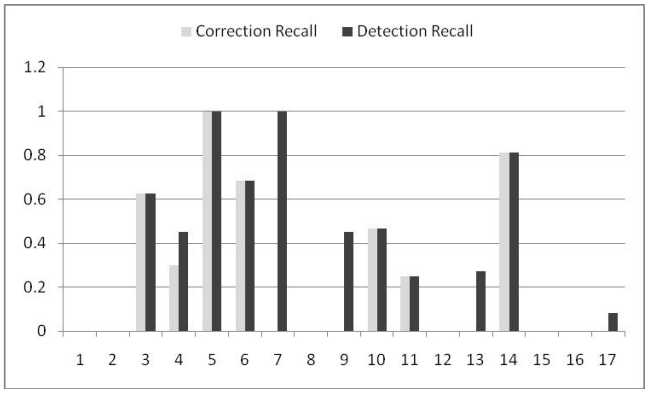


Fig. 2. Results for Rule-based Grammar Checker

 ISBN: 978-1-61208-139-7

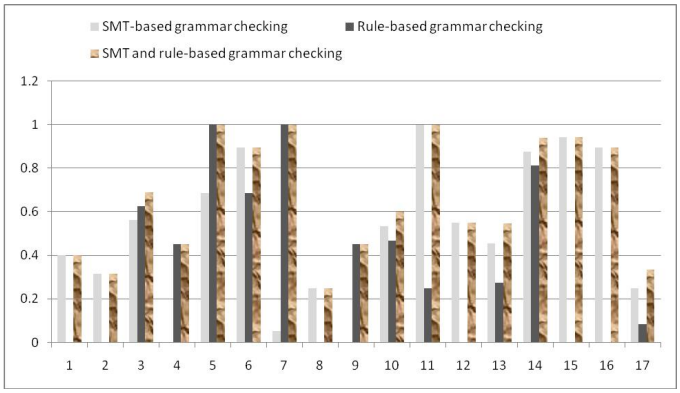


Fig. 3. Combining SMT and rule-based results for grammar detection

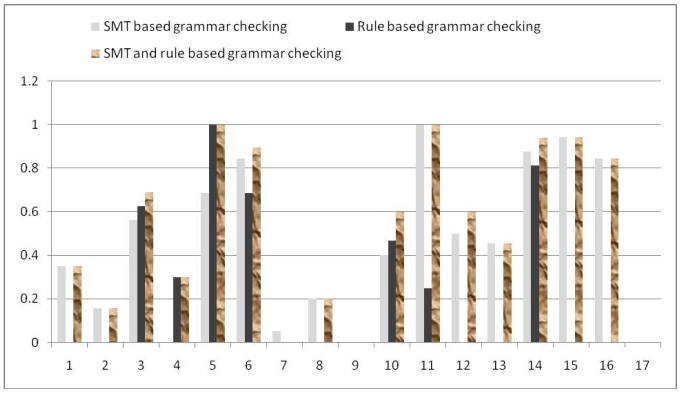


Fig. 4. Combining SMT and rule-based results for grammar correction

In the real world some types of errors are more probable to happen than the others. The conditional probability $P(r_i|C)$ is defined as the probability of making the error r_i when it is applicable to a correct sentence. Although there are learner corpora for some languages like ICLE (Cambridge Learner Corpus) and JLE (Japanese Learners of English Corpus) that contain annotated errors, there is no such annotated corpora available to date for Persian language to compute the probability occurrence of each type of error. The new method to process the free resource of revision histories of Wikipedia to create error corpora, have shown in experiments that even large revision histories contain rather scarce information about errors [26]. In this survey we asked from 19 native speakers the probability of making each type of error in the texts. The answers were classified to high, medium and low. Computing the weighted average of the answers the likelihood occurrence of each error in the texts is estimated. The weights are 80, 50 and 20 percent for high, medium and low classes respectively. The results are given in Table III. The information given in Table III indicates that those errors which could not be recognized by SMT-based grammar checker are the less probable errors of the language.

Similar to the work of [27] which appended the difficult-to-translate phrases with human translations to the training set to reduce the negative impact of these phrases, this time we

TABLE III
OCCURRENCE PROBABILITY FOR EACH TYPE OF ERROR

Rule Number	Probability (%)
1	67.36
2	35.78
3	34.21
4	35.78
5	40.52
6	27.89
7	31.05
8	40.52
9	26.31
10	51.57
11	61.05
12	26.31
13	43.68
14	43.68
15	48.42
16	37.36
17	45.26

made a training set by considering the occurrence probability of errors. If an error was applicable to a sentence it is injected regarding to the relevant probability which results in about 220,000 pair sentences. We refer to our previous train set as train set1 and to our newly produced train set as train set2 which will result in statistical grammar checker1 (SGC1) and statistical grammar checker2 (SGC2). In order to test the results, another experiment is done on 500 erroneous sentences from the test set in which we used the likelihood occurrence information of errors where the error is applicable to the sentence. This test set is evaluated with both SGC1 and SGC2. The language model is same for both models. We refer to this test set as probable test set. The all results are summarized in Table IV. In order to consider the importance of precision for grammar checker [28] we have also evaluated the $F_{0.5}$ measure to weight precision twice as much as recall. In the test sets used so far, all sentences contained just one grammatical error. In order to realize that whether the existence of more than one error would affect the grammar checking process, 20 sentences are tested each containing more than one type of error. The results indicated that the grammar checker may not be able to recognize an error (the error which was previously recognized) if two types of errors happened in the same phrase.

VIII. CONCLUSION

This paper proposed a hybrid model of statistical and rule-based approaches for identifying grammatical errors for Persian language. The statistical part is based on phrasal SMT and the principles are language independent. The studies show that employing SMT framework for grammar checking has the ability to correct some class of errors which are the most probable errors in the sentences. To overcome the negative impact of some types of errors on the recall metric, the system is augmented with the rule-based procedure. The obtained recall was 0.53 for error correction and 0.66 for error detection while the resulted precision is 0.67 without considering the likelihood of occurrences of errors in the text. The likelihood of occurrences of each error type is estimated to be able to

 ISBN: 978-1-61208-139-7

TABLE IV
SUMMARIZED RESULTS OF GRAMMAR CHECKERS

		Correction recall	Detection recall	Precision	F-1	F-0.5
Uniform test set	SGC1	0.44	0.48	0.61	0.53	0.566
	Rule-based grammar checker	0.23	0.35	0.94	0.51	0.581
	SGC1 + Rule-based grammar checker	0.53	0.66	0.67	0.66	0.636
Probable test set	SGC1	0.46	0.5	0.61	0.54	0.572
	SGC2	0.48	0.5	0.62	0.55	0.585
	Rule-based grammar checker	0.25	0.31	0.91	0.46	0.595
	SGC2 + Rule-based grammar checker	0.5	0.57	0.63	0.59	0.599

evaluate the grammar checker more accurately. In this case, the obtained recall is 0.5 and 0.57 with augmentation of rule-based approach. This is the state-of-the-art results on Persian grammar checking so far.

IX. FUTURE WORKS

There are still number of tasks to improve the grammar checking system. We would like to collect grammatical errors from non-native learners which allow us to expand the grammar checker to better distinguish correct and erroneous sentences for language learners. It can also help to find better training examples for the system. Some errors in the sentence are result of real word errors. The SMT-based framework seems to be able to detect one word among the sentence that does not fit. The real-word error detection is going to be tested with this approach. Since the statistical approach is language independent it can be trained and tested on the other languages such as English, considering the errors of the language.

REFERENCES

[1] M. Bhagat, "Spelling error pattern analysis of punjabi typed text," Master's thesis, Computer Science and Engineering Department Thapar Institute of Engineering and Technology Deemed University Patiala, 2007.

[2] C. Leacock, M. Chodorow, M. Gamon, and J. Tetreault, "Automated grammatical error detection for language learners," *Synthesis Lectures on Human Language Technologies*, vol. 3, no. 1, pp. 1–134, 2010.

[3] K. Kukich, "Techniques for automatically correcting words in text," *ACM Computing Surveys (CSUR)*, vol. 24, no. 4, pp. 377–439, 1992.

[4] C. Brockett, W. Dolan, and M. Gamon, "Correcting ESL errors using phrasal SMT techniques," in *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, 2006, pp. 249–256.

[5] N. Ehsan and H. Faili, "Towards grammar checker development for Persian language," in *in 6th IEEE International Conference on Natural Language Processing and Knowledge Engineering (IEEE NLP-KE'10)*, 2010, pp. 150–157.

[6] D. Kies, "Evaluating Grammar Checkers: A Comparative Ten-Year Study," in *in 6th International Conference on Education and Information Systems, Technologies and Applications: EISTA*, 2008.

[7] D. Naber, "A rule-based style and grammar checker," Master's thesis, Diplomarbeit.Technische Fakultat Universitat Bielefeld, 2003.

[8] F. Bustamante and F. León, "GramCheck: A grammar and style checker," in *Proceedings of the 16th conference on Computational linguistics-Volume 1*, 1996, pp. 175–181.

[9] M. Chodorow and C. Leacock, "An unsupervised method for detecting grammatical errors," in *Proceedings of the 1st North American chapter of the Association for Computational Linguistics conference*, 2000, pp. 140–147.

[10] J. Lee and S. Seneff, "Automatic grammar correction for second-language learners," in *Ninth International Conference on Spoken Language Processing*, 2006, pp. 1978–1981.

[11] G. Sun, X. Liu, G. Cong, M. Zhou, Z. Xiong, J. Lee, and C. Lin, "Detecting erroneous sentences using automatically mined sequential patterns," in *Annual Meeting-Association for Computational Linguistics*, vol. 45, no. 1, 2007, pp. 81–88.

[12] M. Shamsfard, H. Jafari, and M. Ilbeygi, "Step-1: A set of fundamental tools for persian text processing," in *In 8th Language Resources and Evaluation Conference*, 2010.

[13] O. Kashefi, M. Nasri and K. Kanani Towards Automatic Persian Spell Checking. *Tehran, Iran: SCICT*, 2010.

[14] K. Megerdoomian, "Finite-state morphological analysis of Persian," in *Proceedings of the Workshop on Computational Approaches to Arabic Script-based Languages*, 2004, pp. 35–41.

[15] D. Jurafsky, J. Martin, A. Kehler, K. Vander Linden, and N. Ward, *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition*. MIT Press, 2010, vol. 163.

[16] B. Sagot and G. Walther, "A morphological lexicon for the Persian language," in *in Proceedings of the 7th Language Resources and Evaluation Conference (LREC'10)*, 2010, pp. 300–303.

[17] P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens *et al.*, "Moses: Open source toolkit for statistical machine translation," in *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, 2007, pp. 177–180.

[18] P. Brown, V. Pietra, S. Pietra, and R. Mercer, "The mathematics of statistical machine translation: Parameter estimation," *Computational linguistics*, vol. 19, no. 2, pp. 263–311, 1993.

[19] F. Och and H. Ney, "A systematic comparison of various statistical alignment models," *Computational linguistics*, vol. 29, no. 1, pp. 19–51, 2003.

[20] A. Stolcke, "SRILM-an extensible language modeling toolkit," in *Seventh International Conference on Spoken Language Processing*, vol. 3, 2002, pp. 901–904.

[21] A. Axelrod, "Factored Language Model for Statistical Machine Translation," Master's thesis, Institute for Communicating and Collaborative Systems, Division of Informatics, University of Edinburgh, 2006.

[22] P. Koehn, "MOSES, Statistical Machine Translation System, User Manual and Code Guide," 2010.

[23] M. Bijankhan, "Naghshe Peykarehaye Zabani dar Neveshtane Dasture Zaban: Mo'arrefiye yek Narmafzare Rayane'i [the Role of Corpus in generating grammar: Presenting a computational software and Corpus]," *Iranian Linguistic Journal*, vol. 19, pp. 48–67, 2006.

[24] D. Nicholls, "The Cambridge Learner Corpus-error coding and analysis for lexicography and ELT," in *Proceedings of the Corpus Linguistics 2003 conference*, 2003, pp. 572–581.

[25] J. Wagner, J. Foster, and J. Van Genabith, "A comparative evaluation of deep and shallow approaches to the automatic detection of common grammatical errors," *Proceedings of EMNLP-CoNLL-2007*, 2007.

[26] M. Miłkowski, "Automated building of error corpora of Polish," *Corpus Linguistics, Computer Tools, and ApplicationsState of the Art. PALC 2007*, pp. 631–639, 2008.

[27] B. Mohit and R. Hwa, "Localization of difficult-to-translate phrases," in *Proceedings of the Second Workshop on Statistical Machine Translation*, 2007, pp. 248–255.

[28] A. Arppe, "Developing a grammar checker for Swedish," in *The 12th Nordic Conference of Computational Linguistics*, 2000, pp. 13–27.

 ISBN: 978-1-61208-139-7

Background Speech Cancellation using a Generalized Subspace Speech Enhancement Method

Radu Mihnea Udrea, Constantin Paleologu, Silviu Ciochina
Telecommunications Department
"Politehnica" University of Bucharest
Bucharest, Romania
mihnea@comm.pub.ro, pale@comm.pub.ro, silviu@comm.pub.ro

***Abstract*— This paper presents a speech enhancement method for reducing undesired babble noise, or background speech, that affects the desired speech, using a generalized subspace approach. The subspace decomposition is obtained with a nonunitary transform based on diagonalization of the clean speech and background distortion covariance matrices. The clean signal is estimating using an optimal subspace estimator that nulls the signal components in the distortion signal subspace and keeps the components in the signal subspace. Objective and subjective measures show a better suppression of background speech that other subspace-based methods that were proposed for white noise.**

Keywords-speech enhancement; colored noise; subspace

I. INTRODUCTION

Over the years, many applications of acoustic noise reduction and speech enhancement require high performance and efficient algorithms. Spectral subtraction [1] is perhaps one of the most popular speech enhancement algorithm due to its low complexity. Even if several methods were proposed [2], [3] to reduce speech distortions and "musical noise" introduced by this algorithm, still there is a compromise to be made between reducing speech distortion and reducing residual noise.

Another approach of more recent speech enhancement algorithms is based on decomposition of the noisy signal in two subspaces: signal subspace and noise subspace. An estimate of the clean signal can be made by nulling the components of the signal in the noise subspace and retaining only the components of the signal in the signal subspace. The subspace decomposition can be done using the eigenvalue decomposition (EVD) [4]-[6] or the singular value decomposition (SVD) [7].

In [4], an optimal estimator that minimizes the speech distortion subject to the constraint that the residual noise fell below a preset threshold is proposed using the eigenvalue decomposition of the covariance matrix. The decomposition of the vector space of the noisy signal into a signal and noise subspace can be obtained by applying the Karhunen–Loéve transform (KLT) to the noisy signal. The KLT components representing the signal subspace were modified by a gain function determined by the estimator, while the remaining KLT components representing the noise subspace were nulled. The enhanced signal was obtained from the inverse KLT of the modified components. This subspace approach was based on the assumption that the input noise was white.

The work in [4] was extended for colored noise. In [5] it is given a proper noise shaping for colored noise without prewhitening, first by classifying the noisy speech frames into speech-dominated and noise-dominated frames and then using a different KLT matrix for these frames to construct the estimator. In [6] a generalized subspace approach with built-in prewhitening for enhancing speech corrupted with colored noise is determined.

In this paper we propose a nonunitary transform, based on the simultaneous diagonalization of the clean speech and the background distortion covariance matrices. No assumptions were made about the covariance matrix of the KLT-transformed noise vectors, hence this estimator is optimal.

This paper is organized as follows. In Section II, the subspace approach using time-domain constraints is presented for white noise and for colored noise (like babble talk). Section II also gives an expression, different that in [6], for the subspace estimator for any type of distortion signal which is uncorrelated to speech. Implementation details are provided in Section III, experimental results are given in Section IV, and the conclusions are given in Section V.

II. SUBSPACE APPROACH FOR SPEECH ENHANCEMENT

The linear model for the clean speech signal assumes that each *K*-dimensional vector *x* can be represented as:

$$\mathbf{x} = \sum_{m=1}^{M} s_m \mathbf{b}_m, \quad M < K \tag{1}$$

where $\{s_1, \ldots, s_M\}$ are zero mean random variables, and $\mathbf{b}_1, \ldots, \mathbf{b}_M$ are *K*-dimensional complex basis vectors, which

 ISBN: 978-1-61208-139-7

are assumed linearly independent. For speech signals, such representation is possible also, when $M<K$.

The model (1) can be expressed as:

$$\mathbf{x} = \mathbf{B} \cdot \mathbf{s} \tag{2}$$

where $\mathbf{B}$ is a $K \times M$ matrix whose rank is M and $\mathbf{s}$ is an M-dimensional vector. The covariance matrix of $\mathbf{x}$ is given by:

$$\mathbf{R}_{\mathbf{x}} \triangleq E\left\{\mathbf{x} \cdot \mathbf{x}^T\right\} = \mathbf{B} \cdot \mathbf{R}_{\mathbf{s}} \cdot \mathbf{B}^T \tag{3}$$

where $\mathbf{R}_{\mathbf{s}}$ is the covariance matrix of the vector $\mathbf{s}$, which is assumed positive definite. Hence, the rank of $\mathbf{R}_{\mathbf{x}}$ is M, and it has $K - M$ zero eigenvalues.

Let $\mathbf{d}$ being the K-dimensional vector of the noise (distortion) signal. Assuming the distortion signal is additive and uncorrelated with the speech signal, we can write the corrupted signal as:

$$\mathbf{y} = \mathbf{B} \cdot \mathbf{s} + \mathbf{d} = \mathbf{x} + \mathbf{d} \tag{4}$$

where $\mathbf{y}$ is the K-dimensional corrupted speech vector.

The clean speech linear estimator will be:

$$\hat{\mathbf{x}} = \mathbf{H} \cdot \mathbf{y} \tag{5}$$

where $\mathbf{H}$ is a $K \times K$ matrix. The error signal resulted from this estimation is given by:

$$\varepsilon = \hat{\mathbf{x}} - \mathbf{x} = (\mathbf{H} - \mathbf{I}) \cdot \mathbf{x} + \mathbf{H} \cdot \mathbf{d} = \varepsilon_x + \varepsilon_d \tag{6}$$

Let

$$\begin{aligned} \overline{\varepsilon_x^2} &= E\left[\varepsilon_x^T \varepsilon_x\right] = tr\left(E\left[\varepsilon_x \varepsilon_x^T\right]\right) \\ \overline{\varepsilon_d^2} &= E\left[\varepsilon_d^T \varepsilon_d\right] = tr\left(E\left[\varepsilon_d \varepsilon_d^T\right]\right) \end{aligned} \tag{7}$$

be the energy of the speech distortion and, respectively, the energy of the residual noise vector. The linear estimator can be obtained [4] by solving the following time-domain constrained (TDC) optimization problem:

$$\begin{aligned} &\min_H \overline{\varepsilon_x^2} \\ &\text{subject to: } \frac{1}{K}\overline{\varepsilon_d^2} \le \alpha \sigma_d^2 \end{aligned} \tag{8}$$

where $0 \le \alpha \le 1$. The estimator derived in this way minimizes the signal distortion over all linear filters which result in the permissible residual noise level. The solution to (8) is given by [4]:

$$\mathbf{H}_{opt} = \mathbf{R}_{\mathbf{x}} \left(\mathbf{R}_{\mathbf{x}} + \mu \mathbf{R}_{\mathbf{d}}\right)^{-1} \tag{9}$$

where $\mathbf{R}_{\mathbf{x}}$ and $\mathbf{R}_{\mathbf{d}}$ are the covariance matrices of the clean speech and noise respectively, and μ is the Lagrange multiplier.

Consider the eigen-decomposition of $\mathbf{R}_{\mathbf{x}}$

$$\mathbf{R}_{\mathbf{x}} = \mathbf{U} \mathbf{\Lambda}_{\mathbf{x}} \mathbf{U}^T \tag{10}$$

where $\mathbf{U}$ is the eigenvector unitary matrix and $\mathbf{\Lambda}_{\mathbf{x}}$ is the diagonal eigenvalue matrix of $\mathbf{R}_{\mathbf{x}}$.

The optimal filter from (9) can be simplified using (10) to:

$$\mathbf{H}_{opt} = \mathbf{U} \mathbf{\Lambda}_{\mathbf{x}} \left(\mathbf{\Lambda}_{\mathbf{x}} + \mu \mathbf{U}^T \mathbf{R}_{\mathbf{d}} \mathbf{U}\right)^{-1} \mathbf{U}^T. \tag{11}$$

A. White Noise Subspace Estimator

For white noise with variance σ_d^2, $R_d = \sigma_d^2 I$ and the estimator from (11) reduces to White Noise Subspace Estimator (WNSE) [4]:

$$\mathbf{H}_{WNSE} = \mathbf{U} \mathbf{\Lambda}_{\mathbf{x}} \left(\mathbf{\Lambda}_{\mathbf{x}} + \mu \sigma_d^2 \mathbf{I}\right)^{-1} \mathbf{U}^T = \mathbf{U} \cdot \mathbf{G}_{WNSE} \cdot \mathbf{U}^T. \tag{12}$$

where

$$\mathbf{G}_{WNSE} = \mathbf{\Lambda}_{\mathbf{x}} \left(\mathbf{\Lambda}_{\mathbf{x}} + \mu \sigma_d^2 \mathbf{I}\right)^{-1}. \tag{13}$$

The gain matrix $\mathbf{G}_{WNSE}$ is diagonal with elements (gains):

$$g_{WNSE}(m) = \frac{\lambda_x(m)}{\lambda_x(m) + \mu \sigma_d^2}. \tag{14}$$

Hence, the signal estimate is obtained by applying the Karhunen-Loève transform (KLT) to the noisy signal, then modify the components of the KLT by a gain function and finally, by inverse KLT of the modified components. A block diagram of this estimator is shown in Fig. 1.

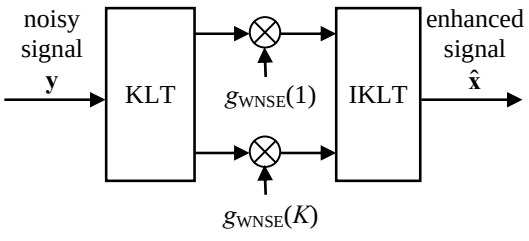


Figure 1. Signal subspace linear estimator

B. Colored Noise Subspace Estimator

If the distortion is not white noise, the matrix $\mathbf{U}^T\mathbf{R_d}\mathbf{U}$ is not diagonal since **U**, being the eigenvector matrix of the symmetric matrix $\mathbf{R_x}$, diagonalizes $\mathbf{R_x}$ and not $\mathbf{R_d}$. There is a matrix **V** that simultaneously diagonalize $\mathbf{R_x}$ and $\mathbf{R_d}$.

As stated in [6], consider the basis matrix $\mathbf{\Sigma}$ satisfying the following equations:

$$\begin{aligned}\mathbf{\Sigma} &= \mathbf{R_d^{-1}R_x}\\ \mathbf{\Sigma V} &= \mathbf{V\Lambda}\end{aligned} \tag{15}$$

where $\mathbf{\Lambda}$ and **V** are the eigenvalue matrix and eigenvector matrix respectively of $\mathbf{\Sigma}$. Applying diagonalizing matrix **V** will result the fully diagonal eigenvalues matrices of $\mathbf{R_x}$ and $\mathbf{R_d}$:

$$\begin{aligned}\mathbf{V}^T\mathbf{R_x V} &= \mathbf{\Lambda_x} \neq \mathbf{\Lambda}\\ \mathbf{V}^T\mathbf{R_d V} &= \mathbf{\Lambda_d}\end{aligned} \tag{16}$$

where $\mathbf{\Lambda_x}$ and $\mathbf{\Lambda_d}$ are the eigenvalue matrices of $\mathbf{R_x}$ and $\mathbf{R_d}$, respectively.

The resulted equations (16) are more general than the relations in [6] (where it is considered that $\mathbf{\Lambda_x} = \mathbf{\Lambda}$ and $\mathbf{\Lambda_d} = \mathbf{I}$). The approach proposed in (16) allows applying the subspace method to any type of distortion signal which is uncorrelated to speech signal.

Applying the eigen-decomposition of $\mathbf{\Sigma}$ from (15) and using (16), the optimal linear Colored Noise Subspace Estimator (CNSE) can be expressed as:

$$\begin{aligned}\mathbf{H}_{CNSE} &= \mathbf{R_d V\Lambda_x}\left(\mathbf{\Lambda_x} + \mu\mathbf{\Lambda_d}\right)^{-1}\mathbf{V}^T =\\ &= \mathbf{V}^{-T}\cdot\mathbf{G}_{CNSE}\cdot\mathbf{V}^T\end{aligned}. \tag{17}$$

where

$$\mathbf{G}_{CNSE} = \mathbf{\Lambda_x}\left(\mathbf{\Lambda_x} + \mu\mathbf{\Lambda_d}\right)^{-1}. \tag{18}$$

In case of the colored noise, the corrupted signal is decorrelated with the non-KLT matrix $\mathbf{V}^T$, then it is modified by the signal subspace gain matrix $\mathbf{G}_{CNSE}$, and, finally, the enhanced signal estimate is obtain by the inverse non-KLT matrix $\mathbf{V}^{-T}$.

Since we have no access to the covariance matrix $\mathbf{R_x}$ of the clean speech signal, the matrix $\mathbf{\Sigma}$ is estimated from the noisy speech signal. Assuming that speech is uncorrelated with noise, we have

$$\mathbf{R_y} = \mathbf{R_x} + \mathbf{R_d}. \tag{19}$$

and so

$$\mathbf{\Sigma} = \mathbf{R_d^{-1}R_x} = \mathbf{R_d^{-1}R_y} - \mathbf{I}. \tag{20}$$

The estimation of μ in the gain function (14) or (18) affects the quality of speech. A large value of μ would reduce the residual noise but would introduce speech distortion. A small value of μ would minimize the speech distortion at the expense of higher values of residual noise. A trade-off between residual noise and speech distortion can be obtained by making μ dependent on the short-time SNR:

$$\mu = \mu_0 - \frac{SNR_{dB}}{s_0}. \tag{21}$$

where μ_0 and s_0 are constants chosen experimentally [6] as explained in the implementation section.

III. ALGORITHM IMPLEMENTATION

The proposed algorithm can be implemented, for each speech frame, as follows:

- The distortion covariance matrix $\mathbf{R_d}$ is computed prior to the starting of the speech signal during speech-absent frames.
- The matrix $\mathbf{\Sigma}$ is estimated using (20) from the noisy signal covariance matrix $\mathbf{R_y}$ and the inverse of $\mathbf{R_d}$.
- The eigen-decomposition of $\mathbf{\Sigma}$ is performed using (15). Extract the eigenvector matrix **V** and eigenvalue matrix $\mathbf{\Lambda}$.
- The dimension of the speech signal subspace is estimated, considering that the eigenvalues of $\mathbf{\Sigma}$ are ordered $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_K$, from:

$$M = \arg\max_{1\leq k\leq K}\left\{\lambda_k > 0\right\}. \tag{22}$$

- The μ factor is computed as a linear function of SNR [6]:

$$\mu = \begin{cases} 5 & SNR_{dB} < -5\\ \mu_0 - \dfrac{SNR_{dB}}{s_0} & -5 < SNR_{dB} < 20\\ 1 & SNR_{dB} \geq 20\end{cases}. \tag{23}$$

  where $\mu_0 = 4.2, s_0 = 6.25, SNR_{dB} = 10\log_{10} SNR$.
- SNR can be computed directly from the eigenvalues λ_k of $\mathbf{\Sigma}$ using the following equation [6]:

$$SNR = \frac{tr\left(\mathbf{V}^T\mathbf{R_x V}\right)}{tr\left(\mathbf{V}^T\mathbf{R_d V}\right)} = \frac{\sum_{k=1}^{M}\lambda_k}{K}. \tag{24}$$

 ISBN: 978-1-61208-139-7

- Compute the optimal estimator $\mathbf{G}_{CNSE}$ using (18) and estimate the desired signal using (5).

The covariance matrices $\mathbf{R}_x$ and $\mathbf{R}_d$ were estimated as Toeplitz matrices using K samples of the unbiased autorcorrelation sequence, without using future or past frames. We choose $K = 40$ samples for speech sampled at 8 kHz. The estimators were applied to frames of the corrupted signal 50% overlapped each other. The covariance matrices were estimated by windowing with rectangular windows. The enhanced speech signal estimation was obtained using overlap and add approach with Hamming windowing.

IV. EXPERIMENTAL RESULTS

We used 20 sentences produced by 10 male and 10 female speakers. For distortion signal we used other speech signals added as babble noise to the clean speech at SNR = 5 dB. For comparative purposes, we also evaluated the algorithm performances applying as a distortion signal white noise at the same SNR.

The Perceptual Evaluation for Speech Quality (PESQ) distance measure and the overall (global) SNR [8] measures were adopted for evaluation of the proposed algorithms. We used the ITU-T Recommendation P.862 (PESQ) [9] to obtain a perceptual evaluation of the enhanced speech quality. The Mean Opinion Score (MOS) obtained in the evaluation process is between 0 and 5 where 0 represents a very annoying distortion of the perceived signal and 5 represents imperceptible quality degradation.

TABLE I. MEAN PESQ AND MEAN GLOBAL SNR FOR WHITE NOISE DISTORTION AT 5dB

	Male Speakers		**Female Speakers**	
	SNR	*PESQ*	*SNR*	*PESQ*
Noisy Speech	4.6 dB	1.78	4.8 dB	1.71
WNSE	11.1 dB	2.36	10.9 dB	2.21
CNSE	11.3 dB	2.47	10.8 dB	2.22

TABLE II. MEAN PESQ MEAN GLOBAL SNR FOR BABBLE SPEECH DISTORTION AT 5dB

	Male Speakers		**Female Speakers**	
	SNR	*PESQ*	*SNR*	*PESQ*
Noisy Speech	5.3 dB	0.72	5.2 dB	0.69
WNSE	6.7 dB	1.06	6.9 dB	0.81
CNSE	7.3 dB	1.45	7.1 dB	1.42

Tables I and II give the mean results for 20 TIMIT sentences corrupted by speechshaped noise at 5 dB. The results are given separately for male and female speakers. As can be seen from Tables I and II, in case of speech corrupted of white noise at 5dB SNR, the proposed approach reduces to Ephraim and Van Trees approach [4]. In case of speech corrupted by background babble talk distortion, our proposed approach (CNSE) outperformed Ephraim and Trees approach [4] for both male and female speakers.

Subjective listening tests confirmed the results in Tables I and II and that with the proposed method, the background noise was imperceptible. Since in our experiments, no voice detection algorithm (VAD) was used to update the noise covariance matrix, we expect further improvements in performance if we use a reliable VAD algorithm to update the noise covariance matrix.

V. CONCLUSIONS

A speech enhancement for reducing undesired babble noise, or background speech, that affects the desired speech, using a generalized subspace approach was proposed. The proposed approach is based on the simultaneous diagonalization of the covariance matrices of the speech signal and the colored noise signal. In case of the colored noise, the corrupted signal is decorrelated with the non-KLT matrix, it is modified by a gain matrix, and, finally, the enhanced speech is estimated by inverse non-KLT matrix. Better SNR and perceptual scores were obtained that applying the standard KLT decomposition used by Ephraim and Van Trees [3] for enhancing speech corrupted by white noise.

ACKNOWLEDGMENT

The work has been funded by the Sectoral Operational Programme Human Resources Development 2007-2013 of the Romanian Ministry of Labour, Family and Social Protection through the Financial Agreement POSDRU/89/1.5/S/62557.

REFERENCES

[1] M. Berouti, R. Schwartz, and J. Makhoul, “Enhancement of speech corrupted by acoustic noise,” in Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, pp. 208–211, 1979.

[2] R. Martin, “Spectral subtraction based on minimum statistics”, Proc. Eur. Signal Process., pp. 1182-1185, 1994.

[3] R. M. Udrea, N. Vizireanu, S. Ciochina, and S. Halunga “Nonlinear spectral subtraction method for colored noise reduction using multi-band Bark scale”, Signal Processing, Volume 88 Issue 5, Elsevier North-Holland, Inc., pp. 1299-1303, May 2008.

[4] Y. Ephraim and H. L. Van Trees, “A signal subspace approach for speech enhancement,” IEEE Trans. Speech Audio Processing, vol. 3, pp. 251-266, 1995.

[5] U. Mittal and N. Phamdo, “Signal/noise KLT based approach for enhancing speech degraded by colored noise,” IEEE Trans. Speech Audio Processing, vol. 8, pp. 159–167, Mar. 2000.

[6] Y. Hu and P. C. Loizou, “A generalized subspace approach for enhancing speech corrupted by colored noise,” IEEE Transactions on Speech and Audio Processing, vol. 11, no. 4, pp. 334-341, 2003.

[7] S. H. Jensen, P. C. Hansen, S. D. Hansen, and J. A. Sørensen, “Reduction of broad-band noise in speech by truncated QSVD,” IEEE Trans. Speech Audio Processing, vol. 3, pp. 439–448, Nov. 1995.

[8] J. R. Deller, J. Hansen, and J. G. Proakis, Discrete-Time Processing of Speech Signals. New York: IEEE Press, 2000.

[9] ITU-T, Perceptual evaluation of speech quality PESQ, an objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs, ITU-T Recommendation P.862, 2000.

 ISBN: 978-1-61208-139-7

University Timetabling Algorithm Considering Lecturer's Workload

Lintang Yuniar Banowosari[1], Vega Valentine[2]
[1]Department of Computer Science and Technology
[2]Department of Industrial Technology
University of Gunadarma
Jakarta, Indonesia
[1]lintang@staff.gunadarma.ac.id, [2]valentvga@gmail.com

***Abstract*—University Timetabling Problem is an allocation or subject to constraints, of given resources being placed in space time, in such a way as to satisfy as nearly as possible a set of desirable university schedule requirements. In this paper, university timetabling algorithm is implemented, considering lecturer's workload in order to have a balance between lecturer's workload as a teaching staff of the university and to actualize the obligation of *Tridharma Perguruan Tinggi,* regulation issued by Indonesia's Ministry of Education. The implementation of faculty timetabling, the workloads summation and the lecture-class timetabling has successfully built in Java Netbeans Swing GUI.**

***Keywords - University Timetabling Problem; lecturer's workload; university schedule requirements; university timetabling algorithm;* Tridharma Perguruan Tinggi; *Java Netbeans Swing GUI*.**

I. INTRODUCTION

Scheduling is a process or a way of organize time according to arrangement of work order plan. It also means a list or activity table or activity plan with a detailed execution time [1]. In university terminology, this scheduling problem is known as University Timetabling Problem.

Every university has their own studying activities organized in such a way to satisfy any requirements they need. In schedule arrangement, universities tend to have a system which can schedule all courses optimally. To have such optimal condition of the course, a well-organized of all scheduling components is needed.

A scheduling system is also the core of university activities because it involves many elements in affiliations to the university, that is human resources (lecturers and students), time slot availability (length of lecture), type of the activity (theory or lab practice), and the facility to support those activities (classroom or laboratory) [6].

Bardadym (2006) classified the university timetabling into five groups [4], they are:

- *Faculty timetabling,* assigns qualified teachers to courses
- *Class-Teacher timetabling*, assigns courses with the smallest timetabling unit being a class of students
- *Course Scheduling*, assigns courses with the smallest scheduling unit being an individual student
- *Examination Scheduling*, assigns examination to students such that students do not have two examinations at the moment
- *Classroom Assignment*, assigns class-teacher couples to classrooms

In fact, there are many algorithms used to organize schedule in university timetabling. Algorithms such as Genetic Algorithm, Simulated Annealing and Tabu Search, are commonly used in university timetabling research.

But, which one is the best algorithm to do university timetabling? This question cannot be generally answered, because the problem is highly institution-specific. Every university has its own way in manage scheduling, with different requirements and regulations. In other words, managing timetable will be dependent on what regulation they hold and what requirements they need.

That is why no specific answer for the question. The best solution will be an algorithm that violate the least constraint or satisfy the most requirements or preferences for a certain university regulation.

One of the constraints in doing timetabling is nonetheless the activity of the lecturer itself because teaching is not always their only activity. Some regulations, such as the one issued from the government, obligate lecturers to do other things in order to dedicate and contribute more in education. In this paper, *Tridharma Perguruan Tinggi*, issued by Education Ministry of Indonesia, is taken as reference in defining activities of lecturer which will lead to some calculations to obtain optimal university timetable.

The outline of this paper is: Section II explains theory of timetabling, Section III describes methodology of workload calculation, and Section IV shows design and implementation of the algorithm.

 ISBN: 978-1-61208-139-7

II. THEORETICAL FRAMEWORK

A. University Timetabling Problem (UTP)

University Timetabling Problem is an allocation or subject to constraints, of given resources that is human resources (lecturers and students), time slot availability (i.e. length of lecture), type of the activity (theory or lab practice), and the facility to support those activities (classroom or laboratory) being placed in space time, in such a way as to satisfy as nearly as possible a set of desirable university schedule requirements.

Edmund Burke in his article titled '*Applications to Timetabling*' [6] specified the timetabling problem as a problem with four parameters, *T* (a finite set of times), *R* (a finite set of resources), *M* (a finite set of meeting) and *C* (a finite set of constraints):

1. *Times*
   A time *t* is an element of the set of times *T* of an instance of the timetabling problem. A time slot is a variable constrained to contain one time.
2. *Resources*
   A resource *r* is an element of the set of resources *R* of an instance of the timetabling problem. A resource slot is a variable constrained to contain one resource. What we called resources are teachers, rooms, items of special equipment, students or group of students that supports a meeting.
3. *Meeting*
   A meeting *m* is a named collection of time slots and resource slot. Assigning values to those slots means that all of the assigned resources attend this meeting at all of the assigned times.
4. *Constraints*
   Constraints divided into two, hard constraint and soft constraint. Hard constraint must be satisfied while soft constraint is desirable, but not necessary, to satisfy—more to optimization objective.
   In university course timetabling, no-clashes constraint would typically be a hard constraint for lecturers but a soft constraint for student as far as optional courses are concerned since it usually impossible to satisfy every student.

B. Algorithms to Solve University Timetabling Problem

Algorithms had been developed and implemented in building a timetable for universities. Literatures about university course timetabling teach us that researchers applied different approaches to tackle the problem [4]. Above many algorithms, there are three most applicable and most widely used meta-heuristic algorithm to make an optimal university timetabling:

1. *Simulated Annealing*

Simulated annealing is a probabilistic method proposed in Kirkpatrick, Gellat, and Vecchi (1983) and Cerny (1985) for finding the global minimum of a cost function that may possess several local minima. It works by emulating the physical process whereby a solid is slowly cooled so that when eventually its structure is 'frozen', this happens at a minimum energy configuration [8].

Simulated annealing started with making the mathematical formulation of the problem that is the hard and soft constraints. After that, define properties of the constraints such as teaching duration, available class, etc. Then a lecture initially placed onto available timeslot.

Energy function, cooling and acceptance probability function also applied. The energy function is derived from the main timetabling objective (considering times, meeting, resource, and constraints), while cooling schedule and acceptance probability function controls accepting new solution with certain energy value. These two functions used to reach the objective of building optimized university timetable.

2. *Genetic Algorithm*

Genetic Algorithm was founded by John Holland in Michigan University, United State (1975) through some researches and David Goldberg introduced [9].

Three main aspects in genetic algorithm are definitions of fitness function, implementation of genetic representation and genetic operation. If the three aspects are defined, then the algorithm will be well-performed.

The algorithm started with a set of randomly selected state called population. Each state defined as a string. It combines two main parent populations. Through some crossover, mutation and fitness function, new children population will be defined as the solution.

3. *Tabu Search*

The basic concept of Tabu Search as described by Glover (1986) is 'a meta-heuristic superimposed on another heuristic' or a higher-level meta-heuristic procedure for solving discrete and continuous optimization problems.

The overall approach is to avoid entrainment in cycles by forbidding or penalizing moves which take the solution, in the next iteration, to points in the solution space previously visited. The solution space that has been visited therefore listed as 'tabu' [10].

Three main strategies of tabu search are [11]:

- Forbidding strategy, control what enters the tabu list
- Freeing strategy, control what exits the tabu list and when
- Short-term strategy, manage interplay between the forbidding strategy and freeing strategy to select trial solutions

 ISBN: 978-1-61208-139-7

Between those three algorithms, Tunçhan CURA, Istanbul University research group, compare the performance by modify those three algorithms [5] into a similar structure design to be proper with the IUFBA requirements. The proposed algorithm has been tested with the 2006-2007 academic year, first term course timetabling data of IUFBA.

From the comparison, he found that simulated annealing has been the best algorithm to solve university timetabling problem. He concludes this based on his experiments in having the three algorithms to do the same case.

Thus for the case of lecturer's workload, simulated annealing algorithm will be used and the equation performed by [5] will be modified to satisfy requirements as explained in the next section.

C. Lecturer's Workload

Lecturers stated as a professional educator and a scientist whose prime objective is to transform, develop, and publish knowledge, technology, and art through education, research, and dedication to public [7].

In Indonesia, lecturer's performance of education always obeying the rule of *Tridharma Perguruan Tinggi* which consists of three dharma. They are Dharma of Education and Teaching, Dharma of Research and Dharma of Public Dedication. Detailed description explained on the calculation part (section three).

III. METHODOLOGY

A. Formulation of the Problem

The following subsection explains the process of listing obligatory rules and constraints and the mathematical formulation of rules and constraints defined.

1. Defining Obligatory Rules, Hard and Soft Constraints

The obligatory rules that generally overdue in universities are:

- No. 1: Each lecture must be assigned to only one class of student at one day and to a single time slot
- No. 2: The lengths of the lectures hours must be taken into consideration while assigning the lectures. For example if the lecture hours are from 9 am to 5 pm and the length of the lecture is 2 hours, this lecture cannot be assigned to 4 pm since it would have exceeded the official lecture hours
- No. 3: More than one lecture cannot be assigned to a given class at the same time interval
- No. 4: A lecturer cannot have more than one lecture assigned in a given time interval

The hard and soft constraints defined as seen in table I.

TABLE I. HARD AND SOFT CONSTRAINTS

Hard Constraints	Soft Constraints
No resources (lecturer and a class of students) may be assigned to different events at the same time	Every lecturer has his/her own availability schedule or submits a plan with desirable time periods that suits him/her best
There is a maximum number of time periods per day, that may not be exceeded	Every lecturer has a minimum and a maximum limit of weekly work-hours
More than one lecture can not be assigned to a given class at the same time slot	Minimize the time gaps within the schedule of each lecturer
Each lecture may be assigned to a lecturer that belongs to a specific set of lecturers that can deliver the lecture	Minimize the time gaps within the schedule of each given class

2. Mathematical Formulation of the Problem

Meeting of lectures, lecturers and rooms that available, denoted by *J, I* and *L* respectively. Lectures can be assigned to any lecture day from Monday to Saturday. Each day consists of 10 hours. Thus, $D = 6$, $H = 10$, denote the number of days and hours of timetable. Thus, the rules will be denoted as follow:

- The general mathematical model for satisfying the lecturer desires represented as:

$$\max \sum_{j=1}^{J}\sum_{i=1}^{I}\sum_{d=1}^{D}\sum_{h=1}^{H}\sum_{l=1}^{L}\sum_{h^*=h}^{\min\{(Y_j+h),H\}} X_{ji} \times P_{idh} \times C_i \times S_{jldh} \quad (1)$$

- Obligatory rule 1 is imposed by:

$$\sum_{l=1}^{L}\sum_{d=1}^{D}\sum_{h=1}^{H} S_{jldh} = 1, j = 1,...,J \quad (2)$$

- Obligatory rule 2 is imposed by:

$$\beta_{jldh} \le H \quad j = 1,...J; h = 1,..., H; l = 1,..., L; d = 1,..., D \quad (3)$$

- Obligatory rule 3 is imposed by:

$$\beta_{jldh} \times S_{jldh} \le \beta_{j^*ldh^*} \quad (4)$$

- Obligatory rule 4 is imposed by:

$$\beta_{jldh} \times S_{j^*l^*dh^*} \times X_{j^*i} \times X_{ji} \le \beta_{j^*ldh^*} \quad (5)$$

Y_j denotes the length of lecture j ($j = 1,..., J$). X_{ji} is a class of students with defined lecture j and lecturer i ($i = 1,..., I$). P_{idh} denotes the desire time slot (a higher value indicating a higher preference) of lecturer i for day d ($d = 1,..., D$) and hour h ($h = 1,..., H$). C_i denotes the lecturer's workload of lecturer i ($i = 1,..., I$). S_{jldh} is space for lecture in the timetable. β_{jldh} is a lecture with defined length in hour (duration).

B. Lecturer's Workload Calculation

Based on *Lampiran II Surat Dirjen Dikti No. 3298/D/T/99* issued on 29 Desember 1999 [7] about lecturer's workload evaluation, the details of workload calculation is described on table II.

TABLE II. LECTURER'S WORKLOAD IN DETAIL

No	Activity	Hour/week	Multiplier Notation
A	***Education***		
1.	Give a lecture 'X' (y credits)	y	$\sum class$
2.	Assess final examination	0.5	$\sum exam$
3.	Assess thesis defense for 3 students	0.5	$(\sum student) \div 3$
4.	Thesis consultation to a student	2	$\sum student$
5.	Student academic adviser for 20 students	1	$(\sum student) \div 20$
B.	***Research***		
1.	Make one research topic per year (as main researcher)	10	$\sum research$
2.	Writing papers to accredited journal, a title per 2 year (as main author)	1	$\sum paper$
C.	***Public Dedication***		
	Giving a workshop for 1 topic per semester	1	$\sum workshop$
D.	***Supporting Activities***		
1.	Active in a committee during a year	1	$\sum committee$
2.	Attend campus event (seminars, meetings, etc)	0.5	$\sum event$

Table III describes the maximum workload can be hold by a lecturer according to ministry's regulation.

TABLE III. WORKLOAD CALCULATION

No	Activity (appropriate to ideal lecturer's workload)	hour/week
A	***Education***	
1.	Teaching a lecture 'X' (3 credits)	9
2.	Teaching a lecture 'Y' (3 credits)	9
3.	Giving consultation to students for (undergraduate) thesis, 3 student per semester	6
4.	Student advisor for 20 students per semester	1
5.	Assessing final examination or (undergraduate) thesis defense , 3 students per semester	0.5
6.	Making one course dictate per year	2
	Total of A	**27.5**
B.	***Research***	
1.	One research topic per year, as the main researcher	10
2.	Writing papers to accredited journal, a title per 2 year as main author	1
	Total of B	**11**
C.	***Public Dedication***	
	Giving a workshop for 1 topic per semester	1
D.	***Supporting Activities***	
	Active in a committee during a year	1
	Sum of Total	**40.5**

C. Timetabling Solver

1. Defining the Number of b Vectors

Let X be the number of different lecture lengths. Thus, each different length, the number of b_k where $hour_k$ equals this length are denoted by λ_x, δ_x, and μ_x ($x = 1,\ldots, X$) respectively.

For example, if there are 3 lectures and their lengths are 2 hours, 2 hours and 3 hours respectively, then the number of different lengths (X) will be 2 ($\lambda_1 =$ 2 hours and $\lambda_2 = 3$ hours), and δ_1 will be 2 and δ_2 will be 1. The number, K, of b vectors imposed by equation (6).

$$\sum_{x=1}^{X} \mu_x \qquad (6)$$

For this study, the sample data was taken from Gunadarma University's Faculty of Psychology for 4th grade class in ATA 2008/2009. For this sample, we got $X = 3$ with $\lambda_1 = 1$, $\lambda_2 = 2$ and $\lambda_3 = 3$. Thus by equation above, we got $K = 7$ with $\delta_1 = 1$, $\delta_2 = 5$ and $\delta_3 = 1$.

2. Filling the b Vectors with Lectures

In this study, the process of assigning defined lecture to b vectors using indirect representation. In such representation, the encoded solution usually represents an ordered list of events, which are placed into the timetable according to some predefined method, or so called timetable builder. The timetable builder can use any combination of heuristics and local search to place events into the timetable, while observing the problem's constraints.

For this work, the indirect representation encodes 3 fields for each event:

- Day and hour (time slot) to allocate the event
- Teacher (1 or more) to be assigned to the event
- Class of students that supposed to take the event

All fields are first encoded as integers and then converted into appropriate variable type for further process in the program. In generating the solution, the solver first decodes it to gain these four fields for every event in the schedule. Then it invokes the timetable builder to works as in Figure 1.

 ISBN: 978-1-61208-139-7

IV. DESIGN & IMPLEMENTATION

A. Lecturer's Workload Implementation in Timetabling Algorithm

In this study, lecturer's workload divided into two different workloads. First is teaching workload (`workload_teach`) and the second is administration workload (`workload_adm`). Teaching workload is the total workload of assigned course calculated by the amount of SKS per course for each class. Administration workload is the total workload of activity but teaching, which is defined in the *Tridharma*.

Pseudocode of Lecturer's Workload implementation to timetable is:

```
get lecturer's workload_adm
get lecturer's workload_teach
if workload_adm + workload_teaching < 40.5
     then put lecturer into S(d,h) matrix
     Do
        insert lecture into S(d,h)
        if any constraint violated
             then search subsequent S(d,h) until no
             violation
        else continue inserting to S(d,h)
     Until workload_teach = 0
else exceed lecturer's max workload
```

For describing how the algorithm works in such a real data, table IV is sample input case of lecturer's activity in a semester:

TABLE IV. LECTURER'S ACTIVITY AS INPUT TEST CASE

Administration Workload	Teaching Workload
Assess 3 final examination (A2)	Teach a lecture 'M' (2 credits) @ 3 classes (A1)
Give workshop for 3 topics this semester (C1)	Teach a lecture 'N' (1 credit) @ 6 classes (A1)
Write a paper to accredited journal (B2)	Teach a lecture 'P' (1 credit) @ 2 classes (A1)
Thesis consultation for 9 students (A3)	

From Table IV, we can calculate weights for administration and teaching workload as follow:

Administration Workload

1. Assess 3 final examination
   $0.5 \times \sum exam = 0.5 \times 3 = 1.5$
2. Give workshops for 3 topics
   $1 \times \sum workshop = 1 \times 3 = 3$
3. Write a paper to accredited journal
   $1 \times \sum paper = 1 \times 1 = 1$
4. Thesis consultation for 9 students
   $0.5 \times (\sum students) \div 3 = 0.5 \times (9 \div 3) = 1.5$

Total administration workload = 7 hours per week

Teaching Workload

1. Teach a lecture 'M' (2 credits) @ 3 classes
   $y \times \sum class = 2 \times 3 = 6$
2. Teach a lecture 'N' (1 credit) @ 6 classes
   $y \times \sum class = 1 \times 6 = 6$
3. Teach a lecture 'P' (1 credit) @ 2 classes
   $y \times \sum class = 1 \times 2 = 2$

Total teaching workload = 14 hours per week

From calculations above, we get the total result of 21 hours workload from administration workload plus teaching workload (7+14). The value is below the maximum workload of 40.5 hours per week. Thus, the lecturer can still be assigned to another teaching assignment through the timetable process (Figure 1) or other administration work. While for some other that reach the total workload of 40.5, they will have the opposite treatment such as workload reduction either from administrational or teaching assignment.

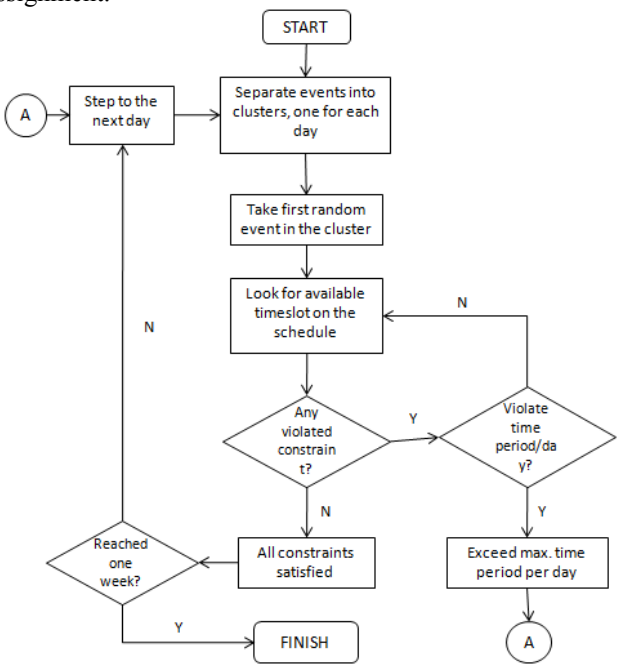


Figure 1. Timetable Builder Algorithm

The constraints involved so far are the hard and soft constraints as listed in Table I. According to the pseudocode, lecturers with maximum workload cannot be assigned to another event anymore. This condition verifies the soft constraint. The time slot is set to be a unique *S(d,h)* matrix including the unique day and hours per week. Therefore, an event-clash for related resources (lecturer and class of student) can be automatically avoided. The treatment for any other constraint will be the same, i.e. search for the next available *S(d,h)* slot.

B. Implementation on Java, Netbeans Swing GUI

University Timetabling Application is a desktop application which is developed to facilitate computerization

 ISBN: 978-1-61208-139-7

in solving the university timetabling problem. It is built in Java programming language using Netbeans Swing GUI for designing the graphical user interface and database, handled by PostgreSQL.

This application built based on previous object-oriented analysis through the system which also developed by the algorithm already explained. The analysis then visualized using Unified Modeling Language (UML), i.e. use case diagram, class and activity diagram (Figure 2)

From the analysis, this application would contain five modules. They are functioned to store subject's data, lecturer data, lecturer's activity, to assign lecturers to subjects and the timetable module.

The main parameter in this application is the credit of a subject, number of class which should get the related subject and the total credit taken by the lecturer. The total credit is calculated by number of credits and class (as explained in IV A) which provides total hours that should be taken by the lecturers. The total hours considered as the lecturer's workload and determine whether the lecturer can still be assigned to another event or not (Figure 1).

In Figure 3, the initial workload calculation of administrative work is calculated by module 'lecturer's activity' (Figure 3a), while the assignment to teach a subject organized by module 'lecture assignment' which shows the detailed parameter of subjects (subject's name, credit) and also the initial administrative workload (Figure 3b).

The first four modules are already set and work properly, while so far, the timetable module is still on progress.

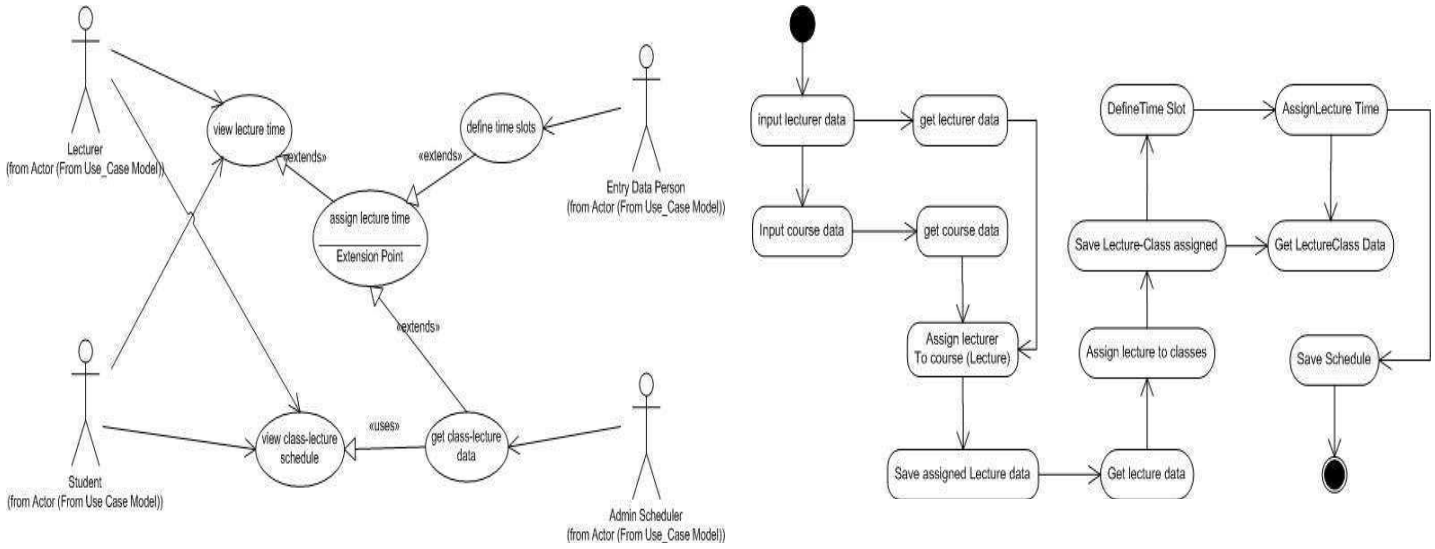


Figure 2. Unified Modeling Language for University Timetabling (a) Use Case for Assigning Lecture (b) Activity Diagram

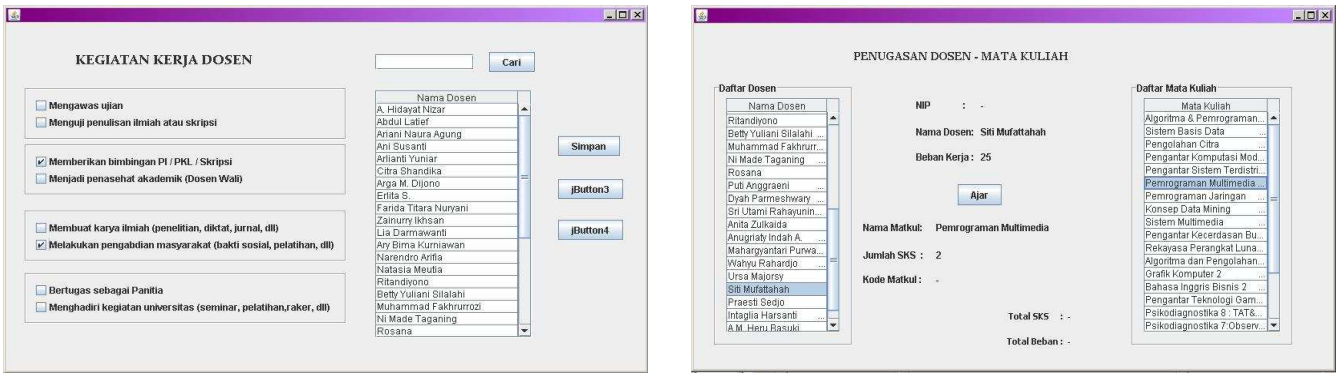


Figure 3. Screenshot of University Timetabling Application's GUI (a) Lecturer's Activity Module (b) Lecture Assignment Module

V. CONCLUSION

In solving university timetabling problem, three algorithms, Simulated Annealing, Genetic Algorithm, and Tabu Search had been theoretically studied. Simulated Annealing supports solving university timetabling problem with consideration of additional variable, such as lecturer's workload, therefore selected for this case.

Lecturer's activity had been categorized and being weighted. It applied to an input test case, simulating calculation of the lecturer's workloads, together results the output of calculation. This output will determine placement of schedule onto the timetable, obeying the obligatory rule, hard, and soft constraints.

However, the implementation of the algorithm using GUI Swing Netbeans has only reached the process of faculty timetabling, the workloads summation and the lecture-class timetabling. Further refinement needed to be done to get the optimal University Timetabling Application. Cooling function for this application is to be considered for the whole timetable because so far it only considers individual lecturer's workload.

REFERENCES

[1] D. Sugono, "Kamus Besar Bahasa Indonesia", Pusat Pembinaan dan Pengembangan Bahasa Indonesia, Departemen Pendidikan dan Kebudayaan, Balai Pustaka PN, 1993.

[2] V. Bardadym, "Computer-Aided School and University Timetabling: the New Wave", Selected and Revised Papers of the 1st International Conference on Practice and Theory of Automated Timetabling, (PATAT 1995), Edinburgh, Springer LNCS 1153, 22-45, 1996.

[3] B. Oestereich, "Developing Software with UML: Object-Oriented Analysis and Design in Practice", Second Edition, 2002, Edinburgh Gate, Harlow CM20 2JE, Pearson Education.

[4] A. Mieke, P. Causmaecker, P. Demeester, and G. Berghe, "Tackling the University Course Timetabling Problem With an Aggregation Approach", KaHo Sint-Lieven Information Technology & Katholieke Universiteit Leuven Campus Kortrijk, Belgium, 2005.

[5] T. Cura, "Timetabling of Faculty Lectures Using Simulated Annealing Algorithm", İstanbul Ticaret Üniversitesi Fen Bilimleri Dergisi, Turkey, 2007.

[6] E. Burke, D. Werra, and J. Kingston, "Applications to Timetabling", University of Nottingham (UK), École Polytechnique Federale de Lausanne (Switzerland), University of Sidney (Australia), 2003.

[7] Lampiran II Surat Dirjen Dikti No. 3298/D/T/99 issued on 29 Desember 1999.

[8] D. Bertsimas and J. Tsitsiklis, "Simulted Annealing", Statistical Science Vol. 8 No. 1 pp 10-15, Sloan School of Management & Electrical Engineering and Computer Science Management, Massachusetts Institute of Technology, Cambridge, 1993.

[9] S. Kazarlis, V. Petridis and P. Fragkou, "Solving University Timetabling Problem Using Advanced Genetic Algorithms", Technological Educational Institute of Serres & Aristotle University of Thessaloniki, Greece, 2002.

[10] R. Battiti, P. Gray, W. Hart, "Tabu Search", 1997, Sandia National Laboratories, Albuquerque, NM 87185. (http://www.cs.sandia.gov/opt/survey/ts.html) [accessed 19 August 2010]

[11] H. Zhang, "Artificial Intelligent: Tabu Search", 22C:145 – Artificial Intelligence, 24 October 2008, The University of Iowa, Fall 2008. (http://www.cs.uiowa.edu/~hzhang/c145/notes/04ts-search-6p.pdf) [accessed 19 August 2010]

 ISBN: 978-1-61208-139-7

Product Features Extraction and Categorization in Chinese Reviews

Shu Zhang, Wenjie Jia, Yingju Xia, Yao Meng, Hao Yu
Information Technology Laboratory
Fujitsu Research & Development Center
Beijing, China
e-mail: {zhangshu, wj_jia, yjxia, mengyao, yu}@cn.fujitsu.com

***Abstract*—With the growing interest in opinion mining from web data, more works are focused on mining in English and Chinese reviews. Product features extraction and categorization are very important for feature level opinion mining. In this paper, we propose a supervised product features extraction method, regard it as an entity recognizing process, and hope to transfer the effective NER techniques to solve this problem. We propose an unsupervised method to group the product features, mine the association of the product features from the intra and inter relationship. With experiments on Chinese reviews, the results show that proposed techniques for product features extraction and categorization are proved effective and promising. The opinion words are very important features both in features extraction and categorization.**

Keywords- opinion mining; product features; categorization

I. INTRODUCTION

With the growing of Web 2.0 platforms such as blogs, forums and various other types of social media, it becomes possible for people to find useful experience and advice from reviews or comments on products or services. Opinion mining has been proposed to analyze reviews and extricate people from wading through a large number of opinions to find their interest. People usually pay more attention to some aspects of product, it is useful to extract and analyze product features from the reviews. Product features extraction belongs to feature level opinion mining, which is finer-grained opinion mining compared with document and sentence level opinion mining.

In recent years, some feature level opinion mining systems have been presented to capture reviews' opinions on different product aspects. Opinion Observer [1] focuses on online customer reviews and provides the visual comparison of customer opinions of products on various product features. Red Opal [2] offers to find products based on features and scores each product on each feature. This information is useful to both potential customers and product manufacturers.

In reviews, people usually describe the same product features by different words. It is necessary to group them together in order to analyze the overall sentiments on one product feature. For example, "photo", "picture" and "image" all refer to the same aspect in digital camera reviews and should be grouped together, otherwise it is too detailed and tedious for customers and merchants to read and summarize all these product features. It is also infeasible and time-consuming to group the massive product features manually.

This paper focuses on extraction and categorization of product features in Chinese reviews. We propose a supervised product features extraction method, regard it as an entity recognizing process, and hope to transfer the effective NER techniques to solve this problem. We propose an unsupervised method to group the product features, mine the association of the product features from the intra and inter relationship. In this stage, we focus on finding the good indicators to reveal the association of product features and show how their influence on the performance of grouping results.

The remainder of the paper is organized as follows: Section 2 describes the related work on extraction and categorization of product features. Section 3 describes product features extraction. Section 4 presents product features categorization. Section 5 gives the experiments and results. Finally, Section 6 summarizes this paper.

II. RELATED WORK

The techniques for identifying product features are primarily based on unsupervised mining. The most representative research is that of [3]. They adopt association rule mining for extracting nouns as frequent features. Compactness pruning and redundancy pruning are used to filter the incorrect features. Popescu and Etzioni [4] utilize relation-specific extraction patterns with web PMI assessor to assess feature candidates. However, using frequency measure tends to prefer to high frequency features. This leads to the low frequency ones might be missed.

Different from their works, we adopt supervised method to extract product features. We combine frequency, syntax tokens and domain knowledge to find the product features. The importing of domain knowledge is aimed to improve the quality of extraction. With the manually tagged training corpus, we transfer the task of product features extraction into traditional information extraction task using CRFs model.

Grouping product features with similar meaning together is a recent focus in feature level opinion mining. Liu [1] employs WordNet to find synonym groups/sets exist among the features. The coverage of the lexicon is bottleneck of the lexicon-based method. Su [5] proposes a mutual reinforcement approach to clusters product features and opinion words simultaneously and iteratively by fusing both their content information and sentiment link information.

 ISBN: 978-1-61208-139-7

The inter link between product features and opinion words are mined to reinforce the clustering quality. Guo [6] constructs latent semantic association model to group words into a set of concepts according to their virtual context documents, then categorizes product features according to their latent semantic structures and context snippets in the reviews. Su [5] and Guo [6] all choose words as the basic smallest units.

Different from their methods, we adopt classical K-Means algorithm to group product features, pay more attention to mine the association of product features. We present morphemes as the smallest linguistic meaningful unit, measure the intra relationship of the product features. We mine different context information to measure the inter relationship of product features.

III. PRODUCT FEATURES EXTRACTION

In feature level opinion mining, the task is to extract product feature associated with its sentiment orientation. The task is typically divided into three main subtasks: (i) identifying product features, (ii) identifying opinions regarding the product features, and (iii) determining the sentiment orientation of the opinions. This paper mainly focuses on the first step to extract product features in Chinese customer reviews.

A. CRFs Model

The product features are mostly noun or noun phrases. In reviews, opinion words mostly appear around the product features in the sentence. The product features are context related, and for a given domain it has the lexical or syntactic similarity. For example:

"相机屏幕大，画面清晰。" (The camera has a big screen, and photo is very clear.)

Here, "屏幕"(screen) and "画面"(photo) are product features. "大"(big) and "清晰"(clear) are opinion words associated with product features.

We transfer the product features extraction to a sequence tagging problem, and hope to utilize effective NER techniques to solve this problem. Another reason for us to adopt the supervised method to implement this task is that the unsupervised frequency-based methods are dependent on the statistic of the corpus, they couldn't execute effectively when given a single sentence.

Conditional Random Fields Model is proposed by Lafferty [7], which has been proved well performance in information extraction field. It has the advantages of relaxing strong independence assumptions made in HMM [8], and avoiding the label bias problem existed in MEMM [9]. We adopt CRFs model to extract product features.

B. Feature Selection

Feature selection has been an active research pattern recognition, statistics and data mining communities. It is often the case that finding the correct subset of features is an important problem. It may significantly improve the performance of supervised learning algorithm.

In this paper, feature selection is based on some criterions: product features are mostly noun or noun phrases, and more appear in an opinion expression. That means the structure and opinion related semantic information are important. So we utilize some shallow semantic features and domain knowledge. The features are shown in the following, which include word, POS and semantic information:

Word information: We consider the neighboring words in a region with the max window 4 in order to get the context information.

POS information: POS is annotated to capture the word-building and simple syntax information. The noun phrase could be exhibited in a neighboring window with the part-of-speech tags--noun, verb, adverb, punctuation, etc.

Semantic information: we utilize some language resource to get the semantic information, such as domain feature lexicon, opinion lexicon, factor words lexicon.

We search whether the word is in product feature lexicon or not, even whether it is the part of an item or not. Because the product features might have the same or similar component in a given domain. For example, "光学变焦"(optical zoom), "数码变焦"(digital zoom), and "变焦镜头"(zoom lens) are all the product features of digital camera domain, they have the same word "变焦"(zoom) as their component.

The appearance of an opinion word or emotional adverb is more likely to indicate the presence of an opinion. As observed, people often like to express their opinions around the product feature. In product reviews, especially in Chinese reviews, people like to express their opinion in short and simple sentence, like the form of "product feature" + "opinion word". The importing of opinion and adverb lexicons aims to utilize more domain knowledge and opinion information. Since we could catch the simple collocation and pattern between the opinion word and the product features in a window by these information.

IV. PRODUCT FEATURES CATEGORIZATION

Product features categorization aims to group product features with similar meaning together. The challenge in product features categorization is how to capture the association among product features from the review. In this paper, we adopt traditional K-Means algorithm to cluster product features, and focus on mining the association of product features. We consider the association from two sides: intra relationship and inter relationship among product features. Intra relationship means the inner linguistic meaningful unit relationship between two product features. The inter relationship means the relationship of context information of two product features.

A. Morpheme Based Intra Relationship

Most researchers choose words as the basic smallest units in opinion mining. With words as basic units, it can't capture the similarity among some product features. For example, we want to measure the intra relationship among product features "电池" (battery), "电源" (power), and "电池续航能力" (battery endurance). Among them, the pair ("电池" (battery), "电池续航能力" (battery endurance)) has an intra relationship for they have the same word "电池" (battery).

 ISBN: 978-1-61208-139-7

The pairs ("电池" (battery), "电源" (power)) and ("电源" (power), "电池续航能力" (battery endurance)) have no intra relationship as they have no same word. In fact, these product features have the similar meaning in reviews.

Looking smaller units than words level, the above three product features all contain the character "电". This is a good indicator to reflect the association among them. Yuen [10] infers semantic orientation of Chinese words from their association with strongly-polarized Chinese morphemes. The conclusion is that morphemes in Chinese, as in any language, constitute a distinct sub-lexical unit, and have greater linguistic significance than words.

So we choose the morphemes to be smallest linguistic meaningful unit to mine the intra relationship among product features, and calculate the inner characters similarity of product features.

In Chinese, morphemes are mostly monosyllabic and single characters, although there are some exceptional poly-syllabic morphemes like "葡萄"(grape), "咖啡"(coffee), which are mostly loanwords.

In reviews, morphemes reflect the core meaning of product features clearly. For example, " 镜 "(lens), "屏"(screen) and "像"(photo) are the important component of product features in digital camera reviews.

B. Opinion Words Based Inter Relationship

Intra relationship only mines the association among product features from their inner characters components. This information is limited. It is not enough to capture the underlying semantic association of various product features.

In feature level opinion mining, product features and opinion words are basic element. The opinion words mostly appear around the product features in the review sentences. They are highly dependent on each other. It is obvious that surrounding opinion words may play an important role in clustering product features. So we mine the inter relationship among product features utilizing the context information, especially the opinion words associated with product features.

There are hidden sentiment association existing between product features and opinion words. For example, "外型"(shape) and "样子" (appearance), they are not similar on morphemes level, and could not be linked with intra relationship though they refer to the same aspect in reviews. However, they may be evaluated by similar opinion word "美丽" (beautiful). The opinion words describing this aspect of "appearance" are often using the words "美丽" (beautiful), "时尚"(fashion), "流行的"(popular) etc. So the opinion words around the product features really contain the semantic information to reflect the inter relationship among product features.

C. Representation

Product features categorization is conducted by representing each data object instance by a feature vector. We represent an instance as a set of following features.

Morphemes units M: all the characters contained by x_i.

Opinion words units O: only the opinion words in the given window size $\{-t, t\}$ are considered.

The weight of each features units f_j^i is calculated by Mutual Information.

$$PMI(f_j^i, x_i) = \log_2 \frac{P(f_j^i, x_i)}{P(f_j^i)P(x_i)} \quad (1)$$

Where, $P(f_j^i, x_i)$ is the joint probability of x_i and f_j^i co-occurred in the corpus. $P(f_j^i)$ is the probability of f_j^i occurred in the corpus. $P(x_i)$ is the probability of x_i occurred in the corpus. The ratio is a measure of the degree of statistical dependence between the x_i and f_j^i.

V. EXPERIMENTS

In this section, we evaluate the proposed methods and analyze the performance of product features extraction and categorization in detail.

A. Performance of Product Features Extraction

This experiment is conducted on the corpus provided by the COAE (The first Chinese Opinion Analysis and Evaluation), which was held in 2008, aims to enable researchers to participate in large-scale experiments and evaluations, make each researcher's result comparable and promote the related technique in Chinese opinion analysis.

The corpus contains automobile and electronic domains, with about 1,500 sentences each. All product features have been annotated by human.

The precision, recall and F-measure will be used to measure the performance. We adopt strict matching, which means the results submitted by systems are exactly same with the human labels.

Table I and Table II present the evaluation results. For the comparison with others, we also give the Avg. and Max. values in the task. There are 13 participants in this task. Our system is named as FRDC. We aim to testify the performance of proposed method and its capability of domain transplant.

TABLE I. RESULTS OF PRODUCT FEATURES EXTRACTION

RunID	Precision	Recall	F-measure
FRDC	0.3798	**0.4172**	**0.3976**
Avg	0.2877	0.2270	0.2331
Max	0.5641	0.4172	0.3976

TABLE II. DETAIL RESULTS ON DIFFERENT DOMAINS

RunID	Precision	Recall	F-measure
Automobile	0.2435	0.3326	0.2811
Camera	0.3512	0.3563	0.3537
Phone	0.3920	0.3539	0.3720
NoteBook	0.3782	0.3880	0.3830

 ISBN: 978-1-61208-139-7

In Table I, compared with the average and maximum value gotten in the COAE, the value of FRDC in F-measure proves that CRFs-based feature extraction is feasible and valid. Our system's precision and recall are similar and not inclined to one parameter excessively, which means our method is more practical and feasible.

Table II shows the detail results of performance on different domain. The test data include automobile and electronic domains. The electronic domain has the Camera, Phone and Notebook sub-domain. The performance on electronic domain is better than that on automobile domain on all parameters. However for the sub-domain on electronic domain the performance is similar. So it could be concluded that the performance of the system is affected by the domain, but is not sensitive. When the difference between the two given domain is not significant, the performance is similar. One reason for low performance on automobile domain might be caused by that the product features is longer and more complex.

B. Performance of Product Features Categorization

With the limited of human efforts and time, we testify the performance of product features categorization on digital camera domain. The corpus contains about 22,000 posts, extracted from the review websites. Three humans categorize product features into the categories, and we choose the label agreed by at least two humans as the standard. The detail of the corpus is shown in Table 3. We suppose that product features are extracted correctly.

TABLE III. CATEGORIZATION EVALUATION SET

Category	Number of product features
Lens	56
Screen	62
Appearance	110
Battery	18
Photography	76
Total	322

The performance of product feature categorization is evaluated using the measure of Rand Index. It is a measure of cluster similarity.

$$Rand(P_1, P_2) = \frac{2(a+b)}{n \times (n-1)} \quad (2)$$

Where, P_1 and P_2 respectively represent the partition of an algorithm and manual labeling. The agreement of P_1 and P_2 is checked on their $n \times (n-1)/2$ pairs of instances, where n is the size of data set *D*. For each two instance in *D*, P_1 and P_2 either assigns them to the same cluster or to different groups. Let *a* be the frequency where pairs belong to the same group of both partitions. Let b be the frequency where pairs belong to the different group of both partitions. Then Rand Index is calculated by the proportions of total agreement.

In our experiment, *D* contains the product features words in the pre-constructed evaluation set. Partition agreements between the pairs of any two product features are checked automatically. This measure varies from 0 to 1. The score of 1 is the best.

We first testify the performance of the proposed techniques from two perspectives:

1 The effectiveness of inducing morpheme as features to measure the intra relationship among product features.

2 The effectiveness of opinion words as feature to measure the inter relationship among product features.

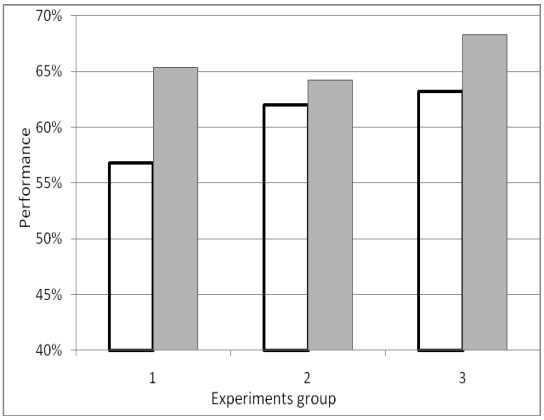


Figure 1. Different feature chosen result.

In Figure 1, the experiment of No.1 group compares the performance of method using full context as features with that of opinion words. It is only considered the inter relationship among product features, no consideration of intra relationship. The left column is the method using full context as features, which is much less than that of opinion words as features (the right column) in accuracy value. That proves the opinion words are good at indicating the semantic similarity of product features associated with them. Compared with opinion words, the full context more likely induce some noise information.

Based on No.1 group, we induce intra relationship measurement. The experiment of No.2 group induces the intra relationship measurement based on word level. The experiment of No.3 group measures the intra relationship with morphemes. No.3 group achieve better accuracy than both No.2 group and No.1. That proves morpheme features are more effective than word features. The inducing of morpheme features to measure intra relationship enhances the performance.

VI. CONCLUSION

In this paper, we probe into the problem of product features extraction and categorization. We propose CRFs-based method to extract product features in reviews. We propose an unsupervised product features categorization method. With the experiments in Chinese reviews, the proposed methods achieve better performance. CRFs-based product features extraction is effective and feasible. Morphemes and opinion words are proved to be the

 ISBN: 978-1-61208-139-7

important features to capture the semantic similarity among product features in process of product features categorization.

However, the methods are only tested on Chinese customer reviews. We will conduct experiments on different languages and domains in future work.

REFERENCES

[1] B. Liu, M.Q. Hu, and J.S. Cheng, "Opinion Observer: Analyzing and Comparing Opinions on the Web," Proc. International World Wide Web Conference, pp. 342–351, 2005.

[2] C. Scaffidi, K. Bierhoff et al. "Red Opal: Product-Feature Scoring from Reviews," Proc. 8th ACM Conference on Electronic Commerce, pp.182–191, 2007.

[3] M.Q. Hu, and B. Liu. "Mining Opinion Features in Customer Reviews," Proc. American Association for Artificial Intelligence, pp. 775–760, 2004.

[4] A. Popescu, and O. Etzioni, "Extracting Product Features and Opinions from Reviews," Proc. on Empirical Methods in Natural Language Processing, pp. 339–346, 2005.

[5] Q. Su, X.Y. Xu et al, "Hidden Sentiment Association in Chinese Web Opinion Mining," Proc. 17th International Conference on World Wide Web, pp. 959–968. 2008.

[6] H.L. Guo, H.J Zhu et al, "Product Feature Categorization with Multilevel Latent Semantic Association," Proc. International Conference on Information and Knowledge Management, pp.1087–1096, 2009.

[7] J. Lafferty, A. McCallum and F. Pereira, "Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data," Proc. International Conference on Machine Learning, pp. 282–289. 2001.

[8] L. R. Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition," Proc. IEEE, 77 (2), pp. 257–286, 1989.

[9] A. McCallum, D. Freitag, and F. Pereira, "Maximum Entropy Markov Models for Information Extraction and Segmentation," Proc. on International Conference of Machine Learning, pp. 591–598, 2004.

[10] W.M Yuen Raymond et al, "Morpheme-based Derivation of Bipolar Semnatic Orientation of Chinese Words," Proc. 20th International Conference on Computational Linguistics, pp. 417–424. 2004.

 ISBN: 978-1-61208-139-7

Low-Density Parity Check Codes for High-Density 2D Barcode Symbology

Ramon Francisco Mejia, Yuichi Kaji, and Hiroyuki Seki
Graduate School of Information Science
Nara Institute of Science and Technology
8916-5 Takayama, Ikoma, Nara 630-0192, Japan
Email: ramon-m@is.naist.jp, kaji@is.naist.jp, seki@is.naist.jp

***Abstract*—Increasing the data density of two-dimensional barcode symbology is an important area of research in Automatic Identification and Data Capture systems because it provides significant improvements on the range and usefulness of barcodes in different applications. The implementation of error-correcting codes for such symbologies is crucial due to the susceptibility of high-density barcode symbols to damage. In this paper, we study the performance of Low-Density Parity Check (LDPC) codes for a two-dimensional barcode symbology with high data density. A high-density barcode symbology is designed, and the characteristics of the communication channel defined by commonly used printers and scanners are modeled and observed. Additionally, the parameters of the symbology are adjusted and the performance of LDPC codes with different code rates is tested. Performance tests show that LDPC codes are effective for certain parameters of the high-density symbology, and limitations of the chosen printing technology severely affect the robustness of the symbology.**

***Keywords-Two-Dimensional Barcodes*; *Low-Density Parity Check Codes*; *Automatic Identification and Data Capture*; *Image Processing*.**

I. INTRODUCTION

A lot of interest has been put on the study of *two-dimensional (2D) barcode systems*, where data is represented as a machine-readable symbol printed on a physical surface called a *barcode symbol* [1]. In such systems, the symbol represents data in a matrix of high and low reflectance regions of the printing surface, and therefore able to carry between 10 to 100 times more data than one-dimensional barcodes. Continued development in this area revolves around improving the *symbology* of 2D barcodes in order to increase its data density; i.e., enhancing the structure and processing methods of a barcode symbol to increase its data capacity while retaining its compact size and portability.

Improvements in the symbology of 2D barcodes further expand its application for different systems. Quick Response (QR) Codes [2], created by Japanese company Denso-Wave in 1994, are used for Medical Information Management systems, where prescriptions are stored in a barcode to reduce human error in the interpretation of handwritten prescriptions and the administration of medicine [3]. In Fu et al. [4], a study of tax-filing methods in Taiwan revealed that the adoption of an electronic filing system which utilizes 2D barcodes increased the processing efficiency of tax returns and reduced error rates versus paper returns. In addition, barcodes also have applications in Mobile Commerce [5], Multimedia Teachware [6] and many others. For these applications, the high data density of 2D barcodes is important as they function not only as an index to external databases, but can hold either files or databases themselves.

A number of 2D barcode symbologies have been proposed so far, and while most are designed with reading speed in mind, some of them are designed for high data capacity [7]. For example, Optar [8] can accommodate 200 kilobytes of data in one A4-paper. PaperDisk [9] can contain 1 megabyte of data in an 8.5×11 inch space. Laboratory tests for High Capacity Color Barcode [10] using eight colors have a capacity of 2,000 bytes of data per square inch in its highest density. However, technical details of these technologies are not disclosed so far. It will be useful if high data capacity barcodes are realized using open technologies only, and if these barcodes can be used with reasonably-priced equipment found in office or home environments.

One disadvantage to increasing the data density of 2D barcodes is its increased susceptibility to errors. Physical damage or inaccuracies caused during the printing or scanning process of the 2D symbol may cause erroneous reading of the stored data bits. To recover from errors, most 2D barcode symbologies use *Reed-Solomon* (RS) codes to encode data prior to generating the symbol [11]. In recent years, it has been shown that well-designed *Low-Density Parity Check* (LDPC) codes [12] perform well compared to RS codes for some communication channels. A remarkable aspect of LDPC codes is that we can perform *soft-decision decoding* for LDPC codes with almost linear-time complexity. Soft-decision decoding is an algorithm for error correction in which inputs to the algorithm can have continuous values. It is more powerful than conventional (referred to as hard-decision) algorithms in which inputs are quantized into two level, but usually requires very large computational complexity. For this reason, soft-decision decoding has been considered as impractical for conventional error-correcting codes including RS codes. On the other hand, this problem can be mitigated with the use of LDPC codes. It has been shown that the performance of appropriately designed LDPC codes asymptotically approaches to Shannon limit as we extend the code length [13], and that LDPC codes with a practical code length show better performance than

 ISBN: 978-1-61208-139-7

conventional error-correcting codes such as RS codes and convolutional codes [14], [15].

This study investigates the error-correcting performance of LDPC codes in a high-density 2D barcode symbology. Given the advantageous properties of LDPC codes over RS codes, our goal is to design a symbology capable of storing large amounts of data, and determine appropriate symbology parameters and LDPC codes to ensure robustness against errors. In order to achieve this, the symbology was designed to allow larger data block sizes (i.e., longer LDPC codes) to be placed within the symbol. The performance of LDPC codes were then tested with barcode symbols of different data densities and geometric parameters. This symbology also takes into consideration the limitations of printing and scanning technologies present in most office environments today.

The remainder of this paper is organized as follows. In Section II the design of the symbology is introduced. Next, methods for evaluating the performance of LDPC are described in Section III. Finally, Section IV discusses the results of the evaluation.

II. SYMBOLOGY

The proposed barcode symbology for this study includes common structural features found in other 2D barcode symbologies, such as finder patterns and timing patterns. However, the focus of the proposed symbology is to increase data density, where the number of bits represented in a given printing area is significantly greater than in other 2D barcodes designed for fast reading. Therefore, instead of using traditional barcode printers and scanners with limited resolutions and computing power, the symbology is designed to work with equipment available in most office environments; namely, a laser printer, a flatbed scanner, and a desktop computer. The following subsections describe the structure of the symbology and the processes involved in generating and scanning symbols.

A. Data Area

A symbol is composed of a $d \times d$ matrix of *data cells* printed as a monochrome image on approximately a 25.4 mm $\times$ 25.4 mm data area, where d is the dimension or the number of cells per row, and d^2 is the density of the symbol. Each data cell represents a single bit of data, printed as a solid square. For the purposes of this study, a white cell is printed when a '0' bit is required and a black cell for a '1' bit.

Because of the size restrictions for the data area, the dimension d of the matrix determines the printing area for a data cell. The *cell size* in millimeters is computed as $(25.4 \text{ mm}/(d + 2 \text{ cells}))^2$. The additional 2 cells are allocated for timing patterns, which will be explained in the following subsection. Since data cells are small (around 0.257mm $\times$ 0.257mm for $d = 97$), it is probable that the printer toner used to draw black cells "spill out" and blot neighboring white cells. This is called *inter-pixel leakage*, and this is mechanically unavoidable when laser printers are used in printing. This may not be the case when using professional-quality image setters, however as mentioned, the environment considered is that of a typical office setting.

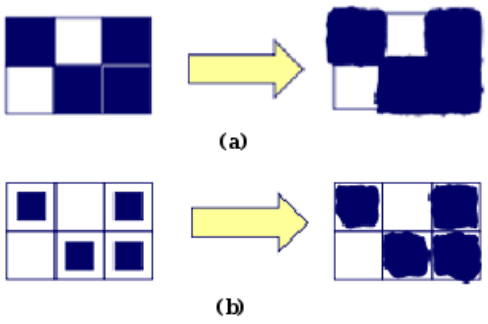


Figure 1. Inter-pixel leakage for (a) black cells printed without adjustments ($mf = 1.0$) and (b) with margin factor ($mf = 0.6$).

To reduce this effect, we propose that the printing size of black cells is reduced to a certain percentage by a *margin factor*. Figure 1 shows the effect of inter-pixel leakage and adding a margin factor.

B. Timing and Finder Patterns

The data area is bordered by four timing patterns. A *timing pattern* is used in most 2D barcodes as a way to calculate the size and location of data cells. In this symbology, the timing pattern is a consecutive series of data cells starting and ending with a black cell, connected with alternating white and black cells. Note that since timing patterns must start and end with a black cell, the dimensions of the data area must be odd.

In order to detect the location of the timing patterns, finder patterns are positioned on the four corners of the symbol. A *finder pattern* consists of three black rings which are co-centric to the first cell of each timing pattern. The rings intersect with the first three black cells of the timing pattern. The diameter and thickness of the rings depend on the dimensions and margin factor of the symbol. To improve the detection of the finder pattern, a *quiet zone* of white cells is placed around the location of the finder patterns. As a result of placing quiet zones in the corners of the symbol, the locations of overlapping data cells are adjusted. Figure 2 illustrates an example of a symbol with timing and finder patterns, and a complete barcode symbol.

C. Encoding Procedure

Data in a barcode symbol is protected by an error-correcting code which is known as an LDPC code. LDPC codes are a class of linear block codes that have very sparse parity check matrices. The sparse structure of the

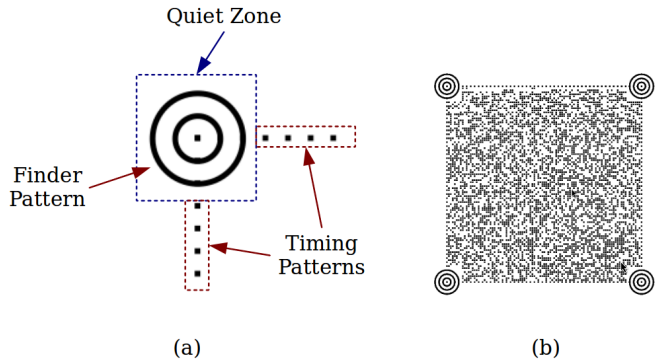


Figure 2. (a) Construction of a finder pattern, quiet zone and two timing patterns. (b) A complete symbol created using the symbology.

check matrix makes the "belief-propagation" principle very effective for finding and correcting errors which are involved in a received signal.

In this study, we consider to use LDPC codes which are designed for IEEE 802.16e standard (also known as mobile WiMAX). Using these codes is advantageous since they have smaller complexity for encoding operation. In general, the encoding operation of an LDPC code requires quadratic-order complexity in the code length; however, the IEEE codes defined in the standard are designed so that they have *quasi-cyclic* structure, which enables the realization a linear-order encoding algorithm.

Another advantage of these codes is that the code parameters can be changed in a flexible manner. The standard defines several classes of LDPC codes with code rates 1/2, 2/3, 3/4 and 5/6, and code length ranges from 576 to 2304 bits. These parameters have a strong relation to the efficiency and the error-correcting capability of the code [16]. Generally, low-rate codes are more powerful than high-rate codes, but more parity bits need to be added for such codes. This implies that, with the same code length, high-rate codes can accommodate more data than low-rate codes. It is also known that long codes usually show better performance than short codes even if they have the same code rate, but the computational cost of longer frame-lengths need to be considered. These parameters are adjusted according to the dimension of barcodes and desired level of reliability.

D. Symbol Processing

We refer to the process of printing and scanning barcodes as *symbol processing*. During symbol processing, the symbol is printed on a piece of paper using a laser printer. A flatbed scanner scans the symbol and the resulting image is passed to an image processing program. The program performs five steps:

1) Finder patterns are located using a basic template matching algorithm [17]. The centers of all finder patterns are then computed.
2) Lines connecting the centers of adjacent finder patterns are connected, and the resulting closed rectangle is masked. This enhances the detection of timing patterns in the following step.
3) The lines connecting adjacent finder patterns are scanned through pixel by pixel (note that one cell consists of several pixels). Given the position of the finder patterns, the path each line passes through is also the location of a timing pattern. When the value of the pixel changes from white to black, this pixel is marked as a *transition point*.
4) Transition points from opposite timing patterns are connected, forming a grid of *sampling cells*. The ratio of black pixels to the total number of pixels from each sampling cell is computed.
5) The ratio r obtained from each sampling cell is mapped to a *soft value* using the function $f(r) = -(2r - 1)$. Soft values are grouped into codewords, which are passed to an LDPC decoder program.

E. Decoding Procedure

The decoding, or error correction, is performed by using a belief-propagation algorithm for LDPC codes [13]. In this algorithm, we consider representing the mathematical structure of the code with a bipartite graph whose incident matrix coincides with the check matrix of the code. The nodes of the bipartite graph are grouped to *variable nodes* and *check nodes*. A variable node receives information from neighbor check nodes, and it attempts to estimate which symbol ('0' or '1' bit) has been transmitted. During the estimation, the statistical information of the communication channel, such as the variance of the Gaussian channel, is considered to derive various probabilities. A check node receives the estimated symbols from neighbor variable nodes, monitors parity constraints, and gives check nodes suggestions for the transmitted symbol. The accuracy of the estimation improves as nodes exchange messages iteratively. It is known that, in most cases, the decoding algorithm reaches the correct codeword with a small number of iterations. The number of iterations needed is rather independent from the code length, thus the decoding algorithm can be regarded as "almost linear-order" complexity.

III. Evaluation

In order to evaluate the performance of the symbology, two experiments were conducted on samples of barcode symbols. The experiments assessed the behavior of the communication channel from different perspectives.

All experiments used the same equipment and symbol processing steps for testing. First, a set of 24 barcode symbols was generated using an encoder program written in the C programming language. The symbols were then printed on plain white bond paper using a Canon LBP3410 laser printer with the default settings. Next, symbols were scanned using

 ISBN: 978-1-61208-139-7

an Epson GT-F720 flatbed scanner at 720dpi. In both the printing and scanning process, monochrome color settings were used. Finally, each symbol was read using an image processing and decoding program, also written in C.

A. Test 1: Analysis of Channel Characteristics

In the first experiment, the characteristics of the communication channel with respect to symbol processing were investigated. The image data read by the scanner are not identical to the virtually constructed barcode image. Many factors, such as the inter-pixel leakage of printers or blobs on bond paper, causing differences between the scanned image and the ideal image. For reliable data recording, it is essential to eliminate these effects or noises. Furthermore, the communication channel which is defined by a printer and a scanner is different from conventional communication channels, therefore statistical analysis of the channel is vital in determining how to implement LDPC codes for the symbology.

As a result, a set of symbols with dimension $d = 117$ and margin factor $mf = 0.6$ were generated and sampling cells were scanned. The computed ratios (before mapping to a soft value) were separated into two groups, based on the expected value (white or black) of the cell. Histograms for both groups were then plotted to infer observations on the channel model. From the distribution, we can observe how the values in the sampling cells match the encoded data bits after they have gone through symbol processing.

B. Test 2: Performance of LDPC Codes

In the second experiment, the error-correcting performance of LDPC codes was evaluated against changes to parameters of the symbology. To measure performance, the bit-error rate (BER) [18] of each sample set was analyzed, where

$$BER = \frac{\text{Number of erroneous bits after decoding}}{\text{Total number of bits in the set}} \quad (1)$$

There are two phases in this experiment. In the first phase, we wish to observe the error correction performance of LDPC codes as data density increases. Sample sets with increasing dimensions were created using LDPC codes with code rates $1/2$, $2/3$, $3/4$, and $5/6$. As previously mentioned, the LDPC codes used are those defined for IEEE 802.16e, and the code rate indicates the error correction capability of the code (lower code rate means stronger error-correcting capability). The margin factor for all symbols was set at $mf = 0.6$.

In the second phase, sample sets with different margin factors were created using the same LDPC code rates as above. The dimensions of all symbols was set at $d = 127$. As discussed earlier, the printing size of a black cell affects the impact of inter-pixel leakage and the ability of the image processing stage to identify the finder and timing patterns correctly. It is therefore of interest to see how the printing size of black cells affects error-correcting performance.

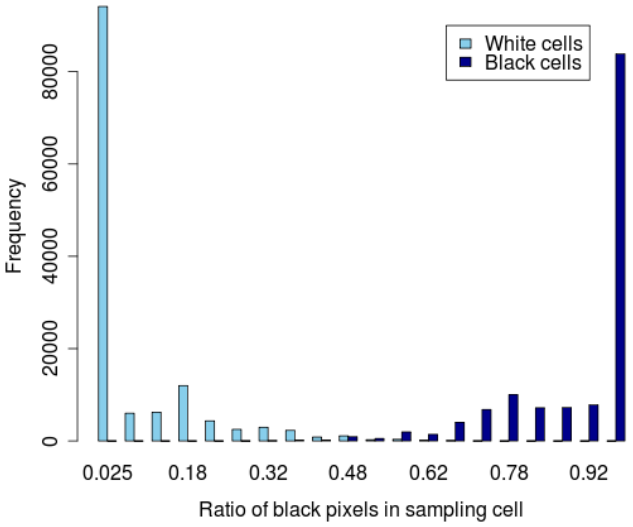


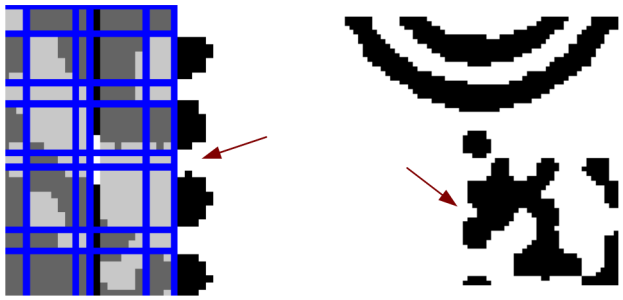
Figure 3. Histogram of ratios for black cells and white cells for symbols with dimension $d = 127$ and margin factor $mf = 0.6$

IV. Results and Discussion

The histograms generated for Test 1 are presented in Figure 3. Statistical analysis of the distribution showed that the channel can be modeled as an *additive white Gaussian noise* (AWGN) channel, which is suitable for the soft-decision decoding of LDPC codes. Also, by analyzing the distribution, the AWGN variance which will be used for the LDPC decoder in the subsequent test was determined to be 0.05637.

During the execution of the first experiment, it was observed that the printing process had another effect on the printing of the data cells. In addition to inter-pixel leakage, cells along the same row are printed with similar sizes, but cells in other rows may be printed with different heights. The same effect was also observed for the widths of cells from the same columns and cells in different columns. It is not clear if this effect is caused by the mechanical constraints of the printer or scanner, or other known factors; the printed barcode image is not as precise and uniform as stated in the specifications of the devices. Nevertheless, the timing pattern considered in this study can cope with this effect; since timing patterns are also printed on the same row or column as data cells, the grid of sampling cells formed by transition points adjusts to the changes in the printing size of the rows or columns. However, this also means that

 ISBN: 978-1-61208-139-7



(a) (b)

Figure 4. (a) Timing patterns are too close and image sampling cells are too small. (b) Timing patterns have no clear separation and creation of the sampling cells fail.

the accuracy of timing pattern detection is crucial to the performance of the symbology.

In Test 2, the parameters for sample sets were varied and the BER for each set was analyzed. The results of the BER analysis for phase 1, where symbols with increasing dimensions were tested, are presented in Table I. From the results, it can be concluded that LDPC codes with different code rates have similar performance for lower dimensions of the symbol. All codewords from the set of symbols were decoded correctly. For $d = 117$, errors appeared for the symbols with LDPC code rate 5/6. This was expected, as this code rate has the weakest error-correcting capability. However, there is a rapid increase of bit-errors for $d = 127$. Inter-pixel leakage of black cells caused timing patterns to become too close to each other, thus the sampling cells formed were too small and there were not enough pixels to compute accurate soft values. Moreover, note that the BER for $d = 137$ is "Undetermined". Due to the high density of cells in the symbol, some images were distorted by inter-pixel leakage in such a way that transition points for adjacent timing patterns could not be detected by the image processing algorithms, and the creation of sampling cells failed. Figure 4 shows some examples of image processing errors encountered during the experiment.

Table I
BER FOR SYMBOLS WITH MARGIN FACTOR $mf = 0.6$ AND INCREASING DIMENSIONS

	Code rate			
d	1/2	2/3	3/4	5/6
97	0.000%	0.000%	0.000%	0.000%
107	0.000%	0.000%	0.000%	0.000%
117	0.000%	0.000%	0.000%	0.847%
127	0.709%	2.143%	5.674%	1.429%
137	Undetermined			

Table II shows the BER analysis for phase 2, where symbols with different margin factors were tested. We can see that the error-correcting performance degrades if the margin factor is too small or too large. If the margin factor is too small, then a small blot on the paper can cause more white cells to be recognized as black cells. More importantly, the printing process may fail to draw data cells and timing patterns, as was observed in some cases. On the other hand, if the margin factor is too large, then problems related to inter-pixel leakage may arise. If these undesired phenomena occur, then error-correcting codes would not be effective. Indeed, there seems to be little relation between the BER and the capability of codes for margin factors 0.50 and 0.60, which are too small and too large, respectively. The margin factor 0.55 is the best for this resolution, and we can see a monotonic relation between the BER and the code rate.

Table II
BER FOR SYMBOLS WITH DIMENSION $d = 127$ AND INCREASING MARGIN FACTORS

	Code rate			
mf	1/2	2/3	3/4	5/6
0.45	Undetermined			
0.50	2.128%	1.429%	8.511%	7.143%
0.55	2.128%	2.857%	3.546%	3.571%
0.60	0.709%	2.143%	5.674%	1.429%
0.65	Undetermined			

Another remark is that the parameters chosen in this test are close to the performance limit of commonly used laser printers. For the symbols of $d = 127$, the printing size of each cell is around 0.197mm $\times$ 0.197mm. Setting the margin factor to 0.50 means that each black cell is drawn by the size 0.098mm $\times$ 0.098mm. Due to the small size, a printer may not be able to control the image. In a 600dpi setting, the minimum unit a printer can control is $25.4/600 = 0.042$ mm. Therefore, a black cell with size 0.098mm $\times$ 0.098mm is composed of two by two units. Furthermore, with such a small scale, the size of the toner particles used in laser printers cannot be ignored. It is said that the diameter of toner particles is around 0.005mm to 0.010mm, which is not very small compared to the cell size. In addition, the toner particle is firmly pressed on paper surface during the printing process, and it is difficult to fully control the position of toner particles on the paper.

V. Conclusion and Future Work

In order to evaluate the effectiveness and robustness of LDPC for high-density 2D barcodes, we first designed a new symbology. Two types of experiments were conducted on the symbology; these were done to observe the characteristics of the channel, and to analyze the changes in the BER of different LDPC codes as the parameters of the symbology were adjusted.

This study has concluded that LDPC codes are feasible for our proposed symbology. The results showed that the communication channel defined by printing and scanning the barcode symbols can be modeled as an AWGN channel, which is suitable for soft-decision decoding of LDPC

 ISBN: 978-1-61208-139-7

codes. Also, the error correction performance of LDPC codes defined in the IEEE 802.16e standard are effective on the channel for certain symbology parameters. Finally, properties of laser printing technology severely affect the performance of the symbology, particularly with regards to the small printing size of data cells. In these cases, LDPC codes were not effective in error correction.

Future work includes performance evaluation of RS codes for the same high-density barcode symbology, and rigorous testing of the robustness of the symbology against physical damage to the symbol.

REFERENCES

[1] Japanese Standards Association, "JIS X 0500-2:2009 – information technology – automatic identification and capture $AIDC$ techniques – harmonized vocabulary – part 2 Optically readable media ORM," 2009.

[2] ——, "JIS X 0510:1999 – two dimensional symbol – QR code – basic specification," 1999.

[3] Y. L. Yeh, J. C. You, and G. J. Jong, "The 2D barcode technology applications in medical information management," *Intelligent Systems Design and Applications, International Conference on*, vol. 3, pp. 484–487, 2008.

[4] J. Fu, C. Farn, and W. Chao, "Acceptance of electronic tax filing: A study of taxpayer intentions," *Information & Management*, vol. 43, no. 1, pp. 109–126, 2006. [Online]. Available: http://www.sciencedirect.com/science/article/B6VD0-4H0BSXY-1/2/eb9f5899522023887bc24f3e111ea556

[5] J. Gao, V. Kulkarni, H. Ranavat, L. Chang, and H. Mei, "A 2D barcode-based mobile payment system," in *Proceedings of the 2009 Third International Conference on Multimedia and Ubiquitous Engineering*, ser. MUE '09. Washington, DC, USA: IEEE Computer Society, 2009, pp. 320–329. [Online]. Available: http://dx.doi.org/10.1109/MUE.2009.62

[6] D. Kim and Y. Mun, "Design and performance analysis of multimedia teachware making system using 2D barcode," in *Computational Science and Its Applications - ICCSA 2006*, ser. Lecture Notes in Computer Science. Springer Berlin / Heidelberg, 2006, vol. 3981, pp. 195–203. [Online]. Available: http://dx.doi.org/10.1007/11751588_21

[7] Y. Kaji, "Two-dimensional barcode system with extreme density," Institute of Electronics, Information and Communication Engineers, Tech. Rep. IT2009-56, January 2010.

[8] Twibright Labs, "Optar," http://ronja.twibright.com/optar/ 14.04.2011.

[9] Cobblestone Software, "Paperdisk technology," http://www.paperdisk.com/id1.html 14.04.2011.

[10] Microsoft Research, "High capacity color barcode technology," http://research.microsoft.com/en-us/projects/hccb/about.aspx 14.04.2011.

[11] K. Tan, D. Chai, and H. Kato, *Barcodes for Mobile Devices*, 1st ed. New York: Cambridge University Press, 2010.

[12] R. Gallager, "Low-density parity-check codes," *Information Theory, IEEE Transactions on*, vol. 8, no. 1, pp. 21–28, 1962. [Online]. Available: http://dx.doi.org/10.1109/TIT.1962.1057683

[13] D. J. C. MacKay, "Good error correcting codes based on very sparse matrices," *Information Theory, IEEE Transactions on*, vol. 45, no. 2, pp. 399–431, 1999.

[14] N. Andreadou, C. Assimakopoulos, and F. N. Pavlidou, "Performance evaluation of LDPC codes on PLC channel compared to other coding schemes," in *Power Line Communications and Its Applications, 2007. ISPLC '07., IEEE International Symposium on*, March 2007, pp. 296–301.

[15] J. Chen, L. Wang, and Y. Li, "Performance comparison between non-binary LDPC codes and reed-solomon codes over noise bursts channels," in *Communications, Circuits and Systems, 2005. Proceedings. 2005 International Conference on*, vol. 1, May 2005, pp. 1–4.

[16] M. G. Luby, M. Mitzenmacher, M. A. Shokrollahi, and D. A. Spielman, "Improved low-density parity-check codes using irregular graphs," *Information Theory, IEEE Transactions on*, vol. 47, no. 2, pp. 585–598, February 2001.

[17] M. Sonka, V. Hlavac, and R. Boyle, *Image Processing, Analysis, and Machine Vision*, 2nd ed. Pacific Grove: PWS Publishing, 1999.

[18] S. Lin and D. J. Costello, *Error Control Coding: Fundamentals and Applications*, 2nd ed. Upper Saddle River: Prentice Hall, 2003.

 ISBN: 978-1-61208-139-7

Research on Adaptative Concession Strategies in Argumentation-based Negotiation

Guorui Jiang
The Economics & Management School
Beijing University of Technology,
Beijing, China, 100124
e-mail: jianggr@bjut.edu.cn

Bo Hao
The Economics & Management School
Beijing University of Technology,
Bejing, China, 100124
e-mail: kujoir@emails.bjut.edu.cn

***Abstract*—The paper discusses an application of multi-agent based on theory of argumentation-based on negotiation, and provides adaptive concession strategy model for beginning a negotiation. Firstly, this paper defines hypotheses of model and a frame of negotiation based on argumentation. Secondly, for comparing purpose, two generating models of concession strategy are also studied as:model based on time constraint and model based on opposite's preferences, the process of demonstration and result of experiment has been shown that the latter designed by PSO-RBFNN (RBF Neural Network optimized by Particle Swarm Optimization) algorithm has better abilities of learning and reasoning, which is dominant strategy in bilateral negotiations,and has a certain feasibility and application value.**

Keywords-Multi-Agent; Self-Learning; Argumentation-based Negotiation; Adaptive Concession Strategies.

I. INTRODUCTION

In the process of e-commerce industrialization, Multi-Agent technology is crucial to the big change of e-commerce, and negotiation as a process of dynamic interaction is considered to be an important factor in multi-agent system. At present, some researches in the technology of automated negotiation based on multi-agent have been pursued and some achievements have also been developed. These researches often have been made in many different aspects, which are the design and implementation of automated negotiation system, the model of negotiation support system and the key technologies in automated negotiation based on multi-agent, etc. In accordance with difference of research methods, three types of negotiation have been presented, which are the negotiation based game theory, the negotiation based on heuristic algorithm and the negotiation based on argumentation, respectively. In recent years, argumentation-based negotiation has been accepted as a promising alternative to game-theoretic or heuristic-based negotiation [1]. In a argumentation-based negotiation, the problem of conflict must been brought up because of the different beliefs between buyer and seller, so that one of the main goals of negotiation is to make one negotiator agent has the ability to infer opposite negotiator 's thinking, and revise own belief through learning opposite's preferences during the process of interaction to avoid negative dialogue between both parties, which make agent be able to adjust negotiation strategies to changing environment. Up to this point, the problem of adaptive strategy has become a new topic in the field of argumentation-based negotiation.

While many researchers had developed some achievements in the field of adaptive negotiation strategies, there are two main problems with current researches, which are largely based on on-line learning. Usually, the large numbers of negotiated transaction records in history are not used effectively. At the same time, the mechanism of argumentation are not fully introduced into current most adaptive negotiation strategies. Consequently, if a negotiator can offer a proposal with argumentation based on concession strategies according to opposite's preferences reasoned from negotiating records in history, then persuasiveness of the proposal will be improved and process of negotiation will be advanced effectively. In this paper, the generating model of concessional strategies in the argumentation-based negotiation is discussed. Firstly, the paper defines hypotheses of model and a frame of argumentation-based negotiation. Secondly, the model based on time constraint and model based on opposite's preferences are presented respectively, and the latter designed by PSO-RBFNN algorithm can be demonstrated to be dominant strategy in a negotiation. Finally, the model based on opposite's preferences which can be proved to have a certain feasibility and application value via experiment.

II. MODEL HYPOTHESES AND FRAMEWORK

The paper will conduct the research on adaptive concession strategies in argumentation-based negotiation based on the following hypotheses and framework.

A. Model Hypotheses

Hypothesis 1. Agent is completely selfish, i.e., agent will pursue individual maximum utility.

Hypothesis 2. Agent has limited rationality, i.e., agent can change the mental state of the other party through offering proposal with argumentation.

Hypothesis 3. Agent with incomplete information don't know the other Agent's preference information, i.e., one agent will can not directly control other agent unless through negotiating between both parties.

Hypothesis 4. Time is precious to both parties.

 ISBN: 978-1-61208-139-7

Hypothesis 5. The both negotiating parties are sincere to reach agreement through negotiations, that is to say there isn't deceit during negotiation process.

Hypothesis 6. The failure of the negotiation is the worst result of negotiations between both parties.

B. Framework

In the accordance with hypotheses above, a framework of bilateral multi-issue negotiation based on argumentation is constructed (see Figure 1), which make agents have ability of self-learning. For example, in a negotiation, if there is a obvious differences between information of buyer agent(a_i)'s order and expectation of seller agent(a_j) after the seller agent offers a proposal, then a_i asks a_j to accept the proposal, but a_j will can not accept it in order to keep individual utility at maximum. Meanwhile, intermediary agent can acquire a_i ' s satisfaction degree of issue in the proposal through neural network module, and then the satisfaction degree will be transformed into the information that understood by a_j through the model of output processing. Further, a_j can acquire the preference condition of the buyer according to the received information, generate concessional strategy, and determine the content of argumentation to improve the persuasiveness of new proposal and advance the process of negotiation.

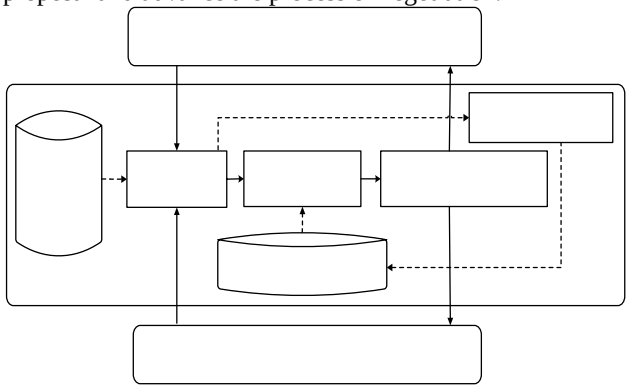


Figure 1. Framework of argumentation-based Negotiation

where dotted line show the training process of network, and solid line show the reasoning process of network.

Negotiator Agent (Seller/Buyer Agent): During the process of negotiation, Negotiator Agent make seller and buyer acquire data information that reflect interaction event in negotiation, and carry out some action that affect the negotiation process.

Intermediary Agent: Intermediary Agent is introduced into this research in order to ensure authenticity of the interaction event and avoid fraud in negotiating environment, which make Negotiator Agent acquire more effective and operational information from Intermediary Agent compared with information obtained directly from the opposite negotiator.

Negotiation Case Base: This base is used for storing finished negotiating records in history, and Intermediary Agent can take advantage of these records to learn negotiators' private information (e.g., preference) .

Knowledge Base: This base is used for storing negotiators' private information for different negotiating mission, in which one negotiating mission is stored for each record, and each record includes the important parameters needed by Neural Network Module.

Training and Learning Module: Finished negotiating records stored in Negotiation Case Base can be processed in this module. This processing is also the process of learing negotiators' private information to adjust the important parameters needed by Neural Network Module, and results will be put into Knowledge Base.

Neural Network Module: During the process of negotiation, Neural Network Module can obtain important parameters of neural network from Knowledge Base, and then reason negotiators' private informtion through processing input data information (have been pretreated through Pretreatment Module) from Negotiator Agent.

Pretreatment Module: The input data from Negotiation Case Base or Negotiator Agent can be normalized in this module so that input data fit the constraints of format required by Neural Network Module and Training and Learning Module.

Output Processing Module: Through this module, negotiators' private information reasoned by Neural Network Module will be tranferred into the information easily unstandable for Negotiator Agent.

III. CONCESSION STRATEGIES IN ARGUMENTATION-BASED NEGOTIATION

Next this paper will discuss adaptive concession strategy model in argumentation-based negotiation.

A. Parameters Settings

1) Let $Ag\{a_1, a_2, a_3, ..., a_t\}$ denotes negotiator agents vector, where a_t is the negotiator agents, t is the number of Agent, i.e., "t=2" denotes bilateral negotiation.

2) Let $r, r \in N$ denotes the current round of negotiation. Let R denotes time constraint of negotiation. One negotiator offer a proposal and the opposite offer a counter-proposal, or the one side agent only sends "Accept", which denotes completing a round of negotiation and then let $r = r + 1$. Let r_{end} denotes the total rounds of completed negotiation. Therefore the following conditions hold $0 \le r \le r_{end} \le R$.

3) Let $Q =< q_s, q_x >$ denotes the information of negotiating mission participated by both agents in a negotiation, where q_s consists of main information of negotiation mission, e.g., the name and identity of goods, etc. The elements in $q_x =< q_{x1}, q_{x2}, q_{x3}, ..., q_{xu} >$ denotes u issues of the goods involved in the negotiation.

 ISBN: 978-1-61208-139-7

4) Let $C(a_i,r,q_x)=\prec c(a_i,r,q_{x1}),c(a_i,r,q_{x2}),...,c(a_i,r,q_{xu})\succ, a_i \in Ag, 0 \le r \le r_{end}$ denotes the issues vector of Agent a_i in the (r)th round, where $c(a_i,r,q_{xu})$ denotes the value of issue.

5) Let $C_r(a_i,q_x)=C(a_i,r,q_x), a_i \in Ag, q_x \in Q, 0 \le r \le r_{end}$ denotes the vector of issues in current round of negotiation, where $c_r(a_i,q_{xu}) \in C_r(a_i,q_x)$ represents the current value of issue q_{xu}.

6) Let $C^h_{rend}(a_i,q_x)=C(a_i,r_{end},h,q_x), a_i \in Ag, q_x \in Q, 0 \le r_{end} \le R, 1 \le h \le n$ denotes the vector of final deal agreed by both parties during the (h)th negotiated transaction in history, where n is the number of negotiated transaction about goods Q in history, and $c^h_{rend}(a_i,q_{xu}) \in C^h_{rend}(a_i,q_x)$ represents the final value of issue q_{xu} in the (h)th negotiated transaction in history.

B. Description of Definition

Definition 1(Satisfaction Degree). A contrast of buyer's feelings generated by purchasing a goods and their own expectation(preference), which is a relative concept. It can be used to assist the seller to investigate the match condition between seller's goods and buyer's expection, and can also be quantized through reasoning in accordance with initial information of goods'issues. The range of satisfaction degree is [0,1], and the greater value, the better degree of satisfaction.

Definition 2(Preference Coefficient). Representing a negotiator' s favorable attitude toward negotiating goods. A negotiator can sort the important degree of negotiating issues in accordance with own expectation. The sorting order can reflect the needs, interests and hobbies of negotiators. The range of preferences coefficient is [0,1], the greater value, the more important for negotiators.

The formal description of satisfaction degree and preferences coefficient is as follows.

1) Let $S(a_i,r^h_{end},q_{xu})=\prec s(a_i,r^1_{end},q_{xu}),s(a_i,r^2_{end},q_{xu}),...,s(a_i,r^n_{end},q_{xu})\succ, a_i \in Ag, h=1,2,...,n$ denotes the vector of satisfaction degree about single issue q_{xu} during the (h)th negotiated transaction in history, where n is the number of negotiated transaction about goods Q in history.

2) Let $s(a_i,r,q_{xu}), a_i \in Ag$ denotes the value of the satisfaction degree about issue q_{xu} in current round of negotiation.

3) Let $S(a_i,r^h_{end},\sum q_{xu})=\prec s(a_i,r^1_{end},\sum q_{xu}),s(a_i,r^2_{end},\sum q_{xu}),...,s(a_i,r^n_{end},\sum q_{xu})\succ, a_i \in Ag, h=1,2,...,n$ denotes the vector of satisfaction degree about all issues $\sum q_{xu}$ during the (h)th negotiated transaction in history.

4) Let $s(a_i,r,\sum q_{xu}), a_i \in Ag$ denotes the value of the satisfaction degree about all issues $\sum q_{xu}$ in current round of negotiation.

5) Let $\alpha_{q_x}=\{\alpha_{q_{x1}},\alpha_{q_{x2}},...,\alpha_{q_{xu}}\}$ denotes the information of preferences about issue q_{xu}, where $\alpha_{q_{xu}}$ is the preference coefficient of issue q_{xu}, which is acquired through reasoning based on negotiated transaction records about issue q_{xu} in the past. $\alpha_{q_{x1}} < \alpha_{q_{x2}} < ... < \alpha_{q_{xu}}$ can be explained that a negotiator agent will lay too much stress on the value of issue q_{xu}, while just the opposite for q_{x1}.

C. Concession Strategies

In a argumentation-based negotiation, no matter which form of argumentation (Reward, Threat, Defense) is presented [6], negotiator agent should consider to make concession on proposal at first. During the process of practical negotiation, a concession strategy based on time constraints may make negotiator loss opportunities for reaching an agreement because both sides are in a hurry to complete negotiation within the time constraint, while concession strategy based on opposite's preferences can help negotiator agent to have a definite object in view generate new proposal in accordance with the least action principle of belief revision, which will improve the efficiency of negotiation. This strategy is dominant strategy in a bilateral negotiation. By compared with concession strategy based on time constraint, the concession strategy based on opposite's preference has the high feasibility and application value. This conclusion will be formalized through the following propositions.

Under Hypothesis 1, agent will pursue individual maximum utility. The individual utility of a negotiator agent a_i in the r th round of negotiation is as follow,

$$U_r(a_i,q_x)=\sum U_r(a_i,q_{xu}) \quad (1)$$

where $U_r(a_i,q_{xu})$ represents the utility what Agent a_i acquire from the value of issue q_{xu}.

But if take the inherent correlation between two issues into consideration (e.g., price and quality: the higher price, the better quality), then (1) can be corrected through introducing a correlative coefficient as show below,

$$U_r(a_i,q_x)=\sum_{u\in K} U_r(a_i,q_{xu})+\sum_{u\in G} U_r(a_i,q_{xu})\left(1+\sum \eta\right) \quad (2)$$

where K is independent set of issues, G is the set of issues associated with other issues, η is the correlative coefficients. Besides, the utility value of single issue can be clarified as the following formula,

if the utility value of single issue increases with the increase of $c_r(a_i,q_{xu})$, then the formula is

$$U_r(a_i,q_{xu})=\left(c_r(a_i,q_{xu})-c^{\min}(a_i,q_{xu})\right)/\left(c^{\max}(a_i,q_{xu})-c^{\min}(a_i,q_{xu})\right) \quad (3)$$

 ISBN: 978-1-61208-139-7

if the utility value of single issue decreases with the increase of $c_r(a_i, q_{xu})$, then the formula is

$$U_r(a_i,q_{xu}) = \left(c^{\max}(a_i,q_{xu}) - c_r(a_i,q_{xu})\right) / \left(c^{\max}(a_i,q_{xu}) - c^{\min}(a_i,q_{xu})\right) \quad (4)$$

where $c^{\min}(a_i, q_{xu})$ represents the minimum value of issue q_{xu} what agent a_i would accept, $c^{\max}(a_i, q_{xu})$ represents the maximum value of same issue what opposite negotiator agent would accept that is considered by agent a_i. Then the Lemma can be presented through analyzing as show below.

Lemma 1. In the generating models of concession strategy based on single time constraint, the individual utility value of agent ($U_r(a_i, q_x)$) will increase with the increase of the value of single issue ($U_r(a_i, q_{xu})$).

Proof. The following conditions hold in accordance with (2),

$$\begin{aligned} U_r(a_i,q_x) &= f\left(U_r(a_i,q_{xu})|_{u\in K}, U_r(a_i,q_{xu})|_{u\in G}\right) \\ &= \sum_{u\in K} U_r(a_i,q_{xu}) + \sum_{u\in G} U_r(a_i,q_{xu})\left(1+\sum\eta\right) \\ &= \sum_{u\in K} U_r(a_i,q_{xu}) + \sum_{m=1}^{u} U_r(a_i,q_{xm}) + \sum_{m=1}^{u}\sum_{n\neq m}^{u} U_r(a_i,q_{xm})U_r(a_i,q_{xn})\eta_{mn} + \sum_{n=1}^{u} U_r(a_i,q_{xn}) \end{aligned}$$

Demanding $U_r(a_i, q_x)$ on the $U_r(a_i, q_{xu})|_{u\in K}$, $U_r(a_i, q_{xm})$ and $U_r(a_i, q_{xn})$ partial derivatives as follow,

$$\begin{aligned} &\frac{\partial U_r(a_i,q_x)}{\partial U_r(a_i,q_{xu})|_{u\in K}} = 1 > 0 \\ &\frac{\partial U_r(a_i,q_x)}{\partial U_r(a_i,q_{xm})} = 1 + \sum_{\substack{m=1 \\ n\neq m}} U_r(a_i,q_{xn})\eta_{mn} > 0 \\ &\frac{\partial U_r(a_i,q_x)}{\partial U_r(a_i,q_{xn})} = \sum_{\substack{m=1 \\ n\neq m}} U_r(a_i,q_{xm})\eta_{mn} + 1 > 0 \end{aligned}$$

therefore,

$$\begin{aligned} &\frac{\partial f\left(U_r(a_i,q_{xu})|_{u\in K}, U_r(a_i,q_{xu})|_{u\in G}\right)}{\partial U_r(a_i,q_{xu})|_{u\in K}} > 0 \\ &\frac{\partial f\left(U_r(a_i,q_{xu})|_{u\in K}, U_r(a_i,q_{xu})|_{u\in G}\right)}{\partial U_r(a_i,q_{xu})|_{u\in G}} > 0 \end{aligned}$$

The process above indicates the individual utility function ($f\left(U_r(a_i,q_{xu})|_{u\in K}, U_r(a_i,q_{xu})|_{u\in G}\right)$) is a increasing function for $U_r(a_i, q_{xu})$, i.e., $U_r(a_i, q_x)$ will increase with the increase of $U_r(a_i, q_{xu})$. Proof finished.

Further, the following proposition hold in accordance with (3) and (4).

Proposition 1. At the beginning of a negotiation, the initial proposal given by a negotiator agent is the proposal that make the individual utility value of negotiator reach maximum. The initial proposal is the following,

if the utility value of single issue increases with the increase of $c_r(a_i, q_{xu})$, then

$$c_{r=1}(a_i, q_{xu}) = c^{\max}(a_i, q_{xu}) \quad (5)$$

if the utility value of single issue decreases with the increase of $c_r(a_i, q_{xu})$, then

$$c_{r=1}(a_i, q_{xu}) = c^{\min}(a_i, q_{xu}) \quad (6)$$

Proof. Under Hypothesis 1**,** negotiator agent will pursue individual maximum utility. In the generating models of concession strategy based on single time constraint, the individual utility value of a negotiator agent is relevant to the current round of negotiation and the value of issue q_{xu}. Obviously, at the beginning of negotiation, if let

$$c_{r=1}(a_i,q_{xu}) = \begin{cases} c^{\max}(a_i,q_{xu}), \textit{Utility of single issue increases with the increase of } c_r \\ c^{\min}(a_i,q_{xu}), \textit{Utility of single issue decreases with the increase of } c_r \end{cases}$$

then $U_{r=1}(a_i, q_{xu}) \equiv 1$ according to (3) and (4), i.e., the utility value of single issue reaches maximum, at the same time, the individual utility value of agent a_i can also reach maximum in accordance with Lemma 1. Up to this point, (5)and(6)are proved to be reasonable. Proof finished.

Proposition 2. In the generating model of concession strategy based on single time constraint, the negotiator agent should generate new proposal in next round of negotiation($C_{r+1}(a_i,q_x) = \prec c_{r+1}(a_i,q_{x1}), c_{r+1}(a_i,q_{x2}),...,c_{r+1}(a_i,q_{xu}) \succ$)when the negotiation comes to a deadlock, where $c_{r+1}(a_i, q_{xu})$ is as follows,

if the utility value of single issue increases with the increase of $c_r(a_i, q_{xu})$, then

$$c_{r+1}(a_i,q_{xu}) = c_r(a_i,q_{xu}) - \xi(r)\left(c^{\max}(a_i,q_{xu}) - c^{\min}(a_i,q_{xu})\right) \quad (7)$$

if the utility value of single issue decreases with the increase of $c_r(a_i, q_{xu})$, then

$$c_{r+1}(a_i,q_{xu}) = c_r(a_i,q_{xu}) + \xi(r)\left(c^{\max}(a_i,q_{xu}) - c^{\min}(a_i,q_{xu})\right) \quad (8)$$

where $\xi(r)$ is the time strategy function based on single time constraint.

Proof. Time is precious to both parties from Hypothesis 4. When both parties can not reach an agreement after the current round of negotiation(r), one negotiator agent firstly

need to reduce the individual utility value($U_r(a_i, q_x)$)in order to break a deadlock and complete negotiation within time constraint(R), So the negotiator should make concession in certain issues. The new proposal should be made certain concession based on the value of issue ($c_r(a_i, q_{xu})$) in the current round of negotiation, which meet the description about limited rationality of agent in Hypothesis 2. Besides, the individual utility value reduction through concession should be as small as possible after each round of negotiation in order to make the individual utility of negotiator keep on a high level.

In accordance with Lemma 1, the individual utility value reduction of agent relates to the utility value reduction of single issue. So the individual utility value reduction can be determined by the utility function of single issue, see Figure 2, where the solid lines denote the utility function of single issue, the utility value reduction of single issue (ΔU) depends on the length of line segment d, d is determined by $c^{\max}(a_i, q_{xu}) - c^{\min}(a_i, q_{xu})$ (i.e. AB) and the time strategy function($\xi(r)$) based on single time constraint, i.e., $d = \xi(r)|OB - OA| = \xi(r)(c^{\max}(a_i, q_{xu}) - c^{\min}(a_i, q_{xu})) = |c_{r+1}(a_i, q_{xu}) - c_r(a_i, q_{xu})|$. Further, (7) and (8) are proved to be reasonable through combining with typical time strategies in the process of a pratical negotiation (uniform, radical and conservative).

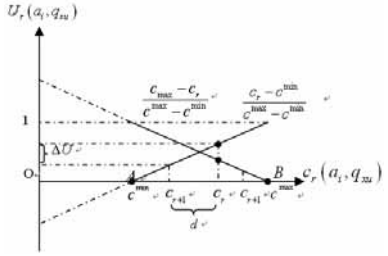


Figure 2

In particular, three typical time strategy function can be used in a negotiation based on single time constraint, as follows,

1) Uniform type. It is the time strategy function that make uniform concession with advance of negotiating process, e.g., $\xi(r) = \frac{1}{R}$.

2) Radical type. It is the time strategy function that make monotonic increasing concession with advance of negotiating process, e.g., $\xi(r) = \frac{1}{2^{(R-r)}}$, $2 \le r \le R$.

3) Conservative type. It is the time strategy function that make monotonic decreasing concession with advance of negotiating process, e.g., $\xi(r) = \frac{1}{2^{(r-1)}}$, $2 \le r \le R$.

Proof finished.

Proposition 3. In the model of concession strategy based on opposite's preferences, when the negotiation come to a deadlock, one negotiator agent can deduce the opposite's preferences from negotiated transaction records in history, and then determine issue and floting value of concession in the next round of negotiation according to the opposite's preferences. $c_{r+1}(a_i, q_{xu})$ is as follows,

if the utility value of single issue increases with the increase of $c_r(a_i, q_{xu})$, then

$$c_{r+1}(a_i, q_{xu}) = c_r(a_i, q_{xu}) - E\left(c^h_{r_{end}}\right)\left(c^{\max}(a_i, q_{xu}) - c^{\min}(a_i, q_{xu})\right) \quad (9)$$

if the utility value of single issue decreases with the increase of $c_r(a_i, q_{xu})$,then

$$c_{r+1}(a_i, q_{xu}) = c_r(a_i, q_{xu}) + E\left(c^h_{r_{end}}\right)\left(c^{\max}(a_i, q_{xu}) - c^{\min}(a_i, q_{xu})\right) \quad (10)$$

where $E\left(c^h_{r_{end}}\right)$ denotes step size of concession.

Proof. The both parties are sincere to reach agreement through negotiations, and there is not deceit during negotiation process from Hypothesis 5. Besides, in accordance with Hypothesis 3, the negotiator agent with incomplete information don't know the opposite's preferences information, and the preferences information can be acquired only through learning and reasoning based on negotiated transaction records in history. In the negotiation, if one negotiator agent can determine own behaviors according to the opposite's preferences, determine issue of concessionand step size of concession, and further generate a new proposal with argumention, then deceit in the negotiation can be avoided. In particular, after determining the issue of concession, $E\left(c^h_{r_{end}}\right)$ can be classified into two categories, relative step size of concession and absolute step size of concession, which are formally represented as follow,

1) Relative step size of concession,

$$E\left(c^h_{r_{end}}\right) = \frac{\sum_{h=h-k}^{h} c^h_{r_{end}}(a_i, q_{xu})}{k}$$

denotes the average concessional range of the issues'value($c_{r_{end}}(a_i, q_{xu})$)of k records before the(h)th negotiated transaction records in hitory, where $1 \le h-k < h \le n$.

2) Absolute step size of concession,

$$E\left(c^h_{r_{end}}\right) = \frac{c^{h=last}_{r_{end}}(a_i, q_{xu}) - c^{h=first}_{r_{end}}(a_i, q_{xu})}{n}$$

where $c^{h=last}_{r_{end}}(a_i, q_{xu})$ denotes the value of issue q_{xu} in final round of the (h)th negotiated record in history, $c^{h=first}_{r_{end}}(a_i, q_{xu})$ denotes the value of issue q_{xu} in first round of the (h)th negotiated record in history, n denotes the length of negotiated records in history which must meet the following condition, $n > 3$.

 ISBN: 978-1-61208-139-7

Besides, when a negotiation come to a deadlock, one negotiator agent can sort the preference coefficient of all issues after determining opposite's preferences of all issues, make concession in the certain issue with minimal priority among $\alpha_{q_{x1}}, \alpha_{q_{x2}}, ..., \alpha_{q_{xu}}$ according to principle of utility maximization, and then offer a new proposal in the next round of negotiation. If the new proposal is rejected, then the agent can adjust the proposal according to following steps,

1) Determining whether the preference coefficient ($\alpha_{q_{xu}}$) of certain issue (q_{xu}) has minimum priority or not, if so, the issue with the second-smallest priority will be used as new issue of concession.

2) The belief of issue with the second-smallest priority should be revised according to function of concession;the negotiator agent should make concession the issue with the second-smallest priority.

Facts show that agent can acquire much more individual utility if the agent comply strictly with the process of revising belief[5]. Up to this point, (9)and(10)is proved to be reasonable.

Besides, the value of every element in the preference coefficient set is determined by satisfaction degree in historical negotiated transaction records, so this paper will fit the mapping relation from satisfaction degree to preferences coefficient through PSO-RBFNN with the ability of approximation to random nonlinear functions in order to adapt to environmental changes.

IV. EVALUATION OF MODEL

The main task of this concessional strategy model is to determine the opposite's preferences to help agent to generate proposal with argumentation and further advance the process of negotiation. The preferences information can be acquired through PSO-RBFNN, so the neural network module in Figure 1 is the core module in the framework of argumentation-based negotiation. And this paper will select the following simulation experiment to demonstrate the feasibility of this module design.

A. *Traning*

Firstly, using a group of historical negotiated transaction records(including 21 records) that have been normalized as training samples of the RBF neural network. And only 4 typical issues (i.e., price, warranty, delivery date and method of payment) will be involved in the following experiment

In the training samples, the values of the 4 issues are used as input($c^h_{r_{end}}(a_i,q_{x1}), c^h_{r_{end}}(a_i,q_{x2}), c^h_{r_{end}}(a_i,q_{x3})$ and $c^h_{r_{end}}(a_i,q_{x4})$), i.e., there are 4 input neurons, the satisfaction degree of each issue and the global satisfaction degree of 4 issues are used as output($s(a_i,r^h_{end},q_{x1}), s(a_i,r^h_{end},q_{x2}), s(a_i,r^h_{end},q_{x3}), s(a_i,r^h_{end},q_{x4})$ and $s(a_i,r^h_{end},\sum q_{xu})$), i.e., there are 5output neurons. So the structure of the neural network is "$4-l-5$", and l is the number of hidden nodes which can be adjusted during the process of learning. Optimum parameters of this neural network are searched by using PSO algorithm to improve speed and accuracy of training. The convergence situation of key parameters is as shown in Figure 3 (via Matlab2010). The MMSE (Minimum Mean-Square Error) decreases with the increase of iteration times, finally the error converge to 10^{-3} in the 40462th generation. Then the RBF neural network that is used to fit negotiators' preferences can be constructed by using these optimum key parameters.

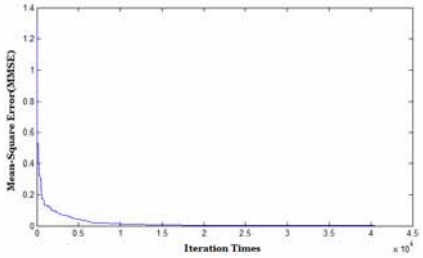


Figure 3 Process of parameters convergence

B. *Testing*

A group of testing samples including 11 records (see Table 1) are inputted into the neural network. The comparation between predictive output and ideal output of the testing samples is shown in Figure 4.

TABLE 1 TESTING SAMPLES

No.	q_{x1}		q_{x2}		q_{x3}		q_{x4}		*P.S*				
	d	χ	d	χ	d	χ	d	χ	q_{x1}	q_{x2}	q_{x3}	q_{x4}	Whole
1	300	0	6	0	7	0	1	0	0.87	0.25	0.14	0.2	0.471
2	290	0.25	12	0.33	1	1.00	1	0	0.9	0.5	1	0.2	0.73
3	280	0.50	24	1	3	0.67	1	0	0.93	1	0.33	0.2	0.758
4	280	0.50	18	0.67	2	0.83	2	0.33	0.93	0.75	0.5	0.4	0.737
5	280	0.50	6	0	1	1	3	0.67	0.93	0.25	1	0.85	0.732
6	275	0.63	24	1	3	0.67	1	0	0.95	1	0.33	0.2	0.766
7	275	0.63	12	0.33	1	1	1	0	0.95	0.5	1	0.2	0.75
8	270	0.75	18	0.67	3	0.67	3	0.67	0.96	0.75	0.33	0.85	0.76
9	265	0.88	18	0.67	2	0.83	4	1	0.98	0.75	0.5	1	0.817
10	265	0.88	12	0.33	1	1	3	0.67	0.98	0.5	1	0.85	0.827
11	260	1	18	0.67	1	1	4	1	1	0.75	1	1	0.925

In Table 1, q_{x1}, q_{x2}, q_{x3} and q_{x4} represent price , warranty, delivery date and method of payment , respectively. d and χ represent original value and normalized value. *P.S* represents predictive output. In column of payment method, "1~4" represent delivery on pay(D.O.P), upfront payment,credit card and pay on delivery(P.O.D), respectively.

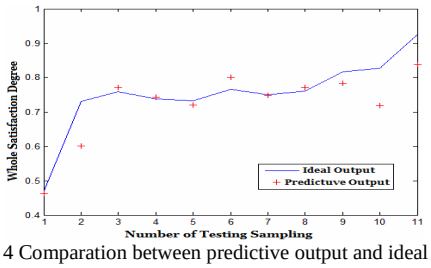


Figure 4 Comparation between predictive output and ideal output

In accordance with the output result of testing samples above, when the ideal(actual) output value of whole satisfaction degree is beyond 0.7, the predictive output value

of whole satisfaction degree is also beyond 0.7, thus the predictive effect is good. Besides, the preferences' characteristics shown in the testing samples are similar to the preferences'characteristics shown in the training samples, e.g., the whole satisfaction degree decreases with the increase of price, the buyer negotiator set a high value on other issues when the price has begun to level off. So the effect of simulation for preferences is good, this model has a certain feasibility and application value. Then negotiator agent can sort the preferences coefficient of all issues after determining opposite negotiator's preferences information, result as follows, $\alpha_{q_{x3}} < \alpha_{q_{x4}} < \alpha_{q_{x2}} < \alpha_{q_{x1}}$. Next, agent can determinate concessional issue (q_{x3}) through adhering to principle of individual utility maximization, and make concession on the issue in accordance with the model of concession strategy based on opposite's preferences mentioned above, further generate a new proposal with argumentaion and advance the process of negotiation.

V. Related Work

In recent years, there is an increasing amount of works on adaptive strategy in negotiation. Richter, Klusch and Kowalczyk suggested an adaptive strategy that bases on multistage fuzzy decision making. The bilateral negotiation strategies can allow agent to adapt its negotiation strategies and improve its individual payoffs by constructing a modelling of individual preferences as fuzzy goal and fuzzy constraints [4]. Wong and Wang proposed an ontology-mediated approach to organize the agent-based supply chain negotiation. Through equiping the agents with sophisticated negotiation knowledge that is structured by the usage of ontology, agents' negotiation behaviors will be more adaptive to various negotiation environments [5]. A number of researchers had attempted to use machine learning methods to optimal adaptive interaction strategies and their researches have several similarities to our own (e.g., each negotiator agent is given ability to reason opposite negotiator's private information). Oliverira and Rocha designed a virtual market and generate negotiation proposal by using a continuous reinforcement learning algorithm to enable agents to adjust themselves to the changing environment, including the opponent agents [6]. Sim and Guo presented a method that use the synergy of Bayesian learning (BL) and genetic algorithm (GA) to determine an agent's optimal strategy in negotiation (N) with incomplete information, called BLGAN. One agent can learn opponent's researve price (RP) and deadline through BL, reduce the size of search space for GA, then search and generate a optimum strategy at each negotiation round [7]. Although there are several similarities, our research differs in that we construct a concessional strategy model designed by PSO-RBFNN to allow negotiator agent to reason opposite negotiator's preferences information, determinte concessional issues and adapt negotiation strategies.

VI. Conclusion and Future Work

This paper proposes some hypotheses that support argumentation-based negotiation at first, then presents the generating model of concessional strategy based on single time constraint and the model based on opposite's preferences, respectively, the process of demonstration and result of experiment show that the latter designed by PSO-RBFNN has better abilities of learning and reasoning, which is dominant strategy in bilateral negotiations, and has a certain feasibility and application value. But the research in this paper is only a preliminary result, there are a lot of works to do in the future, and the emphasis is on appliance of the model to the environment of pratical business negotiation. In addition, how to introduce the factor of credit and trust into the model mentioned above to evaluate argumentation based on negotiation transaction records in history is also an important research direction in the future.

Acknowledgment

This research has been sponsored by NSFC Grant #71071005, 70940005.

References

[1] P. Philippe, R. Hollands, I. Rahwan, F. Dignum, and L. Sonenberg. An empirical study of interest-based negotiation. Journal of Autonomous Agents and Multi-Agent Systems, 2011, Vol.22, No.2, pp. 249-288.

[2] L. Amgoud and H. Prade. Formal handling of threats and rewards in a negotiation dialogue. NewYork, NY, USA: ACM Press, 2005, pp. 529-536.

[3] RYK. Lau. Context-sensitive text mining and belief revision for intelligent information retrieval on the Web. Journal of Web Intelligence and Agent Systems, 2003, Vol.1, No.1, pp. 1-22.

[4] J. Richter, M. Klusch, and R. Kowalczyk. A multistage fuzzy decision approach for modelling adaptive negotiation strategies. Proceeding of IEEE International Conference on Fuzzy Systems, 2010, pp.1-8.

[5] G. Wang, T. Wong, and X. Wang. Research on multi-lateral multi-issue negotiation based on hybrid genetic algorithm in e-commerce. Proceeding of 2nd IEEE International Conference on Information and Financial Engineering, 2010, pp. 706-709.

[6] E. Oliverira and A. Rocha. Agents advanced features for negotiation in electronic commerce and virtual organisation formation process. In: Dignum F Sierra C(eds): Agent Mediated Electronic Co-mmerce, The European AgentLink Perpective, Lecture Notes in Computer Science,Springer, 2001, pp. 78-97.

[7] K. Sim and Y. Guo, and B. Shi. BLGAN: Bayesian learning and genetic algorithm for supporting negotiation with incomplete information. Proceeding of IEEE Transactions On Systems, MAN, And Cybernetics Part B: Cybernetics, Vol. 39, No. 1, February 2009, pp.198-211.

 ISBN: 978-1-61208-139-7

The Empirical Analysis of a Web 2.0-based Learning Platform

András Benedek
Department of Technical Education
Budapest University of Technology and Economics
Budapest, Hungary
E-mail: benedek.a@eik.bme.hu

György Molnár
Department of Technical Education
Budapest University of Technology and Economics
Budapest, Hungary
E-mail: molnargy@eik.bme.hu

***Abstract*— The paper presents and analyses the new elements of instruction which have recently appeared as a consequence of the development of ICT (Information and Communications Technology) such as interactive and mobile communication tools or the applications contributing to 3D visual viewing at a university of significant traditions. The author also deals with the use of the special environment of networked learning based on atypical learning modes, the possibilities of instructional material and management development issues in details, as well as its typical effects on the active participants paying special attention to the forming new teaching and learning styles. Some extremely important parameters and their tendencies during the development decision making processes have been considered as essential by the lecturers involved by means of the application of different statistical algorithms and the webmining technology. Examining the learning activity of students of mainly traditional full-time courses in atypical learning environments, the possibilities of effective learning, network-based cooperation, students' attitudes and the development of key qualifications were evaluated in connection with our blended learning courses supported by electronic learning management systems. (An example of such a course is *Digital Pedagogy*.) By introducing this course attracting a high number of students (3-400 students per course), we wish to present our latest experimental developments, such as course topics, students' assignments, microcontent, blogs, and e-portfolio. The learning patterns of the present – day student popular on demands the integration of the atypical modes in to the teaching/ learning process. Such as electronic new generation learning environment is Moodle (Modular Object - Oriented Dynamic Learning Environment), whose selection was supposted by its open source character modular structure to be easily further – developed and its word wide popularity. The popular on using the atypical learning environment presented in the paper cannot be considered to be representative, however, can be the starting point of the further - development of our learning environments.**

Keywords - ICT; learning support systems; Moodle e-portfolio; web 2.0; online learning.

I. INTRODUCTION

The mathematical model of the description of network structures, such as the Erdős - Rényi model of the 1960s, has been known for a long time [1]. This model was used for the description for random networks consisting of nodes and the connecting lines, and could be describe by means of Poisson distribution. The BACON game [1], mainly popular in the US, was based on such an algorithm.

In accordance with Albert - László Barabássi's theory based on recent researches [1], the operation of complex networks can only be understood through the structures, which implies the interpretation of complex systems and the preparation of maps. All these, require the paradigm shift of human thinking to be able to interpret the available databases appropriately. The complex systems are more than the behaviour of a network. Such a system can be the human mind or an economic system, or the systems of learning processes. Furthermore the given complex systems changes with time, and we need to understand how it changes. The measurement and study of human behaviour patterns, learning behaviours can help to achieve this goal. It can also imply the measurement of human mobility or learning environments. The study of up-to-date ICT tools (as like smartphones, interactive systems, pc-s, multimedia systems), social webpages or electronic learning environments contribute to these measurements. These complex systems can be interpreted as scale independent networks, an example being the worldwide web or the learning network, whose understanding requires modern theories in addition to the already existing up-to-date ICT. So our task is to create such models. This is the science of the future. Our paper wishes to present on one hand the ICT environment, on the other hand student behavior patterns [1].

The classic educational environments have radically changed. It could be best characterized by the increasing use of atypical learning modes. The learning environments these learning modes rely on are supported by ICT technologies in a great degree. The 'learning space' can be compared to the physical behavior of light. Encouraging learning by activities can be considered as such impulses whose effects have been examined in the case of learning networks in the Moodle environment for a relatively great population of students. The impulses are such "flashes" which result in significant, although short student activities. It is supposed that the "flashes" can be developed into a flow of steady light in the learning space by means of appropriate techniques especially by encouraging horizontal activities. The paper calls attention to the above mentioned

 ISBN: 978-1-61208-139-7

phenomenon and formulates statements concerning the sustainability of the active phases of learning based on Moodle statistical analyses. The intensive growing of the students' horizontal communication (use of forums, chat rooms) and the increase of the frequency of learning communication in the informal learning space directly related to the formal learning process can have such a stabilizing effect on formal learning that can lead to the recognition of design and methodological relationships [2].

II. PROPOSAL

A. *ICT tools supporting instruction and learning*

The new strata of the information society grow up in the environment of the most up-to-date technology, so they might be called digital natives. They are familiar with IPTV, smart mobiles and homes, 3D TV, IPv6 network, thin client terminals, network communities (facebook, twitter, hi5) and Web 3.0 protocols [3]. This fact has such important potentials from the viewpoint of teaching and learning that should be relied on when designing and implementing courses. However, to be able to realize it successfully and effectively, we need to analyse the available results and experiences.

One of the new and at present innovative fields is the world of 3Ds. The problem of how 3D reality can be illustrated in one dimension has long been studied. The main goal of the technique is to provide an exact representation, a geometrical analysis, interpretation of spatial figures. In this sense, the two most important forms of visual communication are the representation and the interpretation of figures. Cognition, thinking and creation are interrelated in both forms of visual communication. Visual cognition means using pictures, images, that is, we wish to understand the ability of visual thinking starting from basic procedures. They imply such mental activities as analysis, synthesis or abstraction. Visual images are made dynamic by means of such procedures [4].

A classic example is the visionary turning, transformation of a spatial figure. Procedures by means of mental images can be interpreted as analogical processes of thinking, as opposed to the logical processes of conceptual thinking. If these procedures can become experience-like in a real physical environment, then the above mentioned thinking can be developed. Such a real environment is provided by the ICT system, the "Leonardo3Do", whose development increases Hungary's reputation. Students are able to create, formulate, cut off, and distort spatial figures, then improving their abilities of visual thinking and spatial visual abilities. In addition to it, the most important tool of visual seeing is the so-called bird, which is directly connected to its operator. In addition, the realization of visual thinking is strengthened by the so called active eye glasses the camera system and the control PC, most important tool of visual seeing (see Figure 1).

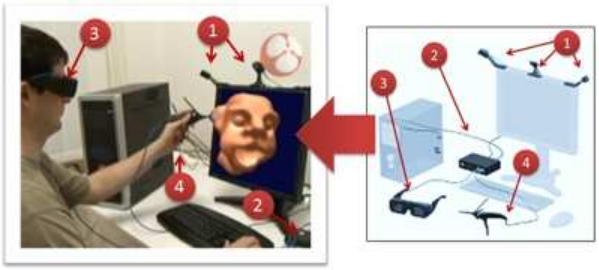


Figure 1. Leonar3Do [17]

Smart phones, IPADs and kinect units providing real life-like simulation, as well as kinect units, games, and the network-based web 2.0 and its applications (Google documents, Google, slide share, Google questionnaires, Google mobiles, Google calendars, Goggle blog, (Facebook, Twitter, Hi5, LinkedIn), whose peripheries are replaced by human fingers. The group of interactive online integrated tools (6^{th} sense technology), in which peripheries are replaced by human fingers is also a new direction of developments, equipped with a mobile small monitor instead of the traditional monitor, the PC by a PDA (See Figure 2) [15].

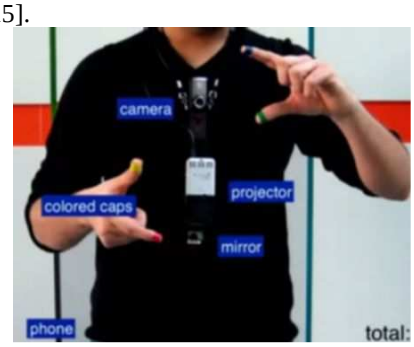


Figure 2. Interactive mobile system [18]

B. *Networked learning and its characteristics*

Having taken into consideration student attitude, learning characteristics, platform independence and the possibilities of development, the Department of Technical Education decided on introducing Moodle LMS system a few years ago. The word Moodle is an acronym of the notion Modular Object - Oriented Dynamic Learning Environment. Moodle is actually a kind of LMS (Learning Management System) application, more exactly a learning management system or an e-learning system shell integrated into the web 2.0 environment. The general task of the LMS is to identify its users and to match them with the adequate subject-matters (courses) according to their roles and rights. The network servers of these systems provide the right databases for system users and simultaneously log their activities and relevant learning data that can be used for generating statistics in the future. This data provide important information on the proficiency of students and also suitable for evaluation of the curricular effectiveness

 ISBN: 978-1-61208-139-7

[16]. Moodle is a web-based system, as it was mentioned before, so its use requires a PC with Internet/intranet and a browser access, a server and its URL given by the service provider. For example: http://mpt.moodle.appi.bme.hu/ (See Figure 3).

Figure 3. Opening page of the Moodle LMS of the Department, Source: own photo

The main page of the Department's Moodle LMS provides a list of all the courses of the Department structured by different categories such as academic years, types of programmes and whether the course is mandatory or optional.

The Moodle platform of the Department offers learning management services for 3,869 participants at present, and their number is increasing semester by semester.

The LMS provides detailed statistical data about the activity of the participants. The following figure shows the activity tendencies covering the past three years, with the red curve presenting the learners' activities. It is to be seen that the maximums of the amplitudes overlap with the dates of performance measurement, that is, the dates of assignment submission or test writing.

The paper presents the structure, characteristics, and the students' activities of an optional course, Digital Pedagogy preferred by the students and attracting a huge number of students. The course has been supported by the Moodle LMS for four years. The course is offered in a blended learning mode, as in addition to the face-to-face lectures, the students are provided the opportunity of networked content creation and that of synchronous communication.

As the above suggests networked communication starts at the beginning of the semester to harmonize topics, to have it approved by the lecturer and to have it assessed on both the students' and the lecturer's side. For this purpose, the system provides a set of activities, such as tests, activities, forums. The communication carried out on these channels can be followed in Moodle, but the participants also get a copy of each via e-mail, which makes the flow of information faster in a great degree [5].

It is possible to prepare log files of a given course – the last lesson of the course, activities, participation, as well as simple statistical curves - by means of the reporting motor of the electronic learning environment.

Figure 4 shows all the activities of all participants (teachers, students, and administrators) in a given semester, which well illustrates that the activity levels are the highest around the dates of assignments submission and tests (see Figure 4).

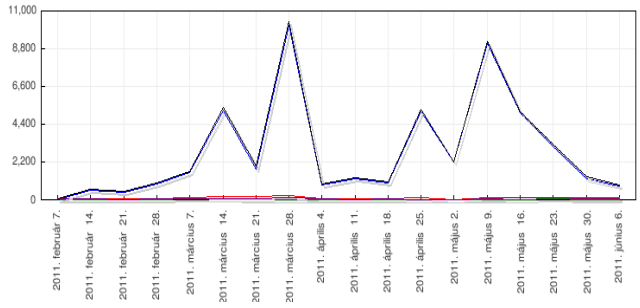


Figure 4. Activities of one semester of Digital Pedagogy course, Source: own photo

Out of the administration blocks of a given course the points report is an excellent way of having information about the full performance of a student in a given semester in a given course – assignments, topic selection, on-line tests, - with given points and comments. It also provides an opportunity of monitoring the student-student and student-lecturer horizontal and vertical communications. All student activities can also be summarized.

Our experiences of several years show that while the lecturers are mainly active in the daytime, the students are active more often than not late in the evening or at night. Since the system is connected to the participants' e-mail systems, most of the students stay online for 10-12 hours a day, so interactivity and quick feedback can be realized. The scope of activities to be carried out suggests that the students prefer the new generation activities, such as microcontent/related activities or the e-portfolio. The assignments are stored by the system, consequently they can be found any time. The following figure is an example of microcontent. The students' activities are stored at a delegated page, and are used to prepare an illustrative collection of them (see Figure 5).

 ISBN: 978-1-61208-139-7

Figure 5. An example of microcontent

The e-portfolio integrated into the Moodle system is very popular among the students. Each student was required to log in and join the group of students of Digital Pedagogy. After entering, the participants could share the uploaded materials and their contributions. The e-portfolio group was used by 253 students in a course, and each of them prepared 4-5 files and introductory portfolios in their profiles.

Both the e-portfolio and the profile editions are user-friendly, as they can be prepared easily and fast, e.g., by means of the clicking and dragging technique (see Figure 6).

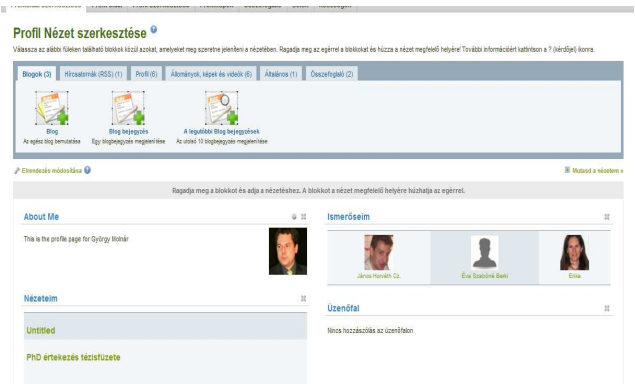


Figure 6. Editing e-portfolio profiles

The students can view their own uploaded and edited contents, on the other hand their own introductory profiles with the photos, as well as the students' acquaintances, groups, notice boards, which are the main forces of the new connectivity learning theories for the participants. By using them, the students can discuss important questions, requirements, solutions related to the course with their peers and their lecturers as well.

C. Web Mining Algorithms

The main objective of the investigation was to explore the most important characteristics of student behaviour or, in other words, the learning activity over the semester by means of web mining methods. The results presented in the previous chapter do not describe classic e-learning but blended learning integrating networked learning phases into the mainstream face to face education and combining the best of both forms of teaching and learning.

In introducing learning activities, two approaches were followed. On one hand, the role of the students' learning process was explored (macroanalysis), and on the other hand the microstructure of processing electronic syllabuses was investigated (microanalysis).

So-called offline web mining methods such as Google Analytics and SPSS Web mining for Clementine [7] and its web mining node are able to provide the possibility of a more profound, more comprehensive and more scientific analyses, far beyond descriptive statistics. It is an aggravating circumstance for the application of online tools that within the virtual learning environment several simultaneous courses are running. The isolation of related results as well as their survey is often too complicated. It is also difficult to perform analyses concerning the levels of the objects comprising the module and the pages comprising SCORM module and the html - based syllabus. Clementine is also capable of extracting related data from the log file. [9].

The application of web mining in VLE is an iterative cycle in which the excavated information and data should "enter the loop of the system and guide, facilitate and enhance learning as a whole, not only turning data into knowledge, but also filtering mined knowledge for decision making". The CRISP-DM (CRoss Industry Standard Process for Data Mining) as a well-known process model consists of six phases (See Figure 7) [12], [14].

 ISBN: 978-1-61208-139-7

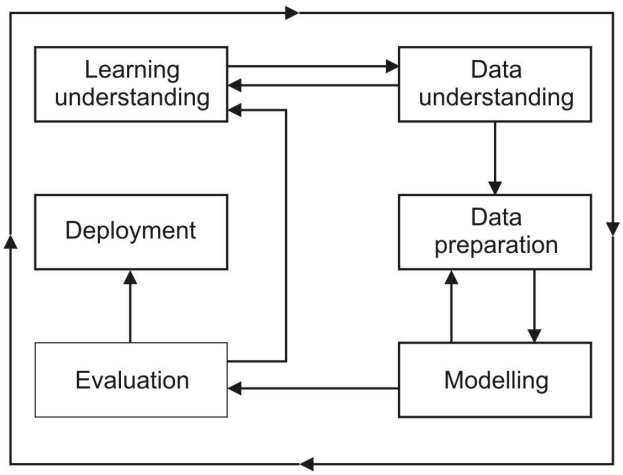


Figure 7. The CRISP-DM model [13]

III. CONCLUSION

The tracking of the work during the semester is supported by log files of Subject sheets. The tracking tools are ready-made and partly built into the learning environment; users' activities, their completed tasks and indices of achievements can be monitored by using statistics and graphs. Moreover, we carried out additional data analyses to obtain information on the progress in reaching the pedagogical goals connected with the course sheet usage. The Google statistical analyzer was used to prepare Figure 8 and Figure 9). The following table shows that how many students visited the the webpage of the course of Digital Pedagogy, how often and the amount of time spent there.

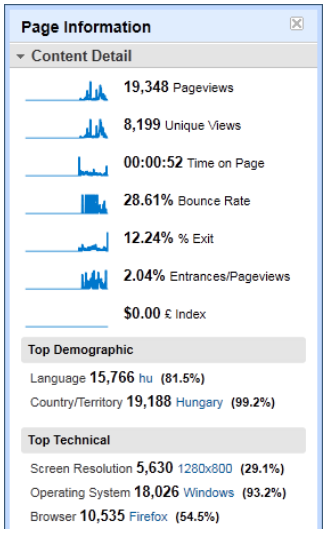


Figure 8. Participant activities of one semester of Digital Pedagogy course

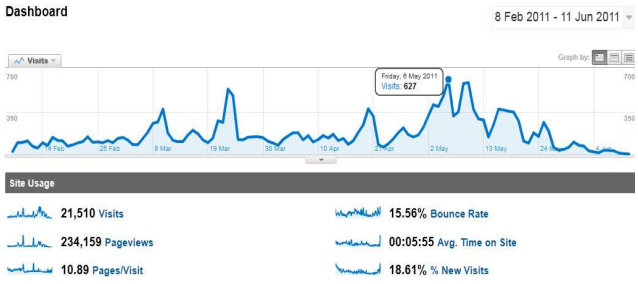


Figure 9. Student activities of one semester of Digital Pedagogy course with google analytics

The report on one single student's performance is an interesting and useful set of statistical data, as it contains the student's points collected during the course on the basis of his/her performance. By means of this, student performance can be evaluated individually and in a realistic way.

Both the students and the lecturers have the opportunity to have access to the detailed report of a student's activity in the function of actual measurements and their points and total points. This is also useful from the viewpoint of the students and the lecturers.

When running a course, all activities of the learning environment are recorded and stored in an appropriate database. Thanks to this data mass, the learning environment becomes suitable for preparing simple statistics without involving any special knowledge. For obtaining reports based on more complex data-correlations, additional data filtering and data combinations need to be done. Therefore, we used data mining techniques for deeper data analyses (SPSS Clementine 12.0 software with WebMining for Clementine) [7].

Using the SPSS Clementine data mining software the content of the log files produced during the use of course sheets can be processed. By doing so, more information can be obtained on the students' behaviour, their learning patterns and the usability of the interface to the electronic learning environment. By means of data mining it is possible to obtain information that could not be gained through other usability tests or queries. By the utilization of the results coming from data mining supported analyses of course sheets and curricula usage we are likely to identify effective education and learning processes and we will able to use the outcome during the development of the electronic curricula as well.

First of all, we formulated questions that could not be answered by means of integrated statistical tools of the learning environment. Then we attempted to answer these questions by using the appropriate data mining software (SPSS Clementine 12.0 software with WebMining for Clementine).

Most LMS have a built - in statistical module which makes possible the fixing of particular items of user activities. Such items can be entry data or download figures per subject item. However, these analysing methods are not qualified for revealing deeper relations or prediction of

 ISBN: 978-1-61208-139-7

learning/related events. This latter fact is of utmost importance for us, as it makes the learning process more predictable pedagogically. The user interface and usage of the program do not require IT or mathematics skills but that scope of interest could be an advantage (see Figure 10) [8], [10].

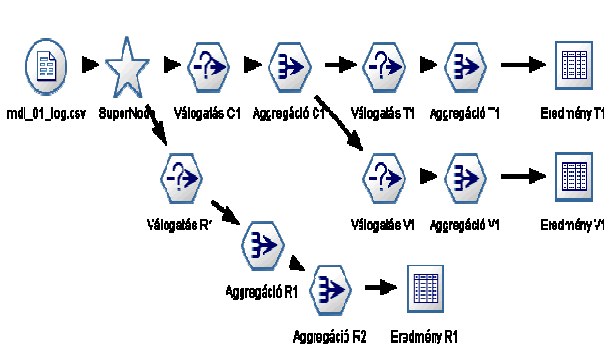


Figure 10. Data analysis with SPSS Clementine 12.0

The students can directly communicate with each other in the system by means of e-mailing, chatting or using forums, the direct communication channels are primarily built around the teacher-student relationship in the electronic earning environment. Our hypothesis, that on one hand the discrete flash-like momentums of networked learning are increasingly becoming continuous and steady, on the other hand the independent branches of the network will also be interrelated, is underlined by our experience and the earlier theories of learning.

We also made an interesting observation that a great deal of the services provided by the electronic learning environment, such as the communication channels of chatting and blogging were not at all used by the students during the learning process. On the contrary, the same services of informal social networking homepages, such as chat module, noticeboard, were used daily, furthermore every one-two hour. For example in the case of facebook, the content of the message can also be related to course content. In addition, the informal network-based learning processes in most cases are carried out either in an open or in a close group in relation to, e.g., a course [11].

Focusing on the future, the Department plans to integrate new forms of network-based communication, that is, community homepages preferred and frequently used by the new generation, such as, e.g., facebook, Hi5, Twitter. These can offer opportunities for fast and wide information provision for the students, furthermore by means of such a channel they like using. This way they learn in an informal way without it seeming to be some sort of external motivation. The Twitter has already been integrated both into the homepage and the Moodle system as well by the Department of Technical Education.

The tools described in the paper and the students' activity help understanding the scale independent networks, by means of which we can get to know the behaviour patterns of students, which can contribute to the orientation in the field of defining future routes of educational development. Educational policy could lead to the formation of a new learning model to be adapted to the new learning environment, based on networks and taking into consideration the learners' characteristics as well. Educational institutions could adopt this model.

Our experiences and our students' opinion suggest that this new – type networked learning support is needed log the present – day student population. Its signs are to be scan in the greatest student activity and in the continuous student – teacher interaction, which had a positive impact on student performance.

REFERENCES

[1] Barabási Albert-László: Behálózva – a hálózatok új tudománya (Networked – the new science of networks) (Magyar Könyvklub, 2003., ISBN 963547895X)

[2] András Benedek (ed.): Curriculum plan on the development of distance education and e-learning, National Institute for Vocational Education, Budapest, 2006.

[3] György Molnár: The requirements and development areas of the ICT aided learning environment, In: András Benedek (ed.): Digital pedagogy – Typotext Budapest 2008., pp. 225-255

[4] Péter Tóth: Seeing and Thinking: Spatial Intuition, Mental Operations, Visual Learning Lab, Budapest, 03.11.2010.

[5] András Benedek - Vidékiné Reményi, Judit (2010): Pedagogy studies in high-tech environment. ATEE 35th Conference, Budapest, Hungary, Responsibility, Challenege and Support in Teacher's Life-long Professional Development, Budapest, 26th-30th August, 2010, Programme and Abstracts, Ed.: Erzsébet Golnhofer, Magdolna Kimmel, Budapest, 2010. p.224.

[6] András Benedek (2009): Mobile Learning: New Horizons and Unstable Summits. Mobile Communication and the Ethics of Social Networking. Communications in the 21st century. Engagement and Exposure. Mobil Communication and Ethics of Social Networking. ed. by Kristóf Nyiri, Vienna: Passagen Verlag, pp. 247-253.

[7] András Benedek - György Molnár, János Horváth Cz. (2010): Moodle-based E-portfolio used in teacher training. (In: Sixth EDEN Research Workshop, User Generated Content Assessment in Learning. Enchanching Trasparency and Quality of Peer Production. Emerging Educational Technologies and Digital Assessment Methods, 24-27 October, 2010, Budapest, Hungary. Edited by Morten Flate Paulsen and András Szűcs, Budapest, 2010. ISBN 978-963-87914-4-3, pp. 131-133.

[8] András Benedek - György Molnár, János Horváth Cz. (2010): Jump over the shadow? From the traditional education to the non-typical one or the experiences of an electronic learning environment. IADIS Multi Conference on Computer Science and Information Systems, Freiburg, Germany, 26-29 july, Proceedings of the IADIS International Conference e Learning, Volume II (Ed. Miguel Baptista Nunes, Maggie McPherson) ISBN:978-972-8939-17-5, pp. 243-246.

[9] Pahl: Data mining technology for the evaluation of learning content interaction. International journal of E-Learning, Vol. 3, 2004, pp. 47-55

[10] SPSS Inc.: Web Mining for Clementine 1.5. User's Guide. NetGenesis, Chicago, 2005, p. 89

[11] György Molnár: ICT, network and mobile communication's solution in mirror of non tipical forms, In: Dr. Anikó

Kálmán(szerk.): Mellearn Conference Proceedings, 5. Hungarian National and International Lifelong Learning Conference - Strategies, Technologies and Methods in a Learning and Knowledge society - Sopron 2009. April 16-17., ISBN: 978-963-87523-9-0, Rexpo Kft., Debrecen, 2009. pp. 85-98

[12] Cristóbal Romero, Sebastián Ventura, Enrique García: Data mining in course management systems: Moodle case study and tutorial. Computers & Education, Vol. 51, No. 1. 2008, pp. 368-384

[13] Gregory Piatetsky - Shapiro: CRISP-DM: A Proposed Global Standard for Data Mining. The On - Line Executive Journal for Data - Intensive Decision Support, Vol. 3, No. 15, 1999

[14] Morten Flate Paulsen : Online Education Systems: Definition of Terms, In: Web - Education Systems in Eu, ZIFF Papiere 118, FernUniversität Hagen, Okt. 2002, pp. 23-28.

[15] McPherson, Maggie and Nunes, J.M. (2008). "Critical Issues for e Learning Delivery: what may seem obvious is not always put into practice". JCAL, **24(6)**, pp. 433-445.

[16] Piet Kommers, Piet A.M. (2010). ICT as explicit factor in the evolution of life-long learning. International journal of continuing engineering education and life-long learning, 20(1/2010), pp. 127-144.

[17] http://about3dtv.com/2010/07/leonar3do-create-3d-virtual-art-in-your-own-home/, downloaded: 15.07.2010.

[18] http://www.mobileUncle.com, downloaded: 15.07.2010.

 ISBN: 978-1-61208-139-7

The Effectiveness of Business Software Systems Functional Size Measurement

Beata Czarnacka-Chrobot
Department of Business Informatics
Warsaw School of Economics
Warsaw, Poland
e-mail: bczarn@sgh.waw.pl

***Abstract*—Execution of software Development and Enhancement Projects (D&EP), particularly those delivering Business Software Systems (BSS) as a product, encounters many problems, which still makes fulfilling of client requirements appear a big challenge for BSS providers. This may be proved by numerous analyses indicating exceptionally low effectiveness of BSS D&EP as compared to other types of IT projects, what - with their significant costs being considered - leads to the substantial financial losses. Author's analysis of fundamental BSS D&EP success factors indicates that one of the most important of them is objective and reliable measurement of such projects' product size, with particular consideration given to client's perspective. Further analyses led author to the conclusion that these conditions are only met by the software product size measure based on its functionality, what is confirmed by the acknowledgement of the so-called software Functional Size Measurement (FSM) concept along with several FSM methods by the ISO and IEC. The paper analyzes and evaluates the potential of effective usage of FSM with regard to BSS, with particular consideration given to the two most popular normalized FSM approaches, namely International Function Point Users Group (IFPUG) method and Common Software Measurement International Consortium (COSMIC) method. This issue is important not only for practical, but also for theoretical reasons, which are caused by the need to satisfy requirements of software engineering as a knowledge discipline having scientific grounds.**

Keywords-business software systems development and enhancement projects; functional size measurement; IFPUG method; COSMIC method; software engineering

I. INTRODUCTION

Majority of application Development and Enhancement Projects (D&EP) fail to meet criteria of their execution effectiveness. As indicated by the results of the Standish Group analyses, success rate for such projects has never gone beyond 35% [1]. It means that majority of them either end up with total failure, or they exceed costs and/or time estimated as well as they lack critical functions/features.

Analyses by T.C. Jones plainly indicate that those application D&EP, which are aimed at delivering Business Software Systems (BSS), have the lowest chance to succeed [2]. The Panorama Consulting Group, when investigating in their 2008 study the effectiveness of ERP (Enterprise Resource Planning) systems projects being accomplished worldwide revealed that 93% of them were completed after the scheduled time while as many as 68% among them were considerably delayed comparing to the expected completion time [3]. Comparison of actual versus planned expenses has revealed that as many as 65% of such projects overran the planned budget. Only 13% of the respondents expressed high satisfaction with the functionality implemented in final product while in merely every fifth company at least 50% of the expected benefits from its implementation were said to be achieved. Interesting comparisons of resolution results, cost overrun, time overrun, and expected ROI, made by the Standish Group with regard to three types of order processing application D&EP, are presented in Table 1 [4]. All these sample analytical results unequivocally indicate that, from the provider's perspective BSS D&EP are particularly difficult in terms of management.

Meanwhile BSS are not only one of the fundamental IT application areas; also their development/enhancement often constitutes serious investment undertaking: spending on BSS may considerably exceed the expense of building offices occupied by companies commissioning such

TABLE I. COMPARISONS OF RESOLUTION RESULTS, COST OVERRUN, TIME OVERRUN, AND EXPECTED ROI FOR THREE TYPES OF ORDER PROCESSING APPLICATION D&EP

Resolution	New application development	Package application with modifications	Application modernization
Resolution results comparison			
Successful	4%	30%	53%
Challenged	47%	54%	39%
Failed	49%	16%	8%
Cost overrun comparison			
Below 20%	43%	22%	46%
20% to 50%	21%	36%	29%
51% to 100%	10%	29%	14%
Over 100%	26%	13%	11%
Avarage overrun	44%	47%	34%
Time overrun comparison			
Below 20%	38%	27%	59%
20% to 50%	19%	32%	21%
51% to 100%	30%	31%	12%
Over 100%	13%	10%	8%
Avarage overrun	44%	45%	29%
Expected ROI comparison			
High	11%	34%	52%
Average	66%	57%	37%
Low	23%	9%	11%

Source: [4, pp. 4-6].

 ISBN: 978-1-61208-139-7

systems, and in extreme cases, even 50-storey skyscraper, roofed football stadium, or cruising ship with a displacement of 70.000 tons [5]. Yet quite often client spends these sums without supporting his decision on getting engaged in such investment by proper analysis of the costs, based on the rational, sufficiently objective and reliable basis. The above situation manifests itself in the difference in costs spent by various organizations on similar applications that may be even fifteen fold [6]. It comes from the fact that objective and reliable BSS D&EP effort estimation still appears to be a great challenge to the software engineering. Thus rational *ex ante* as well as *ex post* pricing of BSS, being of key significance to clients, encounters serious problems in practice. In the author's opinion the main reason of that problem is effort calculation on the basis of resources, while such activity should ground on the required (*ex ante* pricing) and actually delivered (*ex post* pricing) BSS size, which determines work effort, what was proved in [7].

Exceptionally low effectiveness of BSS D&EP as compared to other types of IT projects (e.g., maintenance, support, package acquisition, implementation projects, projects delivering other types of software), especially with their costs being considered, leads to the substantial financial losses, on a worldwide scale estimated to be hundreds of billions of dollars yearly, sometimes making even more than half the funds being invested in such projects (see e.g., [8][9]). What is more, analyses of The Economist Intelligence Unit indicate that there is strong correlation between delays in delivery of software products and services and decrease in profitability of a company therefore failures of BSS D&EP, resulting in delays in making new product and services available and in decreasing the expected income, represent threat also to the company's business activity [10].

Author's analysis, which concerned numerous studies on factors of BSS D&EP effectiveness, available in the subject literature, leads to the conclusion that among fundamental factors are [11]:

1) Realistic planning, with particular consideration given to the reliable and objective estimates for key project attributes (work effort, execution time and cost), what requires BSS size measurement (see e.g., [12]).

2) Proper project scope management, above all consisting in undertaking small projects, that is projects whose product is characterised by relatively small size (expressed in appriopriate software product size units), what also requires BSS size measurement.

3) Authentic involvement of client in the project (both users and managers) - thus BSS size measurement should be carried out by taking into consideration mainly the perspective of the client, that is with the use of software product size units that are of high significance to him.

The above considerations lead to the conclusion that significant factor of BSS D&EP effectiveness is objective and reliable measurement of their product size, with particular consideration given to client's perspective. As proved in [11], these conditions are being fulfilled only by the measure of product size based on its functionality.

What underlay the search for the right measure of software product size, having been undertaken for several decades already, was not only the need to increase the effectiveness of software D&EP execution but also the requirements of software engineering as a discipline of knowledge where quantifiable approach to software development/enhancement should be of key significance. Definition of software engineering adopted by the Institute of Electrical and Electronics Engineers (IEEE) reads that it is „the application of a systematic, disciplined, quantifiable approach to the development, operation, and maintenance of software" [13]. Quantifiable approach means that the measurement of software processes and products should constitute immanent feature of this discipline of knowledge. That's why over the last couple of years significant intensification of works could have been observed, which aimed to standardize the best practices, especially with regard to software products. Various ISO/IEC (International Organization for Standardization/ International Electrotechnical Commission) norms have been developed as a result of these works, filling an important gap in the software engineering [14]. One of the most important groups of ISO/IEC standards concerns software product size measurement.

II. STANDARDIZATION OF FUNCTIONAL SIZE MEASUREMENT

Many years' verification of reliability and objectivity of various approaches towards software size measurement showed that what for now deserves standardization is just the concept of software size measurement based on its functionality – being an attribute of first priority to the client. The concept of the so-called software Functional Size Measurement (FSM) was normalized in the six-part standard ISO/IEC 14143 [15]. First of all, this standard specifies definition of functional size, which is understood as „size of the software derived by quantifying the Functional User Requirements", while Functional User Requirements (FUR) stand for the „sub-set of the User Requirements describing what the software does, in terms of tasks and services" [15, Part 1]. Hence functional requirements in this norm, due to their importance and need to ensure objectivism of measurement, are treated disjointly when combined with other requirements of non-functional character. The elementary unit of FUR used for measurement purposes is called Base Functional Component (BFC). The example of a FUR could be "Maintain Customers", which may consist of the following BFC: "Add a new customer", "Change customer details" and "Delete a customer" [15, Part 1]. On the other hand, Functional Size Measurement Method (FSMM) in the discussed standard was defined as a specific FSM implementation defined by a set of rules, which conforms to the mandatory features of such measurement.

According to the ISO/IEC 14143 norm the process of using FSMM should comprise the following steps [15, Part 1]: (1) defining the scope of FSM, (2) identifying the FUR contained within the scope of FSM, (3) identifying the BFC contained within the FUR, (4) classifying the BFC with

 ISBN: 978-1-61208-139-7

regard to their type, (5) assigning appropriate value to each BFC, and (6) calculating functional size.

There are about 25 variants of the FSM techniques having been developed, however only five of them have been now acknowledged by the ISO/IEC as conforming to the rules laid down in the ISO/IEC 14143 norm and certified as international standard, namely: (1) International Function Point Users Group (IFPUG) method, which is approved in the ISO/IEC 20926 standard [16]; (2) Mark II (MkII) function point method proposed by the United Kingdom Software Metrics Association (UKSMA), which offers more detailed measurement comparing to the IFPUG method and is normalized in the ISO/IEC 20968 standard [17]; (3) Netherlands Software Metrics Association (NESMA) function point method, being the simplified version of IFPUG method, which is approved in the ISO/IEC 24570 standard [18]; (4) Common Software Measurement International Consortium (COSMIC) method, which is certified in the ISO/IEC 19761 standard [19]; and (5) FSM method developed by the Finnish Software Metrics Association (FiSMA), which is normalized in the ISO/IEC 29881 standard [20].

The first three methods listed above are accepted by the ISO/IEC not in full versions, as proposed by the organizations developing them, but in part, however in the most important part with respect to the software functional size measurement [15, Part 6] – that is why they are called the first-generation FSMM. In the approaches proposed by IFPUG, UKSMA and NESMA these methods involve also delineating of the so-called Value Adjustment Factor (VAF), which is supposed to adjust functional size being measured with the use of Unadjusted Function Points (UFP) to the environment of specified project by taking technical and quality requirements (i.e., requirements of non-functional character) into consideration [21, Part 5]. Yet this part of these methods has not been approved by the ISO and IEC – as these organizations' assumptions exclude the fact of FSM depending on requirements of this type. On the other hand, the COSMIC and FiSMA methods were recognized as international standard entirely [15, Part 6][20] – that is why they are called the second-generation FSMM.

FSM methods accepted by ISO/IEC differ in terms of software measurement capabilities with regard to various categories of software (i.e., so-called functional domains). Thus prior to choosing given method one should assess its adequacy to the type of software product. In the ISO/IEC 14143 norm it is stated, that [15, Part 6]:

- There are no functional domains constraints for the accepted part of the IFPUG and NESMA methods, although they were developed as intended for measurement of BSS functional size, nor for the FiSMA method.
- The UKSMA method is adequate for the measurement of any type of software provided that the so-called logical transactions may be identified in it. The rules were developed as intended for BSS therefore the method supports neither complex algorithms characteristic of scientific and engineering software nor the real-time systems.
- The COSMIC method is adequate for: data-driven systems (i.e., BSS), time-driven systems (i.e., real-time systems), and hybrid solutions combining both the above (e.g., real-time systems of airline tickets booking). On the other hand there are constraints for software with complex mathematical algorithms or with other specialized and complex rules (e.g., expert, simulation, weather forecasting systems) and for software processing continuous variables (e.g., computer games, musical instruments software).

The ISO/IEC 14143 norm adheres to the ISO/IEC 15939 standard [22], determining general rules and procedures for the software measurement process in compliance with the ISO/IEC 15288 norm [23], which, on the other hand, defines processes of the system's life cycle. One of the steps of the size measurement process defined in the ISO/IEC 15939 standard is procedure of selecting a method that will be used to measure its size. According to this procedure, selection of FSM method being best tailored to the client's needs should consist of the following activities: (1) characteristics of organizational units of software user with regard to the measurement process, (2) identification of their information needs towards measurement process, and (3) selection of appropriate FSMM on the basis of prospective methods identification (for more details see [24]).

Requirements towards appropriate FSM method vary depending on the organization's character. For example, financial institutions usually choose the method, which correctly measures the BSS while chemical company, by reason of its basic activity, would rather require measurement method being suitable for the real-time systems. Choosing method adequate to the needs would also depend on how its result is planned to be used. If an organization intends to use the measurement results also for the purpose of comparing its productivity against industry data, it is recommended to choose the method being relatively popular in the given industry, for which such data exist. In the case it only needs cursory, rough estimation of functional size, the requirements towards appropriate method of its measurement will get reduced (see also [24]).

ISO and IEC allow for selecting method other than the methods approved by them yet they recommend that it conforms to definitional part of the ISO/IEC 14143 norm. It is also recommended to carry out measurement with the use of relevant supporting tools (see e.g., [12]).

III. FUNCTIONAL SIZE MEASUREMENT OF BUSINESS SOFTWARE SYSTEMS

Thus to measure BSS functional size one may use all FSMM normalized by the ISO/IEC. What's more, this is the need to measure BSS size that was at the basis of FSM concept and methods development. In the context of their FSM it is assumed that software systems of this type are characterized by the following properties [25]:

- Basic purpose of BSS is to acquire, collect and make available data concerning business activity to support this very activity by: keeping data in the ordered way, enabling execution of various inquiries

 ISBN: 978-1-61208-139-7

and delivering information supporting the decision-making.

- Functionality of BSS usually is dominated by the need to collect business data of differentiated level of structure complexity and to ensure their integrity and availability in the long run.
- Overwhelming majority among the so-called functional users of BSS are persons (in contrast to things: other software, devices, hardware) who usually enter into direct interaction with the system through the input/output devices. It means that considerable part of functionality is directed towards right proceedings in case of mistakes being made by this type of users and towards helping to use BSS efficiently. If other, equivalent applications or their components cooperate with the measured system, then they also – next to persons – are functional users of this application.
- Different BSS may cooperate (e.g., exchange data) either on-line or in a batch mode.
- In BSS data are usually collected historically, i.e., after the events that took place in real world, taking into consideration the time of current answer and the fact it is given by a person. Data may be processed also in the batch mode. As a rule BSS do not include software used to drive the real-world events in the real time, which is characteristic of real-time systems although it happens that they receive data in the real time (e.g., prices on the market) – as a result they are forced to respond in similar way.
- Business rules governing data manipulation may be sometimes complicated however BSS rarely include a large number of complex mathematical algorithms.
- BSS usually reside in one layer of software, however application layer software depends on software located in other layers (e.g., operating system, device drivers) – otherwise it could not have been used.
- BSS perceived by functional users being persons as an individual application in fact may consist of several equivalent components. Thus separate measurement for each of them may turn out necessary. This applies to the situation where the goal of the product FSM is to get its size for the effort estimation while each component is based on different execution technology.

All FSM methods normalized by the ISO/IEC allow, among others, for:

- Expressing BSS size from the perspective of its functionality - software attribute being of first priority to the client, what promotes his involvement in the project – and this is a fundamental factor of BSS D&EP success (see e.g., [1][8][9]).
- Comparing the actually delivered BSS size with the size required by client, what enables to evaluate the realized project with regard to the actual value of product delivered (for more details see [7]).
- Making BSS size independent of technology used in the project execution – since functional size reflects actual capabilities of the system, which are independent of programming language used.
- The way of calculating BSS size that is independent of the development methodology and of the project's life cycle models as well as of project constrains and developer's experience [26].
- Obtaining sufficiently objective and reliable estimates not only for BSS size, but also for D&EP work effort, cost and completion time relatively early in the project's life cycle [27] – since early estimates of BSS functional size can be based on incomplete FUR (see e.g., [12][26]).
- Estimating size, effort, cost and time of each change to the BSS functional user requirements.
- Determining the effort, cost and time of all project stages - since the BSS size is based on FUR and these are them that decide on the effort.
- Obtaining appropriate economic indicators - since the use of BSS functional size indicates increased productivity in case of the reduction of total costs, resulting from using more efficient programming language (withdrawing the paradox of software size programming units).
- Supporting CMMI-DEV (Capability Maturity Model Integration for Development [28]) - since the FSM is a factor that makes it easier for an organization to achieve subsequent levels of maturity [29].

IV. THE IFPUG METHOD VERSUS COSMIC METHOD

The two most popular normalized FSMM dedicated to business software systems are IFPUG method and COSMIC method. There are obviously certain similarities between them, which most of all include (see e.g., [30][31][32]):

- Common FSM concept, based on similar understanding of the measurement purpose and scope as well as definition of the measured software boundaries.
- The rules of both methods are based on similar, yet not identical, meaning of the terms related to data. What also is convergent is the concept of data transformation as well as of users perceived as recipients of the measured software functionality.
- Occurrence of two phases of measurement: identification of elements, on the basis of which the functional size is determined, and actual measurement, in which they are mapped into this numerically-expressed size. In the IFPUG method, the first of these phases is not described explicite yet it assumes that the measurement is based on the FUR - data models, functions/processes models or windows, screens, forms and reports designs may also be used for this purpose. In the phase of actual measurement, the explicitly described rules of this method are employed towards these elements. In the COSMIC method, the measurement phase proceeds solely on the basis of FUR.
- Similar way of expressing FUR. In both methods, FUR are expressed by means of BFC. In the

 ISBN: 978-1-61208-139-7

approach developed by IFPUG there are 5 types of BFC: Internal Logical Files (ILF), External Interface Files (EIF), External Inputs (EI), External Outputs (EO), and External inQuiries (EQ) [16], whereas in the COSMIC method there are 4 types of BFC: entry, exit, read, and write [19]. However, there is no simple analogy between them as in the COSMIC method data are not measured explicite and they are not distinguished as a type of BFC.

- Both approaches, although in a different way, meet the requirements imposed on FSM methods in the ISO/IEC 14143 norm therefore both were recognized as international standards of FSM (the IFPUG method not in full version - see [16] vs. [21]).

Differences between the discussed methods mostly concern the following:

- Rules of measurement. Fundamental difference in this respect is the fact that the IFPUG method includes general system characteristics (VAF), representing the influence of technical and quality requirements (i.e., requirements of non-functional character) on functional size. This is the reason why this approach has not been approved by ISO/IEC entirely, however taking them into account in calculations is not necessary. What's more, studies have revealed low practical usefulness of VAF to increasing the quality of prognoses. Characteristics of this type do not exist in the COSMIC method where measurement is based solely on FUR.
- Size boundaries for processes/functions. In the IFPUG method, the size of all five BFC is arbitrarily limited thus the software size depends on their number. While in the COSMIC approach there is no upper limit for the process size as it is determined by the number of data movements. On the other hand, the size of COSMIC data movement is 1 CFP (COSMIC Function Point) and does not depend on data to which it pertains, which is the case of processes in the IFPUG method.
- Data inclusion manner. In the IFPUG method, data are included in calculations in a twofold way: separately as internal/external logical files and as file type referenced affecting the process size. In the case of COSMIC method, data are included with each data movement of read or write type of BFC. Thus the use of IFPUG method requires constructing of data model, which in the COSMIC approach is not indispensable however proves useful. In the IFPUG method, data model also provides basis for early estimates while in the COSMIC approach this is process model that is employed for the approximation purposes.
- Benchmarking data resources. Current version of the largest repository with benchmarking data concerning software FSM, that is repository of International Software Benchmarking Standards Group (ISBSG) [33], includes data in nearly 80% pertaining to the software products being measured with the use of IFPUG method while in only 8% to those measured with the use of COSMIC method.

Moreover, in the subject literature, however, in most cases being supported by COSMIC, the following features of this method are pointed out as deciding on its advantage over IFPUG method:

- Broader range of application. The IFPUG method was developed in order to measure BSS, however in its current version no constraints with regard to the measurement of other functional domains were imposed by ISO/IEC. Meanwhile it is often argued that this method does not prove useful in the case of real-time systems size measurement – unlike COSMIC method [31]. According to the author of this paper, such conclusion goes too far both from theoretical and practical point of view although measurement of this type of software using IFPUG method undoubtedly is more complicated as compared to the COSMIC method and therefore it may be less accurate. In publications on the IFPUG method one may find not only the rules but also the examples of employing it in the measurement of real-time systems size (see e.g., [34]).
- Compliance with object-oriented analysis and programming. In this case it is argued that if the COSMIC method was developed much later than IFPUG method it then takes into account modern techniques of FUR description and systems construction, paying attention mostly to the object-oriented approach [35]. However, this in no way proves that there is no possibility to calculate functional size using object-oriented approach to the development based on the IFPUG method – rules of the method and practical examples do indicate such a possibility exists (see e.g., [36]).
- Broader measurement perspective. With the use of IFPUG method, functional size is measured from the perspective of end user while with the use of COSMIC method – from the point of view of the so-called functional user that next to an end user includes also developers, who perceive other applications and devices interacting with the measured software [31]. Perspective limited to an end user only carries some danger of skipping in the calculations of such functionality, which is imperceptible to an end user, however on condition that it is assumed that only a user being a person can be a recipient of functionality. Meanwhile, recognizing the IFPUG method as complying with the ISO/IEC 14143 standard means that definition of user it currently employs is consistent with this notion's definition included in this norm, wherein a user is understood not only as a person but also as a thing (e.g., other applications, devices) that interacts with the measured software [15, Part 1].
- COSMIC approach assumes that typical software is made of layers, for which the rules of proceedings were expressed explicite therefore this method can

 ISBN: 978-1-61208-139-7

be used to measure complex, layered architectures [26].

- In COSMIC approach there are no artificial limits imposed on the size of functional process, that's why the integrity of measure is very good, while in the IFPUG method artificial limits (e.g., weights) limit the size of BFC, so the integrity is limited [26].
- Possibility of faster delivery of results. COSMIC method happens to be regarded as more intuitive, more concise and simpler than IFPUG approach, which should result in quicker delivery of the measurement outcome. Yet this has not been confirmed by the surveys, which indicated that there are no significant differences in the quickness of measurement made with the use of both methods [32]. What is more, even authors of the COSMIC method admit that in case one needs quick measurement with low-quality user requirements, it is simpler (and quicker) to employ IFPUG method – which results from the limited scope of its BFC size, which are easier to be predicted correctly [35]. In this situation using the COSMIC method would require an expert in order to obtain result on the same level of reliability, while this would increase the effort of measurement process. It is worth noting that it applies to the possibility of employing both methods for the estimation purposes: in the original COSMIC method there are limited possibilities to carry out approximate calculations at the early stage of the project's life cycle, or the way of obtaining such calculations is time-consuming, which results from the necessity to base on FUR specification of high level detailness. However, there are some its variants that allow for early estimates of functional size on the basis of incomplete FUR (e.g., Object-Based Approach, Story-Based Approach, and Event-Based Approach) [26].

Organizations employing IFPUG method may face the need to converse results obtained with the use of this method into the results expressed in COSMIC function points (CFP), e.g., due to client's requirements or because of this other method being recognized as more adequate in specific application conditions. And inversely: it may happen there will be the need to converse sizes expressed in CFP into the functional size expressed in IFPUG function points, e.g., due to the need to use generally available benchmarking data for comparison purposes as a serious argument in negotiations with client. However, there is no possibility of exact conversion of the results of both methods using mathematical formula. This results from the fact that BFC of the IFPUG method cannot be exactly translated into BFC of the COSMIC method, and inversely, as well as from the above mentioned differences in measurement rules.

One of the approaches towards conversion is conversion based on statistical formula. Many studies have been initiated in this area, which aimed at gaining adequate statistical formula that would be expressing correlation between the sizes obtained with the use of both methods, however their outcomes differ to a large extent, which may be seen in the examples shown in Table 2. Thus it is advised that an organization facing the need of conversing sizes of its applications makes its own analysis using the regression method in order to derive equation specific to itself, at the same time relying on the right size of the sample of benchmarking data concerning measurement as well as on statistically representative examples, being specific to given organization.

TABLE II. EXAMPLES OF STATISTICAL FORMULAS FOR THE CONVERSION OF THE IFPUG/NESMA METHOD RESULTS INTO THE COSMIC METHOD RESULTS[A]

Author (year)	Sample size	Size range (UFP IFPUG/ NESMA)	Formula (regression analysis)	R^2
Fetcke (1999)	4	40–77	$CFP = 1.1 \times UFP - 7.6$ (*UFP* – number of IFPUG Unadjusted Function Points)	0,97
Vogelezang Lesterhuis (2003)	11	39–1424	$CFP = 1.2 \times UFP - 87$ $CFP = 0.75 \times UFP - 2.6$ (≤200 *UFP*) $CFP = 1.2 \times UFP - 108$ (>200 *UFP*) (*UFP* – number of NESMA Unadjusted Function Points)	0,99
Abran, Azziz, Desharnais (2005)	6	103–1146	$CFP = 0.84 \times UFP + 18$ (*UFP* – number of IFPUG Unadjusted Function Points)	0,91
Desharnais, Abran (2006)	14	111–647	$CFP = 1.0 \times UFP - 3$ (*UFP* – number of IFPUG Unadjusted Function Points)	0,93
Van Heeringen (2007)	26	61–1422	$CFP = 1.22 \times UFP - 64$ (*UFP* – number of NESMA Unadjusted Function Points)	0,97

[A] R^2 is a coefficient of determination describing degree to which the model explains the shaping of the variable being explained – in this case expressing the proportion of deviation in the COSMIC size (in CFP) being explained by the change in the IFPUG/NESMA size.

Source: Author's own analysis based on [32] and [37].

The issue of conversion of the results of both methods requires further investigations, even more as there are no works concerning their latest versions.

Naturally, the discussed methods are not devoid of shortcomings, though. Among objections being most often raised towards both approaches is their relatively high complexity. In fact, when analyzing the rules of both methods (see [21] and [25]) it is hard not to agree with this argument, even despite the existence of various tools supporting them (see e.g., [12]). However, the studies reveal that the time devoted by a specialist (the so-called scope manager) to the delineating of functional size is estimated to be less than 1% of the project life cycle duration [38]. As specialist makes use of the information, which is collected within any methodically-conducted project (especially of data model and functions/processes model), regardless of whether or not the estimation of its attributes is intended. Whichever option is used, the work effort, cost and time involved in the execution of functional size measurement is

meager comparing to these attributes for the entire life cycle of BSS D&EP, not to mention comparing them with possible effects of erroneous decisions. What's more, it's hard to expect that measurement of products of high complexity - and BSS undoubtedly are such products – is going to be effective yet simple task. Other methods of product size measurement are either simple and ineffective (methods based on programming units) or they are neither simple nor effective, especially with regard to the planning purposes (methods based on construction complexity units), not to mention that they lack usefulness from the client's perspective (see also [11]). Hence it is worth to treat the costs of employing FSMM as an investment in the improvement of the software processes in an organization.

V. THE OBJECTIVITY AND RELIABILITY OF BUSINESS SOFTWARE SYSTEMS FUNCTIONAL SIZE MEASUREMENT

FSM methods, despite relatively high complexity, are used worldwide more and more often (see e.g., [39]), clearly due to their proved effectiveness, especially in case of BSS. For instance, in UK, the Mark II method is a method recommended in the execution of D&EP for the needs of public administration. On the other hand, COSMIC method is a national standard in Japan and in Spain. What's more, these methods are widely employed not only by providers but by clients as well [26].

Research into the objectivity and reliability of software size estimation in functionality units early in D&EP life cycle was carried out by the ISBSG [27]. It analyzed data for 130 projects having product size ranging from 11 to 20 000 function points – and the method decidedly most often used for calculation purposes was the IFPUG method. The researches concerned both prognosis of thus expressed size on the basis of data model as well as methodical calculation of the size on the basis of FUR specification. The ISBSG argued that the product size expressed in functionality units in both cases is estimated in sufficiently objective and reliable way comparing to the calculations made on the basis of end product, whereas estimates derived on the basis of FUR specification are characterized by higher reliability than those obtained on the basis of data model: in the first case 70% of estimates proved not lower than actual product functional size while in the second case this coefficient amounted to 62% of such estimates. In both cases allowable estimation error was assumed only at the level of ±10%. With regard to the objectivity of product size estimation using functionality units it is stated that two specialists obtain results that differ only by ±10%, however on the condition that requirements are specified properly [40]. Conclusions coming from the ISBSG research are also confirmed, among others, by M. Parthasarathy [41, p. 292], who pointed out objectivity and relatively high reliability of the product size estimation based on FSM.

Attempt to carry out similar research was also made by the author of this paper. Within the surveys that aimed at analysing the level of FSMM usage by the Polish BSS providers as well as the reasons behind this *status quo*, she managed to obtain data, which in the case of three BSS providers (small IT company, medium-sized IT company and IT department in a bank) allowed for the IFPUG method reliability analysis (for more details see [39]). For this method the analysis of prognoses accuracy was made in comparison with actual end product size based on the number of unadjusted function points: (1) estimated on the basis of data model and function model, with average complexity being assumed for the function depending on its type; (2) calculated in accordance with method's recommendations on the basis of FUR specification. All products considered in the analysis (provider 1: 11 products, provider 2: 14 products, provider 3: 11 products) are rather relatively small business applications (up to 600 IFPUG UFP). When analyzing reliability, prediction accuracy indicator *PRED(RE)* was employed, which serves to express what in the surveyed cases is the percentage share of these projects/products whose estimates are contained within the assumed estimation Relative Error (RE) related to the actually received value [42]. Thus in order to consider a method reliable the *PRED(30)* was assumed on the level not lower than 80% [43]. What also was calculated is the *PRED(10)*, in order to compare prediction accuracy level with surveys conducted by the ISBSG, in which allowable estimation error was assumed on the level of ±10%.

As indicated by the author's analysis, the IFPUG method in the case of calculations made - according to method's recommendations - on the basis of FUR specification meets the assumed reliability condition. Yet this method does not meet the prediction accuracy condition in any of the analyzed cases, if estimation is made on the basis of data model and function model with average complexity being assumed for the functions. Thus the research results confirm that better effects may be achieved if calculations are made on the basis of FUR specification, which is consistent with the conclusion coming from the ISBSG analyses. Yet the obtained results appear worse in comparison with the ISBSG report - this may result from the fact that in the survey the Polish providers presented the author with data coming from BSS D&EP chosen by chance (not from the best projects, which was probably the case of ISBSG) as well as from scantier experience in using FSMM in Poland.

The FSMM standardized by the ISO/IEC provide also sufficiently objective and reliable basis for BSS D&EP effort, budget and time frame estimating. Results of numerous surveys, including, e.g., those carried out by the State Government of Victoria [44] and International Software Benchmarking Standards Group [27], indicate that BSS D&EP in case of which the FSMM were used for effort planning, are characterized by relatively accurate estimations. Studies by the State Government of Victoria indicate that pricing of BSS on the basis of product size expressed in functionality units results in reducing the average budget overrun to less than 10%. The ISBSG report confirms these results: in the situation where the methods based on product functional size are employed in making cost estimation, in 90% of cases the estimates differ from the actual costs not more than by 20%, and among these very cases 70% are accurate to within 10%. Also analysis of the results of 25 studies concerning the reliability of the most important BSS D&EP effort estimation methods, made

 ISBN: 978-1-61208-139-7

by the author on the basis of the subject literature, revealed that currently the highest accuracy of effort estimations is delivered by the effort estimation methods based on BSS size expressed in functionality units [12].

VI. CONCLUSION AND FUTURE WORK

The measurement of software products is an area of software engineering, which cannot be considered as sufficiently mature not only in terms of practice, but also in terms of knowledge maturity [45]. That's why over the last couple of years significant intensification of works could have been observed, which aimed to standardize the best practices of software products measurement. Some of the undertakings have only just gained recognition, which may be proved by the fact that the latest version of the CMMI-DEV model was strongly focused on measurement. Also various ISO/IEC norms have been developed as a result of these works, filling an important gap in the software engineering (for more details see [14]). One of the most important groups of ISO/IEC standards concerns the software product FSM. Normalization of such measurement is aimed mainly at reducing unnecessary diversity in the area of software size measures, ensuring compatibility between the standardized approaches as well as their usefulness, especially for business software systems.

The ISO/IEC standards for the FSM methods, like the ISO/IEC 14143 norm for the FSM concept, adhere to other standards. The ISO/IEC 15939 offers help in defining the set of measures being adequate to the specific informational needs yet it neither provides the list of such measures nor it recommends specific set of measures for the D&EP. Therefore one may find the opinion that although employing of rules described in this standard is necessary for the measurement process implementation in the organization, these rules *per se*, however, are not sufficient for this purpose [46]. Thus this standard should be linked with other normalized measurement approaches, e.g., the IFPUG method or the COSMIC method.

As indicated by the above analyses, it is hard to unequivocally decide on the advantage of the COSMIC method over the IFPUG method (or inversely) – both have strengths and drawbacks, coming up in the specified problem areas, both have supporters and adversaries. Most probably, COSMIC approach will not totally replace the IFPUG method in the nearest future as this first-generation method has proved being sufficiently objective and reliable approach, at least with regard to the business software systems [47]. Since both approaches prove useful to BSS, this is not the author's intention to solve this dilemma. In any case, from the perspective of requirements made for the methods of BSS size measurement there are no significant differences between the COSMIC and IFPUG method. Generally speaking, functional size obtained with the use of both methods constitutes sufficiently appropriate measure of BSS size and the basis for the estimation of BSS D&EP work effort, cost and duration. These methods, however, are not free of disadvantages therefore they need further improvement, which should benefit to higher accuracy of prognoses being obtained through the methods. Yet the differences between them are significant enough so that they cause problems in proper conversion of their results.

That's why in further works attention should be mostly paid to the possibility of working out the rules of conversion between the results gained with the use of various FSMM, especially these two most popular methods of BSS size measurement. Also surveys that aimed at analyzing the level of using the FSMM by the Polish BSS providers as well as the reasons behind this *status quo* [39] will be continued to keep observing the changes while the research area will be extended as much as possible to other Polish dedicated BSS providers and other BSS D&EP scope management aspects, with particular consideration of the FSMM reliability.

Undoubtedly, the issue of BSS size measurement is important both for pragmatic as well as for theoretical reasons. Pragmatic reasons rise from the need to increase effectiveness of the execution of BSS D&EP. On the other hand, theoretical reasons are provoked by the need to satisfy requirements of software engineering as a discipline of knowledge – without the possibility of measurement of its basic objects, ensuring objective and reliable analytical criteria, it is hard to regard it as a discipline having scientific grounds. Hence strong significance of appropriate software products' size measurement methods arises, especially with regard to BSS having the lowest chance to succeed.

REFERENCES

[1] Standish Group, "CHAOS summary 2009", West Yarmouth, Massachusetts, 2009, pp. 1-4.

[2] T. C. Jones, Patterns of software systems failure and success, International Thompson Computer Press, Boston, MA, 1995.

[3] PCG, "2008 ERP report, topline results", Panorama Consulting Group, Denver, 2008, pp. 1-2.

[4] Standish Group, "Modernization – clearing a pathway to success", West Yarmouth, Massachusetts, 2010, pp. 1-16.

[5] T. C. Jones, "Software project management in the twenty-first century", Software Productivity Research, Burlington, 1999, pp. 1-19.

[6] State Government of Victoria, "southernSCOPE, reference manual", Version 1, Government of Victoria, Melbourne, Australia, 2000, pp. 1-22.

[7] B. Czarnacka-Chrobot, "Rational pricing of business software systems on the basis of functional size measurement: a case study from Poland", Proc. of the 7th Software Measurement European Forum (SMEF) Conference, T. Dekkers, Ed., Libreria Clup, Rome, Italy, June 2010, pp. 43-58.

[8] J. Johnson, "CHAOS rising", Proc. of 2nd Polish Conference on Information Systems Quality, Standish Group-Computerworld, 2005, pp. 1-52.

[9] Standish Group, "CHAOS summary 2008", West Yarmouth, Massachusetts, 2008, pp. 1-4.

[10] Economist Intelligence Unit, "Global survey reveals late IT projects linked to lower profits, poor business outcomes", Palo Alto, California, 2007: http://www.hp.com/hpinfo/newsroom/press/2007/070605xa.html (03.12.2008).

[11] B. Czarnacka-Chrobot, "The economic importance of business software systems size measurement", Proc. of the 5th International Multi-Conference on Computing in the Global Information Technology (ICCGI 2010), 20-25 September 2010, Valencia, Spain, M. Garcia, J.-D. Mathias, Eds., IEEE Computer Society Conference Publishing Services, Los Alamitos, California-Washington-Tokyo 2010, pp. 293-299.

 ISBN: 978-1-61208-139-7

[12] B. Czarnacka-Chrobot, "The role of benchmarking data in the software development and enhancement projects effort planning", in New Trends in Software Methodologies, Tools and Techniques, H. Fujita and V. Marik, Eds., Proc. of the 8th International Conference SOMET'2009, Frontiers in Artificial Intelligence and Applications, vol. 199, IOS Press, Amsterdam-Berlin-Tokyo-Washington, 2009, pp. 106-127.

[13] IEEE Std 610.12-1990: IEEE standard glossary of software engineering terminology, The Institute of Electrical and Electronics Engineers, New York, 1990.

[14] B. Czarnacka-Chrobot, "The ISO/IEC standards for the software processes and products measurement", in New Trends in Software Methodologies, Tools and Techniques, H. Fujita, V. Marik, Eds., Proc. of the 8th International Conference SOMET'2009, Frontiers in Artificial Intelligence and Applications, vol. 199, IOS Press, Amsterdam-Berlin-Tokyo-Washington, 2009, pp. 187-200.

[15] ISO/IEC 14143 Information Technology – Software measurement – Functional size measurement – Part 1-6, ISO, Geneva, 2007.

[16] ISO/IEC 20926 Software and systems engineering - Software measurement - IFPUG functional size measurement method 2009, edition 2, ISO, Geneva, 2009.

[17] ISO/IEC 20968 Software engineering – Mk II Function Point Analysis - Counting practices manual, ISO, Geneva, 2002.

[18] ISO/IEC 24570 Software engineering – NESMA functional size measurement method version 2.1 - Definitions and counting guidelines for the application of Function Point Analysis, ISO, Geneva, 2005.

[19] ISO/IEC 19761 Software engineering – COSMIC: a functional size measurement method, edition 2, ISO, Geneva, 2011.

[20] ISO/IEC 29881 Information Technology – Software and systems engineering – FiSMA 1.1 functional size measurement method, ISO, Geneva, 2008.

[21] IFPUG, "Function point counting practices manual, release 4.3", Part 0-5, International Function Point Users Group, NJ, January 2010.

[22] ISO/IEC 15939 Systems and software engineering - Measurement process, ISO, Geneva, 2007.

[23] ISO/IEC 15288 Systems and software engineering - System life cycle processes, ISO, Geneva, 2008.

[24] B. Czarnacka-Chrobot, "Standardization of software size measurement", in Internet – Technical Development and Applications, E. Tkacz, A. Kapczynski, Eds., Advances in Intelligent and Soft Computing, vol. 64, Springer-Verlag, Berlin-Heidelberg, 2009, pp. 149-156.

[25] COSMIC, "The COSMIC functional size measurement method, version 3.0, guideline for sizing business application software (version 1.1)", Common Software Measurement International Consortium, Québec, May 2008.

[26] G. Rule, "The most common Functional Size Measurement (FSM) methods compared", Software Measurement Services, St. Clare's, Mill Hill, Edenbrige, Kent, UK, 2010, pp. 1-8.

[27] ISBSG, "The ISBSG report: software project estimates – how accurate are they?", International Software Benchmarking Standards Group, Hawthorn, Australia, 2005, pp. 1-8.

[28] CMMI Product Team, "CMMI for Development, version 1.2", Software Engineering Institute, Carnegie Mellon University, Pittsburgh, 2006, pp. 1-573.

[29] C. A. Dekkers and B. Emmons, "How function points support the Capability Maturity Model Integration", in CrossTalk. The Journal of Defence Software Engineering, February 2002, pp. 21–24.

[30] J. Cuadrado-Gallego, D. Rodríguez, F. Machado, and A. Abran, "Convertibility between IFPUG and COSMIC functional size measurements", Proc. of the 8th International Conference on Product-Focussed Software Process Improvement, PROFES 2007, Riga, July 2007, pp. 273–283.

[31] G. Xunmei, S. Guoxin, and Z. Hong, "The comparison between FPA and COSMIC-FFP", Proc. of Software Measurement European Forum (SMEF) Conference, Rome, Italy, 2006, pp. 109–117.

[32] H. van Heeringen, "Changing from FPA to COSMIC. A transition framework", Proc. of Software Measurement European Forum (SMEF) Conference, Rome, Italy, 2007, pp. 143-154.

[33] ISBSG, "Data demographics release 11", International Software Benchmarking Standards Group, Hawthorn, Australia, June 2009, pp. 1-24.

[34] International Function Point Users Group: http://www.ifpug.org/publications/case.htm, Case 4 (03.04.2011).

[35] Common Software Measurement International Consortium: http://www.cosmicon.com/advantagecs.asp (8.06.2009).

[36] International Function Point Users Group: https://www.ifpug.org/publications/case.htm, Case 3 (03.04.2011).

[37] COSMIC, "The COSMIC functional size measurement method, version 3.0, advanced and related topics", Common Software Measurement International Consortium, Québec, December 2007.

[38] P. Morris, "Functional size metrics", Total Metrics, October 2001.

[39] B. Czarnacka-Chrobot, "Analysis of the functional size measurement methods usage by Polish business software systems providers", in Software Process and Product Measurement, A. Abran, R. Braungarten, R. Dumke, J. Cuadrado-Gallego, J. Brunekreef, Eds., Proc. of the 3rd International Conference IWSM/Mensura 2009, Lecture Notes in Computer Science, vol. 5891, Springer-Verlag, Berlin-Heidelberg, 2009, pp. 17–34.

[40] International Software Benchmarking Standards Group: http://www.isbsg.org/Isbsg.Nsf/weben/ Functional%20Sizing%20Methods (21.04.2008).

[41] M. A. Parthasarathy, Practical software estimation: function point methods for insourced and outsourced projects, Addison Wesley Professional, 2007.

[42] D. V. Ferens and D. S. Christensen, "Does calibration improve predictive accuracy?", in CrossTalk. The Journal of Defence Software Engineering, April 2000, pp. 14—17.

[43] A. Abran and P. N. Robillard, "Reliability of function points productivity models for enhancement projects (a field study)", Proc. of Conference on Software Maintenance 1993-CSM-93, Montreal, IEEE Computer Society Press, Los Alamitos, 1993, pp. 80-97.

[44] P. R. Hill, "Some practical uses of the ISBSG history data to improve project management", International Software Benchmarking Standards Group, Hawthorn VIC, Australia, 2007, pp. 26-30.

[45] L. Buglione and A. Abran, "The software measurement body of knowledge", Proc. of Software Measurement European Forum (SMEF) Conference, Rome, Italy, January 2004, pp. 84–94.

[46] L. Bégnoche, A. Abran, and L. Buglione, "A measurement approach integrating ISO 15939, CMMI and ISBSG", Proc. of Software Measurement European Forum (SMEF) Conference, Rome, Italy, 2007, pp. 111–130.

[47] F. Vogelezang, "COSMIC Full Function Points. The next generation of functional sizing", Proc. of Software Measurement European Forum (SMEF) Conference, Rome, Italy, March 2005, pp. 281–289.

 ISBN: 978-1-61208-139-7

A Proposal Metaprocesses as Software Assets in the Telehealth Domain

Javier Fernandez
University of Granada
Department Languages and Informatics Systems
Granada, Spain
jfernandez_9@ugr.es

Maria Hurtado
University of Granada
Department Languages and Informatics Systems
Granada, Spain
mhurtado@ugr.es

Freddy Duitama
University of Antioquia
Department Systems Engineering
Medellin, Colombia
freddy.duitama@udea.edu.co

Jose Garrido
Department Languages and Informatics Systems
University of Granada
Granada, Spain
jgarrido@ugr.es

***Abstract*— The main purpose of this research is to develop and implement a metamodel process (Metaprocess) that serves as a repository to be reused in TeleHealth domain specific processes, being therefore identified as a useful tool for the development of agile and efficient applications. The work will be well justified as it will allow the efficiency validation of metaprocesses use and its reuse in the early stages of software application development in the field of Telemedicine. The intent is to provide a response to the research question regarding how to enhance reusability during the early stages of software development through the use of Meta-processes as software assets, in the particular domain of Tele-Health. This question stems from evidence that even through the great boom of business process system specifications and business process services, these systems have a great number of problems such as the lack of formal standards and appropriate specification metamodels in the business processes logic to meet their rules and specifics domains. Likewise, the existence of multiple work schemes and Metaprocesses representations do not consider the planned reuse of software assets in the early stages of development and formalization schemes that enable its validation for different domains.**

Keywords- Metaprocess; Software Assets; Telehealth.

I. INTRODUCTION

The utilization of meta-processes as software assets that can be used in early stages of software development according to [6] [9] and [26], requires methods and models of formalization and validation that take into account the supported development process, as the intent is to develop them into frames of reference that facilitate the agile construction of software through the use of metamodels without departing from the knowledge of domain specific rules that need to be systematically arranged.

In recent years, meta-process, software assets and reuse have become increasingly popular as an inexpensive and promising means for software development methods. However, without creating a new set of issues and trade-offs. The performance and resource management of software development are becoming a very crucial phase for future generation of applications.

Now, in the framework for the development of this work, progress has been made on the bibliographical review and comparative analysis on the use of Metaprocesses as software assets, resulting in the development of a first article on thematic literature review. Lastly, progress has been made in the Metaprocesses conceptual review and software assets, as well as in the knowledge of the chosen Telehealth domain, particularly information regarding the study of protocols and guidelines for cardiovascular risk programs.

First, Section 2 describes the conceptual background about business process, software assets, business metaprocess, business process modeling and process models automation. Section 3 explains the background research about principal topics in the work and Section 4 presents our work in progress with some results so far.

II. CONCEPTUAL BACKGROUND

In this section, we provide the relevant definitions for the state of the art research in the context of business processes and metaprocesses as software assets and provide a better understanding of the research problem.

A. Business Processes

As outlined in [1], [2] [3] [4] and [5], a business process can be defined as an organizational entity that exists depending on the occurrence of events to accomplish a specific purpose, it is governed by rules that allow the control and monitoring of the activities that shape it, the latter are assigned resources and roles to help meet the corporate objective for which they are defined.

 ISBN: 978-1-61208-139-7

B. Software Assets

In [6] [7] [8], software assets are software components that meet the following criteria: a) They are a collection of artifacts to provide a solution to a problem b) They may be used and reused in different contexts c) They can be extended and customized in several places.

C. Business Metaprocesses

Rolland et al. [9] proposes that business metaprocesses can be defined as process models that serve as benchmarks to be instantiated in different domains.

D. Business Process Modeling

Basu et al. [10] show that the business processes are usually represented as flows or flow charts or activity charts, sometimes as nodes or rules for execution of tasks or activities.

E. Process Models Automation

In [11] [12], a business process system is designed as a generic type of software supported by design techniques focused on operational business processes management, it is known by the requirements that must support the process and it is represented by a graphical scheme.

III. BACKGROUND RESEARCH

Some research studies on business process modeling techniques show us the impact they have had on the construction of business metamodels as representational frameworks and business process building models for their use and implementation at the organizational level.

Along these lines, the approaches described in the work of [13], [14] [15] and [16] present to us how modeling techniques utilize a wide variety of symbolic and representational constructors to model integrated work systems at the conceptual level. While each technique uses a unique method for proposing what it believes to be the essential elements of an integrated work system, many of the concepts used in these techniques are very common: goals, roles, actors (or agents), activities (process tasks), interactions (or conversations), workflow, resources, information and resource interdepencies are primarily based on the concept of interaction or exchange. However, these common elements give us the fundamental components of the proposal and construction of Metaprocesses as frameworks grounded in the use of business process models to instantiate in different domains.

Moreover, the techniques described in [17], [18] and [19] demonstrate that the proposed business process modeling, if indeed they propose common work schemes for business process modeling, exclusively of the platforms used and oriented to specific domains, they mostly lack from mechanisms of validation and transformation of models originating from the definition of business rules to the construction of domain-specific metamodels, that contribute to the construction of metaprocesses, which is also accompanied by the absence of formalisms to propose strategies and domain model validation to ensure consistency between the model and its domain without losing the level of reference to be achieved using metaprocesses. The proposals presented by [17], [18] and [19] focus specifically on the following aspects: [17] recognizes the importance of developing software based on software product lines, whose cornerstone should be the reutilization of previously developed components or software assets that can be instantiated for different domains, although the research conceptually structures the development framework and software evolution towards a more adaptive, reusable and agile process, it does not deliver a schematic framework that allow us to address the problem of reusability at the early stages and especially validation as the key to verify the quality of the software. Meanwhile, [18], as it well indicates that a suitable scheme of work in our case could be (MDE) model-driven engineering, focusing our attention primarily on the use and validation of models, combining the concepts of domain specific languages and models transformation engines to completely cover the life cycle of a software system, it does not go any further than the construction and transformation of application models independently of the platform used; therefore not addressing the issue of validation or the reuse in an appropriate scheme of quality assurance software during the early stages. Finally, [19] along with the above mentioned and including the proposal of [18] on the use of BPMN as a standard notation for modeling construction, it ends up proposing the use of design patterns through language definition of software architectures or ADLs mixed with the use of attribute based architecture styles (ABAS) and ISO quality schemes to ensure software quality of the end product, it primarily focuses in the final stages of construction and ignores the issues of software evolution and quality assurance from the early stages, it also lacks clear verification, validation and reuse mechanisms that would allow the proposed patterns to be software construction metaprocesses.

In addition to the above mentioned and reviewing the contributions of [20], [21] [22] [23], [24] and [25] we see how business process modeling proposals for specific domains such as Metaprocesses, mostly lack from clear validation and formalization mechanisms independently from the use of standards and tools for the representation of multiple domains, leaving aside the principle of models reusability and expression thereof.

Finkelstein et al. [26] show some methodologies of software development focused on processes as: EPOS: Expert System for Program and System Development [27], SOCCA: Specifications of Coordinated and Cooperative Activities [28], MERLIN: Supporting Cooperation in Software Development Through a Knowledge-Based Environment [29], OIKOS: Constructing Process-Centred SDEs [30], ALF: A Framework for Building Process – Centred Software Engineering Environments [31], ADELE-TEMPO: An Environment to Support Process Modeling and

 ISBN: 978-1-61208-139-7

Enaction [32], SPADE: An Environment for Software Process Analysis, Design, and Enactment [33], PEACE: Goal –Oriented Logic – Based Formalism for Process Modeling [34], E^3: Object – Oriented Software Process Model Design [35], PADM: Towards a Tool Process Modeling System [36]. Once the methodologies have been verified, it could be concluded that the EPOS, PWI and PEACE methodologies are the most well rounded methodologies, in terms of the factors such as software process modeling approach, process modeling languages, use of metaprocess and engines. And evaluating elements such as purpose, core component types, modeling core component types, languages, property modeling language, tools as support of software process models, analysis and design process modeling, customization, instantiation and evolution, support for rendering, internal representations and process engine architecture.

Given the above, it is necessary that the proposal to be developed on methodologies and processes as software assets gather the earlier results, in the sense that it must meet primarily but not exclusive to the most relevant parameters presented in the studied methodologies, meaning it should be focused on their implementation and interaction with computerized and human agents, considering the existing relations and entities, having fundamentally a formal and object-oriented component, being expressive and representative, easy to interpret and compiled, also allowing a syntactic definition of the models, presenting a method, allowing through its concepts and mechanisms pave the way to go from a generic model to a model adapted to a specific domain and presenting meta-activities that allow its instantiation, as well as being fundamentally deterministic, that it cannot be affected by the human agents intervention, preserving consistency and allows to easily identify inconsistencies, it also should allow the use of external tools to be integrated and keep the model objects consistency, likewise allow the internal representation of software process modeling and allow the use of several engines for process management.

IV. RESEARCH

The main purpose of this research is to present a strategy by which Meta-processes are developed into Software Assets in the Tele-Health domain. The intent is to provide a response to the research question regarding, How to enhance reusability during the early stages of software development through the use of Meta-processes as software assets in the particular domain of Tele-Health? This question stems from evidence that even through the great boom of business process system specifications and business process services, these systems have a great number of problems such as the lack of formal standards and appropriate specification metamodels in the business processes logic to meet their rules and specifics domains. Likewise, the existence of multiple work schemes and Metaprocesses representations do not consider the planned reuse of software assets in the early stages of development and formalization schemes that enable its validation for different domains.

This research has a social impact in terms about: Health crisis-solutions for resource optimization, possible reduction of medical errors and ease at notational level. And scientific impact in: Reuse in early stages, metaprocess instantiation in different domains and oriented software development models.

This research has specific objectives as: Consider proposals for use of Metaprocess as assets and reuse software in early stages. Conceptualize on the use of Metaprocess methodologies as software assets and its reuse in early stages. Identify the components of telehealth processes for the construction of a Metaprocess. Define the Metaprocess in the context of model driven software development in the field of telehealth and Validate the applicability of the proposed Metaprocess in the reusability of the components in the specific domain of Telehealth. Now, in the framework of the proposed strategic solution for the development of this research, progress has been made on the bibliographical review and comparative analysis on the use of Metaprocesses as software assets, resulting in the development of a first article on thematic literature review. Progress has been continuously made in the review of some aspects related to the Metaprocesses domain; therefore, presenting a first draft of conceptual scheme based on the metamodel of the proposed metaprocess.

Lastly, progress has been made in the Metaprocesses conceptual review and software assets, as well as in the knowledge of the chosen Telehealth domain, particularly information regarding the study of protocols and guidelines for cardiovascular risk programs.While there are existing medical protocols and guidelines for patient care, our research have allowed us to evidence the absence of standard protocols to triage patients with cardiovascular risk; therefore, the metaprocesses proposal along with its respective metamodels allow from its instantiation and personalization, according to the patients characteristics, adapt to the requirements and needs thus contributing to reduce medical errors that may occur in this first phase of telecare.

ACKNOWLEDGMENT

We thank ARTICA research of Colciencias, TICs Ministery and COOPEN research European Community and University of Granada.

REFERENCES

[1] Alfaro, J. Sistemas para la medición del rendimiento en la empresa, México, Limusa, 2008.

[2] Hammer, M. and Champy, J. Reengineering the corporation: A manifesto for business revolution. New York, Mc Graw Hill, 1993.

[3] Hammer, M. and Zisman, M. Design and implementation of office information systems. In Proceedings NYU Symposium on automated

office systems, New York University graduate school of business administration, pp. 13-24, 1979.

[4] Davenport, H. Reengineering a business process. Harvard business school note, 9-396-054, pp 1-16, USA, 1995.

[5] Kettinger, J., Teng, C., and Guha, S. Bussines Process change: a study of methodologies, techniques and tools, MIS Quarterly, 21 (1), pp. 55-80, USA, 1997.

[6] Larsen, G. Model-Driven development: Assets and reuse, USA, IBM Systems journal, 45(3), 2006.

[7] García, A., García, V., Sánchez, R., and García. J. Using software product lines to manage model families in model-driven engineering. In SAC 2007: Proceedings of the 2007 ACM Symposium on Applied Computing, track on Model Transformation, pp 1006_1011. ACM Press, Mar 2007.

[8] Clements, P., and Northrop, L. Software Product Lines: Practices and Patterns. Addison Wesley, Aug 2001.

[9] Rolland, C., and Prakash, N. On the Adequate Modeling of Business Process Families. Université Paris1 Panthéon Sorbonne. Francia, 2000.

[10] Basu, A. and Blanning, R. Model integration using metagraphs. Iinformation systems research, USA, 5(3), 1994.

[11] Carlsen, S. Conceptual modeling and composition of flexible workflow models. PhD Thesis, Information system group. Department of computer and information science,. Faculty of physic, informatics and mathematics, Norwegian university of science and technology, Trondheim, 1997.

[12] Davenport, H., and Short, J. The New Industrial Engineering: Information Technology and Business Process Redesign, Sloan Management Review 31(4), pp. 11-27, 1990.

[13] Winograd, T., and Flores, F. Understanding Computers and Cognition. A New foundation for Design. Norwood (NJ). 1986.

[14] Ould, M.A. Business Processes: Modeling and Analysis for Re-engineering and Improvement, Wiley, Chichester, 1995.

[15] Weigand, H., Verharen E., and Dignum, F. Interoperable transactions in business models - a structured approach. University of Technology, Eindhoven, pp. 1-17, 1996.

[16] Lind, M., and Goldkuhl, G. The Evolution of a Business Process Theory - the Case of a Multi-Grounded Theory, in Beekhuyzen J, von Hellens L, Guest K, Morley Understanding Computers and Cognition. A New foundation for Design. Norwood (NJ). 1986.

[17] Montilva. J. Desarrollo de Software Basado en Líneas de Productos de Software. IEEE Computer Society. Mérida. Venezuela, 2006.

[18] Perez, J., Ruiz, F., and Piattini, M. Model Driven Engineering applicator a Business Process Management. Informe Técnico UCLM-TSI-002. Departamento de Tecnologías y Sistemas de Información. Universidad de Castilla-La Mancha. España, 2007.

[19] Bonillo, P. Metodología para la Gerencia de los Procesos del Negocio Sustentada en el Uso de Patrones. Journal of Information Systems and Technology Management. Vol. 3, No. 2, 2006, pp. 143-162, 2006.

[20] La Rosa, M. Linking Domain Models and Process Models for Reference Model Configuration. Lecture Notes in Computer Science. Vol. 4928, Springer, Berlin, pp. 417-430, 2008.

[21] Mohan, S., Choil, E., and Dugki, M. Domain Specific Modeling of Business Processes and Entity Mapping using Generic Modeling Environment (GME). International Conference on Convergence and Hybrid Information Technology. Vol. 1, pp. 533-538, 2008.

[22] Valdivia, M., and Santana, W. Valoración teórica sobre la gestión de procesos de negocios, los sistemas de información empresariales y los estándares para la modelación, UCI, Cuba. Available: http://semanatecnologica.fordes.co.cu/Evirtual/files/Marbys%20%20William_VALORACION%20TEORICA%20SOBRE%20LOS%20SISTEMAS%20DE%20INFORMACION%20EMPRESARIALES%20%20LOS%20ESTANDARES%20PARA%20LA%20MODELACION.doc. Consulted: 10-Dec-2010.

[23] Zorzan F., and Riesco, D. Automatización de procesos de desarrollo de software definidos con SPEM. Universidad de Rio Cuarto, Argentina. Available: http://www.ing.unp.edu.ar/wicc2007/trabajos/ISBD/070.pdf. Consulted: 10-Dec-2010.

[24] Cánovas C. Un caso de estudio para la adopción de un BPMS. Universidad de Murcia. España. Available: http://alarcos.inf-cr.uclm.es/pnis/articulos/pnis-07-canovas-bpms.pdf. Consulted: 10-Dec-2010.

[25] Delgado, A. Desarrollo de Software con enfoque en el Negocio, Universidad de la Republica, Uruguay. Available: http://alarcos.inf-cr.uclm.es/pnis/articulos/pnis-07-delgado-dsen.pdf. Consulted: 10-Dec-2010.

[26] Finkelstein, A., Kramer, J., and Nuseibeh, B. Software Process Modeling and Technology, Research Studies Press LTD. Londres. 1994.

[27] Conraid, R. Object –Oriented and Cooperative Process Modelling in EPOS.In PROMOTER Book. Ed. Bashar A. Nuseibeh, Imperial College, 1994.

[28] Engels, G. Object –Oriented Specification of Coordinated Collaboration. In Advanced IT tools: IFIP World Conference on IT Tools, Springer. Canberra (Australia). pp. 2-6. September, 1996.

[29] Junkermann, G. Merlin: Supporting Cooperation in Software Development through a Knowledge-based Environment. In Software Process Modelling Technology, A. Finkelstein, J. Kramer, y B. Nuseibeh, eds., ch. 5, pp. 103-130, Somerset, England: Research Studies Press, 1994.

[30] Montangero, C., and Ambriola. V. OIKOS: Constructing Process-Centered SDEs. In A. Finkelstein, J. Kramer, y B. Nuseibeh, editors, Software Process Modelling and Technology, pp 33 – 70. John Wiley & Sons Inc, 1994.

[31] Canals, G. ALF: A framework for building process-centred software engineering environments. In A. Finkelstein, J. Kramer, & B. Nuseibeh, editors, Software Process Modelling and Technology, pp 153 – 185. John Wiley & Sons Inc, 1994.

[32] Belkhatir N. ADELE-TEMPO: An environment to support process modeling and enaction. In A. Finkelstein, J. Kramer, and B. Nuseibeh, editors, Software Process Modelling and Technology, pp 187 – 222. John Wiley & Sons Inc, 1994.

[33] Bandinelli, S. SPADE: An Environment for Software Process Analysis, Design, and Enactment. In A. Finkelstein, J. Kramer, and B. Nuseibeh, editors, Software Process Modelling and Technology. John Wiley & Sons Inc, 1994.

[34] Arbaoui, S., and Oquendo, F. PEACE: goal-oriented logic-based formalism for process modelling. In A. Finkelstein, J. Kramer, and B. Nuseibeh, editors, Software Process Modelling and Technology. John Wiley & Sons Inc, 1994.

[35] Baldi, M. E3: object-oriented software process model design. In A. Finkelstein, J. Kramer, and B. Nuseibeh, editors, Software Process Modelling and Technology. John Wiley & Sons Inc, 1994.

[36] Bruynooghe, F. PADM: towards a total process modelling system. In A. Finkelstein, J. Kramer, and B. Nuseibeh, editors, Software Process Modelling and Technology. John Wiley & Sons Inc, 1994.

 ISBN: 978-1-61208-139-7

Balancing LTE Protocol Load on a Multi-Core Hardware Platform Using EFSM Migration

Anas Showk, Shadi Traboulsi, David Szczesny
Institute for Integrated Systems,
University of Bochum,
44801 Bochum, Germany
Email: [Anas.Showk, shadi.traboulsi, david.szczesny]@rub.de

Attila Bilgic
KROHNE Messtechnik GmbH,
Ludwig-Krohne-Str. 5,
47058 Duisburg, Germany
Email: A.Bilgic@krohne.com

***Abstract*—In the past decade the mobile communication data rate is increased dramatically. Therefore, mobile terminals must have enough processing capability by migrating to multi-core processors. To exploit the multi-core processing power, we focus on a critical component in the future mobile terminal which is the load balancer. Its main role is to partition and balance the load so as to achieve an optimal sharing of load between cores and to eventually reduce power consumption. In this paper, we show how a model driven layered protocol stack can be parallelized and run on a multi-core modem. For instance, we illustrate how the Extended Finite State Machines (EFSMs) concurrency is employed to achieve the protocol stack parallelism. Furthermore, we move the load balancer to the modem subsystem layer by using the EFSM migration between cores. The proposed migration scheme during run time replaces the classic thread migration scheme and reduces the thread context switching overhead which definitely improves the performance. In addition, we present a semi-dynamic load balancer implementation accompanied with customized data pipeline scheduler for future multi-core LTE smart phones.**

***Keywords*-LTE protocol stack, multi-core mobile terminal, parallel embedded software, load balancing, model driven development.**

I. INTRODUCTION

The Long Term Evolution (LTE) is the enhancement of Universal Mobile Telecommunications System (UMTS) and is optimizing its radio access architecture. The targets of LTE are to increase the data rates to 100 $\mathrm{Mbit/s}$ in the downlink and 50 $\mathrm{Mbit/s}$ in the uplink. Due to the exponential growth of data rate in the mobile communication systems during the past decade, more efforts have been invested in order to achieve the required performance that satisfies the increasing processing demand for computational intensive applications. The computational power provided by single processing units, at reasonable power consumption, seems to grow slower compared to the application needs. Therefore, research is focused on migrating to multi-core architectures which can provide a better balance between performance, power consumption, flexibility and scalability. For instance, multi-cores can provide the required computing performance while allowing lower clock rates to achieve power efficiency and offer a second dimension in resource allocation.

Developing next generation mobile communication protocol can gain by reusing established approaches and best practices like Model Driven Development (MDD). For instance, the Specification Description Language (SDL) is commonly used in the previous mobile communication protocols. The dynamic behavior in an SDL system is described in processes using Extended Finite State Machines (EFSM) [1].

Data pipeline scheduling is accomplished by apportioning protocol functionalities into a series independent steps, where the output of one step is the input to the next. However, each step can be executed on a different core in order to constitute separated steps in a pipeline. These parts might be different protocol stack layers or specific protocol functions depending on the design. Pipelining can be very powerful if the degree of parallelization is high. Even though, it may take some efforts in order to fill the pipeline and generate a constant throughput. For instance, pipeline steps should have the same execution time and they must be tuned such that one stage does not become a bottleneck.

A load balancer is a component that distributes the computational load between two or more entities such as cores, clustered computer systems, network links or other resources, in order to get optimal resource utilization. In the embedded domain, load balancing provides an efficient execution of multiple threads on processors with multiple cores for concurrent and parallel applications. Depending on the implementation of the decision making of the load balancer, it can be a static, dynamic, or combination of them [2].

In this paper, we present the parallelization of model driven LTE protocol stack. In addition, we illustrate the design and implementation of a semi-dynamic load balancer accompanied with data pipeline scheduling for future multi-core LTE modems. It is moved to the modem subsystem layer by migrating the EFSMs between cores depending on the required data rate. For example, at very low data rates all protocol EFSMs should run on single core. If the data

 ISBN: 978-1-61208-139-7

rate is increased some of the EFSMs are moved to a second core in order to provide enough processing power. While increasing the data rate further, the load balancer distributes the EFSMs on more cores as needed.

This paper is organized such that, Section II shows the previous work done in this field. The smart phone system architecture is presented in Section III including the software stack and the hardware platform. The data pipelining scheduler for LTE protocol stack is discussed in Section IV. In Section V, the load balancer design and implementation is presented followed by the conclusion in Section VI.

II. Related Work

Embedded multi-core scheduling is investigated by researcher from different perspectives. For instance, the time server technique was developed in order to cope with scheduling several applications on the same platform [3]. This solution is used extensively for single core embedded systems and was also extended for multi-processor systems [4], [5]. Some power-aware scheduling algorithms utilize Dynamic Voltage Scaling (DVS) so as to optimize the power consumption [6], [7]. Moreover, in [8], a scheduling solution was used to achieve fault-tolerance for embedded systems with soft and hard timing constraints.

In [9], an efficient, optimal pipelining algorithm for associating a task chain with a chain of processors was explained. The pipelining algorithm is based on new data structure, called the layered assignment graph. A flow graph scheduling algorithm which considers pipelining, retiming and hierarchical node decomposition is presented in [10]. Research on pipeline scheduling is at a significantly less mature state than on the classical scheduling problem [11].

Researchers investigated load balancing for embedded multi-core systems on the operating system's level using the thread migration technique. However, most of the embedded multi-core system load balancing techniques used the thread level migration. For instance, based on the virtualization concept, the most common approaches in scheduling techniques and load balancing for embedded mobile communication systems are discussed in [2]. However, to the best of our knowledge, researchers have not used the EFSM migration for load balancing. In contrast to the other researchers, we develop an innovative EFSM migration scheme in order to minimize the thread context switching frequency. In addition, we customize data pipeline scheduling and integrate it with our load balancer in the modem subsystem layer so as to have a parallel execution of LTE protocol stack on the multi-core modem of a mobile terminal.

III. System Architecture

A mobile terminal architecture depicted in Fig. 1 is divided into three main parts: the hardware platform, the operating system layer and the modem subsystem layer. ARM processors are widely used in mobile phones and specifically the ARM11 core is representative for a mobile terminal state-of-the-art hardware platform [12]. Therefore, we have chosen ARM RealView® base board including the ARM11 MPCore™ to be our hardware platform. More details on the base board can be found in [13], [14]. In addition the L4/Fiasco microkernel is selected as an operating system. Moreover, the modem subsystem layer is composed of the load balancer and LTE protocol stack as illustrated in Fig. 1. In the next subsections, more details are given about the operating system and the modem subsystem layers.

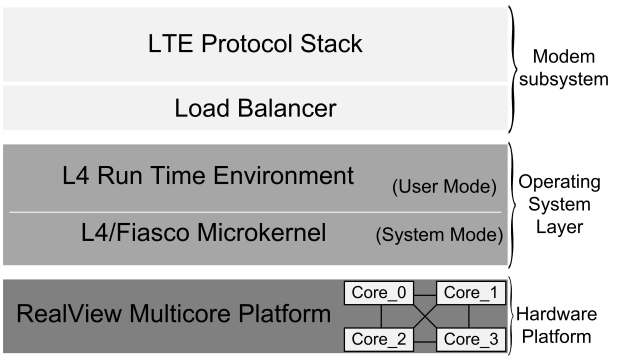


Figure 1. The system architecture of the future mobile terminal.

A. The Operating System's Layer

The L4/Fiasco Microkernel operating system is composed of two layers, the L4/Fiasco microkernel and the L4 runtime environment (L4Re) [15] as shown in Fig. 1. The selection of such a modern operating system is done because it represents a robust Real Time Operating Systems (RTOSs) as well as utilizes the concepts of microkernels [16]. The services of each layer of the deployed OS are illustrated in the following.

The L4/Fiasco microkernel is running in processor privileged mode and is responsible for controlling the underlying hardware. It provides a minimal set of mechanisms like tasks, threads, and Inter-Process Communication (IPC). Fiasco kernel services are implemented in terms of kernel objects. A task constitutes of an address space where one or more threads can execute. Multiple threads are scheduled by Fiasco's priority-based and preemptive scheduler. The scheduler is controlled by the timeslice length, priority and the maximum controlled priority. Each thread is scheduled for maximum time equal to the associated timeslice length. The kernel uses multiple-level round-robin queues such that there is a queue associated with each priority level. The combination of all the queues represents the kernels ready queue. Unlike the timeslice length and priority, the maximum controlled priority is not thread based but rather task based. It is specified for every task at the creation time and all threads in the task will have the same value. In addition, an IPC kernel object provides the basic communication

mechanism in L4-based systems and is used mainly for transmitting arbitrary data between threads.

On the other hand, the L4Re offers a basic set of abstractions and services, which are useful to implement user-level applications on top of the L4/Fiasco microkernel. It consists of a set of libraries mainly responsible for memory and IO resource management. The kernel does not migrate threads dynamically, but only supplies needed functionalities for its applications.

B. Load Balancer

The main job of the load balancer is to monitor the execution of applications and share the load between the available cores. In addition, it should decide where application components are initially started and when they need to be migrated in order to satisfy an optimum criterion like performance, power consumption, etc. Only the load balancer has a global view on the total load of the system and can thus distribute it on cores according to the previously specified requirements.

The load balancing is achieved within the modem subsystem layer by utilizing the EFSM migration scheme. It's design and development procedure can be divided into two main steps offline analysis and online processing. The first step (i.e., offline analysis) is accomplished at compile time to collect the information about LTE protocol stack and generate LTE scheduling tables which include the process-thread mapping tables and LTE configuration table prior to system execution. The process-thread mapping tables define the association of the EFSMs to threads depending on LTE states. LTE states are identified according the required data rate and how many cores should be involved in order to achieve this data rate. On the other hand, during system's run time the online processing includes all the actions needed to reconfigure the system when the LTE state changes. For example, when the load balancer detects an LTE state change, it will flush the pipeline by processing all the scheduled task. Then reconfigure the thread activation scenario and migrate the EFSMs between cores depending on the new LTE state in order to balance the load between cores. In Sect. V detailed explanations of the load balancer design and implementation are given.

C. LTE Protocol Stack

Developing software for LTE mobile terminal can greatly benefit from reusing prevailing approaches and best practices. For over a decade, most global installations have taken advantage of Model Driven Development (MDD) for communication products; oftentimes with tools using the Specification Description Language (SDL). Therefore, an SDL tool is selected to develop the access stratum part of LTE protocol stack in the mobile terminal side. Since the user plane is more computational intensive, the modeling targeted only the user plane part.

The access stratum part of LTE protocol stack includes layer 2 which is divided into three sublayers: Medium Access Control (MAC), Radio Link Control (RLC), and Packet Data Convergence Protocol (PDCP) [17]. Fig. 2 illustrates the LTE data flow from the mobile terminal perspective. In the uplink direction the mobile terminal generates packets and sends them through the air interface to an evolved base station (eNodeB). The building of packets payload and header as well as the downlink processing are modeled according to the 3GPP standard of LTE [18], [19], [20]. On the other hand, the mobile terminal receives Transport Blocks (TBs) in the downlink direction from the air interface and processes them through the MAC, RLC, PDCP and IP layers. The functionalities of the mentioned layers are modeled as described in more details in [21], [22].

The dynamic behavior in SDL systems is described in the SDL processes using EFSMs. Processes in SDL can be created at system start, or created and terminated dynamically at runtime. The concept of process instances that work autonomously and concurrently makes SDL a true real time language. The other advantage of the processes concurrency is making the parallelism easier to identify and exploit in contrast to pure C programming. It is clear, from Fig. 2, that the processing of packets in each layer should be done after the previous layer in a sequential manner. Therefore, the LTE protocol stack is modeled by dividing its functionalities into several EFSMs, which are communicating using the SDL asynchronous messages communication facility. For instance, the output from an EFSM is the input to the next EFSM and the latter's output represents the input to the one after, and so on, formulating a chain of EFSMs. Even though, the protocol stack processing is parallelized by exploiting EFSMs concurrency together with data pipeline scheduling.

IV. Pipeline Scheduling

In general, scheduling algorithms suitable for embedded systems are mapped to two major classes: static and dynamic. The scheduling involves three steps: assigning tasks to processors, ordering execution of these tasks on each processor, and determining when each task fires such that all data precedence constraints are met. To reduce the run-time computation, all the three steps are performed at compile time by a static scheduler. On the other hand, the scheduler which accomplishes these steps during run time is called a dynamic scheduler.

Generally, pipeline schedulers aim to efficiently divide a task into stages, allocate some cores to stages, and create schedules for each pipeline stage. As a consequence, the stage with the longest execution time in the pipeline determines the throughput of the multi-core system. In general, pipelining can considerably increase the throughput beyond what is obtainable by the classical (minimum-make-span) scheduling algorithms. However, this increase in throughput

 ISBN: 978-1-61208-139-7

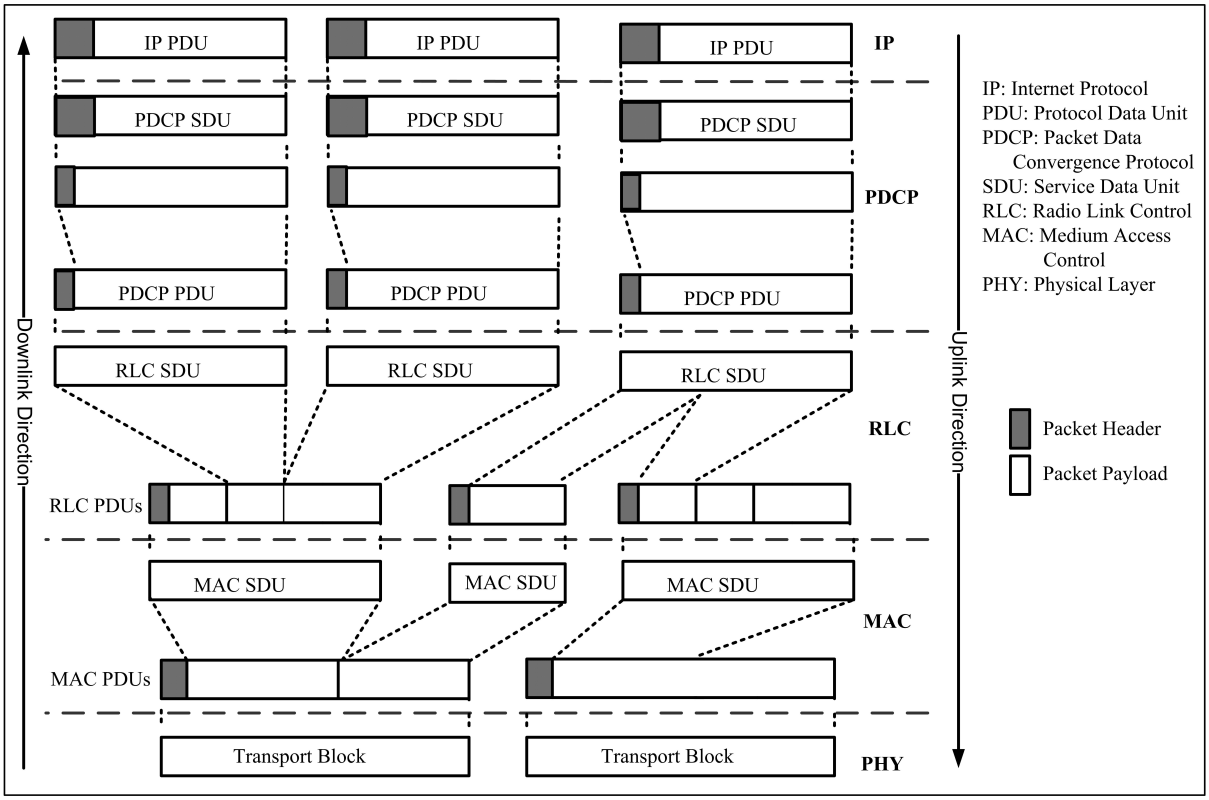


Figure 2. The LTE protocol data flow in the mobile terminal side.

may cost a significant increase of latency compared to the classical (minimum-make-span) schedulers [11].

The new direction that is relevant to embedded multi-core scheduling is the exploration of pipeline scheduling algorithms, which concentrate on throughput as the main performance metric. Hence, the presented data pipeline scheduler is adopted and customized for LTE protocol stack in order to improve the performance of the multi-core hardware platform. In addition, the scheduler is implemented on the SDL level by exploiting the message communication between EFSMs for several thread activation scenarios depending on number of needed cores.

As an example, a thread activation scenario of the LTE protocol running on four cores is depicted as a message sequence chart in Fig 3. All LTE EFSMs are distributed on four cores in such a way that processing a packet within the same core is illustrated as one task before the control is given to the other core for further processing of the same packet. For instance, the first packet (P_1) is processed by the first core (i.e., $Core_0$) while other cores are idle. After that, $Core_0$ sends the packet P_1 to $Core_1$ for activating the thread on $Core_1$ to do further processing and sends the message $Trigger$ to itself in order to start processing the second packet (P_2). consequently, both packets P_1 and P_2 are processed in parallel on cores $Core_1$ and $Core_0$, respectively. After processing P_1, $Core_1$ and $Core_2$ hand it over to the next core and send the $Trigger$ message to $Core_0$ to start working on a new packet. The last core $Core_3$ sends a $Trigger$ message to $Core_0$ when ever it finishes processing a packet in order to keep the pipeline full all the time. The only exception arises when there is an LTE state change and the pipeline should be cleaned thus it should not send this message. This technique of asynchronous SDL message communication is used for multi-core synchronization at a high abstraction level. One of the disadvantages of this scenario is the long latency of filling the pipeline (i.e., from start processing P_1 on the first core up to finish processing of the same packet on the last core). However, the throughput will increase if the load is distributed evenly between cores.

V. Load Balancer Design and Implementation

Power consumption and performance are very important factors for embedded multi-core systems dedicated to wireless communication protocol processing. Thus, in order to have an efficient multi-core mobile modem we concentrate on the load balancer which is one of the most critical components of the system. For example, a load balancer aims to share the load evenly between cores in order to ensure an optimal performance as well as to reduce power consumption. According to the load balancer design, it can

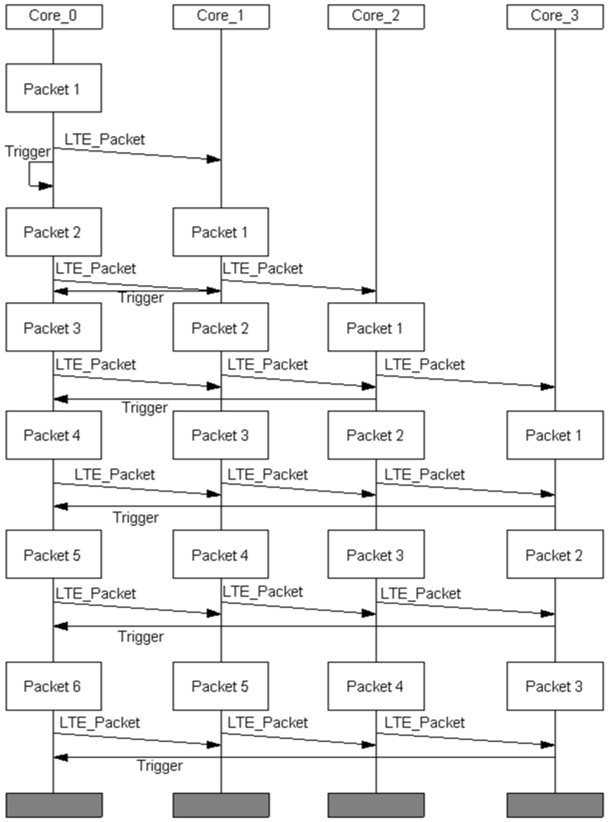


Figure 3. The message sequence chart of a thread activation scenario with four cores.

be a static, dynamic or combination of them. In this work, the load balancer development is divided into two main steps: offline analysis and online processing as depicted in Fig. 4. In the next sub-sections, more explanations of these steps are given.

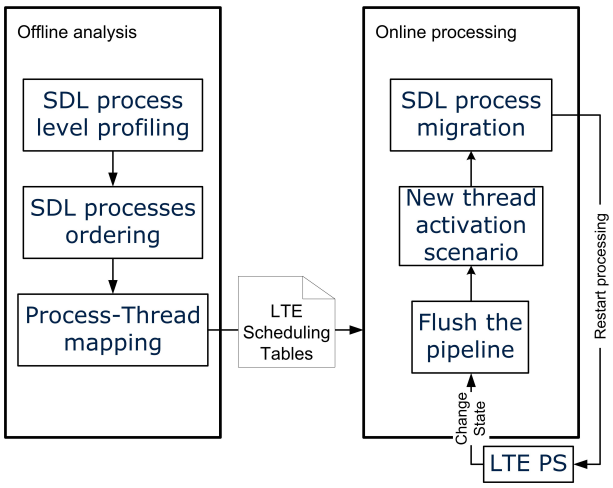


Figure 4. The offline analysis and online processing to balance LTE protocol load.

A. *Offline Analysis*

The LTE protocol stack design is accomplished using the hierarchical decomposition of SDL with system, block, sub-block and process as the main building blocks. Every protocol sublayer is realized by a sub-block (that is MAC, RLC,PDCP and IP). The latter is divided into sub-sub-blocks or processes depending on role or mode of the target layer. The behavior of the protocol is implemented in the SDL process level using concurrent EFSMs. Each EFSM is idle in the current state until it is triggered by an event to execute a transition and move to the next state. This event can be a message from another EFSM or even itself, an expiration of a timer or a change of internal variable.

A directed multi-graph is an ordered pair (A, E), where A is a set of actors (sometimes called nodes or vertices) and the set E is order pair of nodes called edges. Graphically, actors are represented by circles and edges are represented by arrows connecting the circles. Each edge is an ordered pair (a_1, a_2) where $a_1, a_2 \in A$. If $e = (a_1, a_2) \in E$, we say that e is directed from a_1 to a_2; a_1 is the source actor of e, and a_2 is the sink actor of e. In a directed multi-graph two or more edges can have the same source and sink actors and loops from the actor itself are allowed.

The data-flow graph is a directed multi-graph which is a conceptual notion for expressing the function of a system. The actors are the computations and the edges are First-In-First-Out (FIFO) queues. The latter direct the data (or token) as an output from one computation to be an input to another one. In terms of SDL and EFSMs, an actor is an EFSM (e.g., the active transition of a state machine form current state to the next state), an edge is an SDL queue and a token is an SDL message. Therefore, the collection of all EFSMs that constitutes the LTE protocol can be represented by data-flow graphs.

During offline analysis, the LTE protocol stack is profiled to measure the cost of every EFSM. Afterward, the EFSMs are ordered depending on which one will be executed first according to the data precedences to look like the data-flow graph example depicted in Fig 5. In the example, a data-flow graph is illustrated by a set of pairs (A, E), where set A is bounded by the big circle and E is not shown for sake of clarity. The set A includes all the nodes which are part of LTE protocol models where:

$$A = \bigcup_{i=1}^{N} a_i \ .$$

The execution time of each actor is represented by C_i where $i = [1, 2, 3, ..., N]$ is computed so as to calculate the cost of a group of actors or even the total cost of the system. For simplicity, we assume that the cost C_i includes the edges delay (i.e., the transition time between actors when running on the same core). Therefore, the cost of the chain of actors between actor a_x and actor a_y can be calculated by the delay

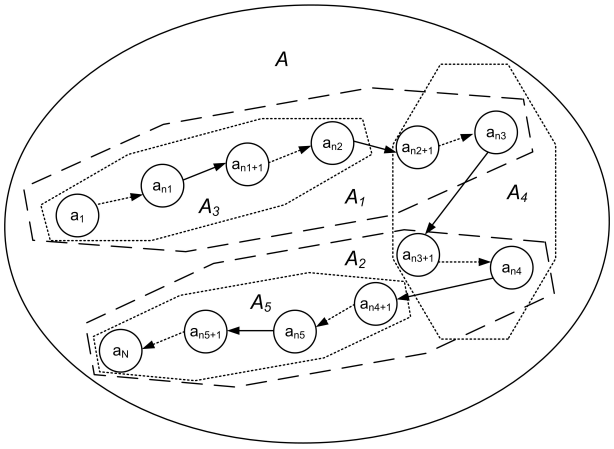


Figure 5. The data-flow graph example illustrates the mapping of actors into different sets in order to allocate them to one, two or three cores.

D_{xy} where:

$$D_{xy} = \sum_{i=x}^{y} C_i \; ; \; where \;\; 1 \leq x, y \leq N \; and \; x \leq y \; .$$

Since the ARM RealView® base board, which is used as a hardware platform, has four cores, the costs C_i are used to partition the set A in order to share the load evenly when one, two, three or four cores are needed. For example, subsets A_1 and A_2 in Fig. 5, which are bounded by octagons with dashed lines, are chosen in such a way that they should have almost equal costs (i.e., delay D_{1n_3} is equal to delay $D_{(n_3+1)N}$). Furthermore, the intersection between both sets should be equal zero and the union of them is equal to set A (or mathematically, $A_1 \cap A_2 = \phi$ and $A_1 \cup A_2 = A$). The same conditions are valid for sets A_3, A_4 and A_5 bounded by the octagons with dotted lines. The work can be extended to distribute the load on four cores by dividing the set A into another four sets by applying the same rules.

The system has only four threads which are created at the system startup and associated to each core. As a consequence, the process-thread mapping table for every LTE state in addition to the configuration tables are generated. The process-thread table illustrates which EFSM will run on which thread (core). The LTE state defines the number of cores needed according to the targeted data rate in addition to some other configuration parameters. For instance, a mobile terminal working at low data rate (like voice calls) leads to only one core that should be active. On the other hand, while streaming a very big video file, the system should utilize all the available four cores.

B. Online Processing

The SDL Suite™ tool is equipped with a deployment editor where the SDL system can be divided into separate threads. For example, at the design time the user can decide how many threads will be included and which EFSM will run in which thread. Therefore, at system start SDL system's run time kernel creates threads and associates different EFSMs with different threads according to the created deployment diagram. If an EFSM receives a message or other event occurs, the associated thread will be woken up to execute the transition. In this paper we develop a method to modify the EFSM parameters and move it from one thread to another during run time. In addition, this is also possible even if the two threads are running on different cores.

In this setup, at system start, four threads are created and each ARM11 core processor is allocated to each thread using L4/Fiasco's thread migration facility. When the LTE state changes during the LTE protocol execution, the SDL system will report the situation to the load balancer by calling a load balancer utility function to reconfigure the system. First of all, the load balancer flushes the pipeline by finishing all the scheduled tasks. After that, it reconfigures the thread activation scenario according to the new LTE state. Finally, it migrates some EFSMs depending on the process-thread mapping tables which are generated offline at compile time so as to balance the load between cores before restarting the LTE system processing of a new packet.

The EFSM migration is tested and the thread activities are monitored and printed in Fig. 6. The LTE threads with the names *lte_ps-main*, *lte_ps10001*, *lte_ps10002*, *lte_ps10003* and *lte_ps10004* are executed on cores $Core_0$, $Core_1$, $Core_2$, $Core_3$ and $Core_0$ respectively. The main thread of the system (i.e., *lte_ps-main*) includes the SDL system's run time kernel as well as the load balancer. At the beginning, all EFSMs are associated to thread *lte_ps10001* and run on $Core_1$. Then, for increasing LTE data rate some EFSMs are moved to thread *lte_ps10002* in order to distribute the load between $Core_1$ and $Core_2$ and the system output is verified. The number of used cores is increased to three and four while higher data rates are achieved.

In order to compare the performance of system when using EFSM migration with same system employing classical thread migration, assume the EFSMs are statically associated to different threads. Fig. 7 illustrate how the system load can be divided to threads which are distributed on multi-core. The X-axis represents the number of active cores in each case. On the other hand, the Y-axis shows the portion of the load on each core and the partitioning into threads. It is clear that, the minimum number of threads needed to distribute the load when activating one, two, three and four cores is six threads.

For example, at low data rate all threads should be migrated to run on only one core. As a consequence, the processing time of a TB, on uplink and downlink together, includes thread context switching costs due to the transition from one thread to the other. In contrast, no thread context switching occurs when using our EFSM migration scheme which definitely improves the performance. Moreover, this is still the case while running the LTE protocol on more than

 ISBN: 978-1-61208-139-7

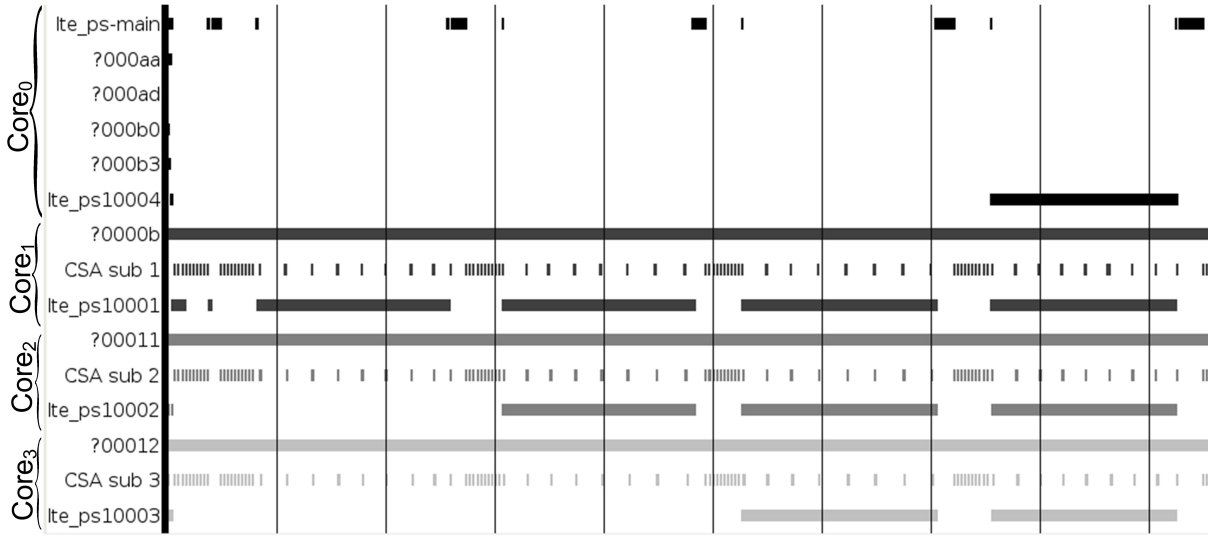


Figure 6. The monitoring of thread activities on ARM11 multi-core.

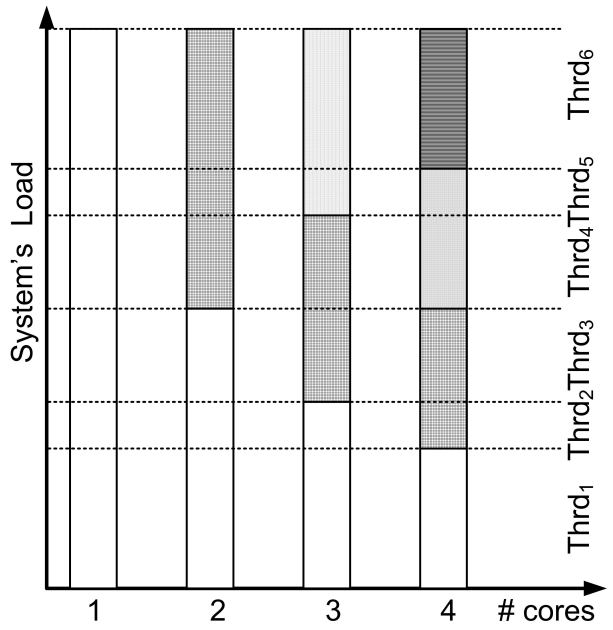


Figure 7. The system load partition into threads and distribution on one, two, three and four cores.

one core. More precisely, six, three and two thread context switches are avoided on one, two and three cores situation respectively. When four cores are active, four threads will execute on two cores (i.e., two threads for each core), thus introduces two thread context switches which can be avoided by employing EFSM migration. After measuring the thread context switching overhead, we found that it costs about 52 μs per switch. As a consequence, by utilizing EFSM migration we save around 312, 156, 104 μs when executing the LTE protocol on one, two, three cores respectively. In addition, in the four cores situation 104 μs from the execution time of processing one TB is reduced.

VI. Conclusion

A light version of the LTE protocol stack for the access stratum user plane is modeled using the SDL Suite™ tool. The SDL model is composed of several EFSMs which are associated with four threads to enable execution in a multi-core platform. The generated code is executed on ARM RealView® baseboard on top of an L4/Fiasco based RTOS. In this paper, we investigate load balancing and scheduling of LTE model driven protocol stack on a state-of-the-art multi-core mobile terminal. In addition, we show how the SDL EFSMs concurrency is exploited in order to achieve a parallel execution of the LTE protocol. As a result, we offer a parallel software architecture for LTE which is not existing today to the best of our knowledge. A new technique for load balancing in the modem subsystem level using EFSM migration is presented and successfully implemented. Moreover, the load balancing is accompanied with an adopted and customized data pipeline scheduling in order to make it suitable for the LTE protocol stack. In addition, we employ the thread activation scenario as a high level synchronization technique for multi-core mobile modem platform by utilizing the asynchronous message communication facility of SDL Suite™ tool. Last but not least, we prove that our innovative EFSM migration technique avoids the thread context switching and therefore, improves the performance in contrast to the thread migration counterpart. For future work, we are planning to continue enhancing the LTE mobile terminal performance by optimizing the IPC cost and decreasing its frequency. Even more, we will consider power consumption

 ISBN: 978-1-61208-139-7

as a parameter for scheduling in order to increase the battery life time of the LTE mobile terminal.

REFERENCES

[1] IBM Rational®, "SDL Suite™ User Manual," SDL Suite™ v6.1.

[2] D. Tudor, G. Macariu, C. Jebelean, and V. Cretu, "Towards a Load Balancer Architecture for Multi-Core Mobile Communication Systems," in *Proceedings of the 5th International Symposium on Applied Computational Intelligence and Informatics*, May 2009, pp. 391–396.

[3] S. Baruah and G. Lipari, "A Multiprocessor Implementation of the Total Bandwidth Server," in *Proceedings of the 18th International Parallel and Distributed Processing Symposium (IPDPS04)*, June 2004, p. 40a.

[4] B. Brandenburg and J. Anderson, "Integrating Hard/Soft Real-Time Tasks and Best-effort Jobs on Multiprocessors," in *Proceedings of the 19th Euromicro Conference on Real-Time Systems (ECRTS07)*, July 2007, pp. 61–70.

[5] S. Baruah, J. Goossens, and G. Lipari, "Implementing Constant-Bandwidth Servers upon Multiprocessor Platforms," in *Proceedings of the Eighth IEEE Real-Time and Embedded Technology and Applications Symposium (RTAS02)*, September 2002, pp. 154–163.

[6] Z. Shao, M. Wang, Y. Chen, C. Xue, M. Qiu, L. Yang, and E. Sha, "Real-Time Dynamic Voltage Loop Scheduling for Multi-Core Embedded Systems," *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 54, no. 5, pp. 445–449, May 2007.

[7] Y. Chen, Z. Shao, Q. Zhuge, C. Xue, B. Xiao, and E. Sha, "Minimizing Energy via Loop Scheduling and DVS for Multi-Core Embedded Systems," in *Proceedings of the 11th international Conference on Parallel and Distributed Systems - Workshops (ICPADS'05)*, vol. 2, July 2005, pp. 2–6.

[8] V. Izosimov, P. Pop, P. Eles, and Z. Peng, "Scheduling of Fault- Tolerant Embedded Systems with Soft and Hard Timing Constraints," in *Proceedings of the Conference on Design, Automation and Test in Europe (DATE '08)*, March 2008, pp. 915–920.

[9] S. H. Bokhari, "Partitioning Problems in Parallel, Pipelined, and Distributed Computing," *IEEE Transactions on Computers*, vol. 37, no. 1, p. 4857, January 1988.

[10] P. Hoang and J. M. Rabaey, "Scheduling of DSP Programs onto Multiprocessors for Maximum Throughput," *IEEE Transactions on Signal Processing*, vol. 41, no. 6, pp. 2225–2235, June 1993.

[11] S. Sriram and S. S. Bhattacharyya, *Embedded Multiprocessors: Scheduling and Synchronization*, 2nd ed. New York, NY, USA: Taylor & Francis Group, CRC., 2009.

[12] C. van Berkel, "Multi-core for Mobile Phones," in *Proceedings of the Conference on Design, Automation and Test in Europe*, April 2009, pp. 1260–1265.

[13] ARM, "RealView® Platform Baseboard for ARM11 MPCore™ User Guide," http://infocenter.arm.com/help/index.jsp?topic=/com.arm.doc.dui0351c, March 2009.

[14] ARM, "ARM11 MPCore™ Processor Technical Reference Manual," http://infocenter.arm.com/help/index.jsp?topic=/com.arm.doc.ddi0360f, October 2008.

[15] TU Dresden, "The Fiasco Microkernel," http://os.inf.tu-dresden.de/fiasco.

[16] S. Traboulsi, F. Bruns, A. Showk, D. Szczesny, S. Hessel, E. Gonzalez, and A. Bilgic, "SDL/Virtual Prototype Co-design for Rapid Architectural Exploration of a Mobile Phone Platform," in *Design for Motes and Mobiles*. Springer-Verlag, September 2009, pp. 239–255.

[17] 3rd Generation Partnership Project (3GPP), "The LTE Protocol Specification, 3GPP Rel8," http://www.3gpp.org/Release-8.

[18] 3GPP TS 36.323, "Evolved Universal Terrestrial Radio Access (E UTRA); Medium Access Control (MAC) Protocol Specification," March 2009.

[19] 3GPP TS 36.322, "Evolved Universal Terrestrial Radio Access (E UTRA); Radio Link Control (RLC) Protocol Specification," March 2009.

[20] 3GPP TS 36.323, "Evolved Universal Terrestrial Radio Access (E UTRA); Packet Data Convergence Protocol (PDCP) Specification," March 2009.

[21] A. Showk, D. Szczesny, S. Traboulsi, I. Badr, E. Gonzalez, and A. Bilgic, "Modeling LTE Protocol for Mobile Terminals using a Formal Description Technique," in *Design for Motes and Mobiles*. Springer-Verlag, September 2009, pp. 222–238.

[22] A. Showk, F. Bruns, S. Hessel, A. Bilgic, and I. Badr, "Optimal resource management for a model driven lte protocol stack on a multicore platform," in *Proceedings of the 8th ACM international symposium on Mobility management and wireless access (MobiWac'10)*. ACM, October 2010, pp. 91–98.

 ISBN: 978-1-61208-139-7

Agent-based Versus Macroscopic Modeling of Competition and Business Processes in Economics

Valentas Daniunas

Institute of Lithuanian Scientific Society
Vilnius, Lithuania
mokslasplius@itpa.lt

Vygintas Gontis, Aleksejus Kononovicius
Institute of Theoretical Physics and Astronomy
Vilnius University
Vilnius, Lithuania
vygintas@gontis.eu, aleksejus.kononovicius@gmail.com

***Abstract*—Simulation serves as a third way of doing science, in contrast to both induction and deduction. The web based modeling may considerably facilitate the execution of simulations by other people. We present examples of agent-based and stochastic models of competition and business processes in economics. We start from as simple as possible models, which have microscopic, agent-based, versions and macroscopic treatment in behavior. Microscopic and macroscopic versions of herding model proposed by Kirman and Bass diffusion of new products are considered in this contribution as two basic ideas.**

***Index Terms*—agent-based modeling; stochastic modeling; competition models; business models.**

I. Introduction

Statistically reasonable models of social systems, first of all stochastic and agent based, are of great interest for wide community of interdisciplinary researchers dealing with diversity of complex systems [1]. Computer modeling serves as a technique in the for finding relation between micro level interactions of agents and macro dynamics of the whole system. Nevertheless, some general theories or methods that are well developed in the natural and physical sciences can be helpful in the development of consistent micro and macro modeling of complex systems [1]. Our own modeling of financial markets by the nonlinear stochastic differential equations is based on the empirical analysis of financial data and power law statistics of proposed equations [2]. Reasoning of proposed equations by the microscopic interactions of traders (agents) looks as a tough task for such complex system. Apparently the development of macroscopic descriptions for the well established agent based models would be more consistent approach in the analysis of micro and macro correspondence. For such analysis one should select simple enough agent based models with established or expected corresponding macroscopic description. In this contribution we discuss few examples of agent based modeling in business and finance with corresponding macroscopic description of selected systems.

Kirman's ant colony model [3] is agent-based model, which explains the importance of herding and individuality inside the ant colonies. As human crowd behavior is ideologically very similar, this model can be applied to and actually was built as framework for financial market modeling [3], [4], [5]. On our website, [6], we have presented interactive realizations of the original Kirman's agent-based model (see [7]) and of it's stochastic treatment by Alfarano et al. [4] (see [8]). Further we follow the works by Alfarano et al. [4], [5] and introduce our own model modifications in order to obtain more sufficient agent-based models of financial markets, which would have an alternative macroscopic description in the terms of Stochastic Calculus.

Diffusion of new products is a key problem in marketing research. Bass Diffusion model is a prominent model in diffusion theory introducing a differential equation for the number of adopters of the new products [9]. Such basic macroscopic description in marketing research can be studied using microscopic agent-based modeling as well [10]. It is a great opportunity to explore the correspondence between the two micro and macro descriptions looking for the conditions under which both approaches converge. Bass Diffusion model is of great interest for us as representing very practical and widely accepted area of business modeling. Web based interactive models, presented on the site [11] serve as an additional research instrument available for very wide community.

Our web site [6] was setup using WordPress webloging software. WordPress is user-friendly, powerful and extensible web publishing platform, which can be adapted to scientist's needs. There is a wide choice of plugins, which enable writing of equations (mostly using LaTeX). Though bibliography management is not as well covered.

Interactive models themselves are independent from WordPress framework. They were implemented using Java programing language [12], which is better suited for stochastic modeling, and AnyLogic multi-paradigm simulation software [13], which provides convenient tools to implement agent based models. Either way by compiling appropriate files one obtains Java applets, which can be included in to the articles written using WordPress. This way articles become interactive - user can both theoretically familiarize himself with the model and test if the claims made in the article describing model were true. This happens in the same browser window, thus, transition between theory and modeling appears to be seemless. As models are implemented as Java applets all computation occurs on client machine, user must have Java Runtime Environment installed (it is available free of charge from Oracle Corp.), and server load stays minimal.

 ISBN: 978-1-61208-139-7

In Sections II and III, we present web-based micro and macro modeling of selected social systems in more details. Conclusions and future work are given in the Section IV.

II. KIRMAN'S MODEL FOR FINANCIAL MARKETS AND ITS STOCHASTIC TREATMENT

There is an interesting phenomenon concerning behavior of ant colony. It appears that if there are two identical food sources nearby, ants exploit only one of them at a given time. The interesting thing is that food source which is in use is not certain at any point of time. As at some times switches between food sources occur, though the quality of food sources remains the same. One could imagine that those different food sources are different trading strategies or simply actions available to traders (i.e., buy and sell). Thus, one could argue that speculation bubbles and crashes in the financial markets are of similar nature as explotation of food in ant colonies - as quality of stock and quality of food in the ideal case can be assumed to be constant. Thus, model [3] was created using ideas obtained from the animal world in order to mimic traders' behavior in the financial markets.

And actually Kirman, as an economist, developed this model as rather general framework in context of economic modeling (see [3] and his later bibliography). Though recently his framework was also used by other authors who are concerned with the financial market modeling (see [4], [5]). Basing ourselves on the main ideas of these authors and our previous results in stochastic modeling (see [2]) we introduce specific modifications of Kirman's model providing a class of nonlinear stochastic differential equations [14] applicable for the financial variables.

Original Kirman's one step transition probabilities [3],

$$p(X \to X+1) = (N-X)\,(\sigma_1 + hX)\,, \tag{1}$$
$$p(X \to X-1) = X\,(\sigma_2 + h[N-X])\,, \tag{2}$$

can be rewritten for continuous $x = X/N$ as

$$\pi^+(x) = (1-x)\left(\frac{\sigma_1}{N} + hx\right), \tag{3}$$
$$\pi^-(x) = x\left(\frac{\sigma_2}{N} + h[1-x]\right), \tag{4}$$

where X is a number of agents exploiting chosen trading strategy, N is a total number of agents in the system. Here the large number of agents N is assumed to ensure the continuity of variable x, expressing fraction of selected agents, X, from whole population. Note that the transition probabilities depend on σ_1, σ_2 parameters, which govern individual switches between trading strategies (thus, appropriate terms depend only on the size of the opposing group), and h parameter, which governs recruitment (thus, appropriate terms depend on both sizes - size of the current and opposing groups). Evidently these probabilities are interrelated

$$p(X \to X \pm 1) = N^2 \pi^{\pm}(x). \tag{5}$$

One can write Master equation for the probability density function of continuous variable x by using one step operators $\mathbf{E}$ and $\mathbf{E^{-1}}$ introduced in [15]. Thus, Master equation can be compactly expressed as

$$\partial_t \omega(x,t) = N^2 \left\{(\mathbf{E}-1)[\pi^-(x)\omega(x,t)] + (\mathbf{E^{-1}}-1)[\pi^+(x)\omega(x,t)]\right\}. \tag{6}$$

With the Taylor expansion of operators $\mathbf{E}$ and $\mathbf{E^{-1}}$ (up to the second term) we arrive at the approximation of the Master equation

$$\partial_t \omega(x,t) = -N\partial_x[\{\pi^+(x) - \pi^-(x)\}\omega(x,t)] + \frac{1}{2}\partial_x^2[\{\pi^+(x) + \pi^-(x)\}\omega(x,t)]. \tag{7}$$

By introducing custom functions

$$A(x) = N\{\pi^+(x) - \pi^-(x)\} = \sigma_1(1-x) - \sigma_2 x, \tag{8}$$
$$D(x) = \pi^+(x) + \pi^-(x) = 2hx(1-x) + \frac{\sigma_1}{N}(1-x) + \frac{\sigma_2}{N}x, \tag{9}$$

one can make sure that the above approximation of the Master equation is actually Fokker-Planck equation (first derived in a different way in [4])

$$\partial_t \omega(x,t) = -\partial_x[A(x)\omega(x,t)] + \frac{1}{2}\partial_x^2[D(x)\omega(x,t)]. \tag{10}$$

It is known, [16], that the above Fokker-Planck equation can be rewritten as Langevin equation (this equation was also presented in [4])

$$\mathrm{d}x = A(x)\mathrm{d}t + \sqrt{D(x)}\mathrm{d}W = [\sigma_1(1-x) - \sigma_2 x]\mathrm{d}t + \sqrt{2hx(1-x)}\mathrm{d}W, \tag{11}$$

here W stands for Wiener process.

By assuming that market is instantaneously cleared Alfarano et al. [4] have defined return as

$$r = r_0 \frac{x(t)}{1-x(t)}\eta(t), \tag{12}$$

where $x(t)$ is assumed to be fraction of chartist traders in the market, while other traders in the market, $1 - x(t)$, are assumed to follow fundamentalist trading strategy, $\eta(t)$ is the change of chartist mood defined in the same time window as return, in the most simple case it could be assumed to be a random variable [4], and r_0 scaling term. Using Ito formula for variable substitution [16] we obtain nonlinear SDE for the middle term, $y(t) = \frac{x(t)}{1-x(t)}$,

$$\mathrm{d}y = (\sigma_1 - y[\sigma_2 - 2h])(1+y)\mathrm{d}t + \sqrt{2hy}(1+y)\mathrm{d}W. \tag{13}$$

Agreement between agent-based model and stochastic model for y is demonstrated in Fig. 1.

Note that the above derivation, and thus, the final equations, does not change even if σ_1, σ_2 or h are functions of x or y. Thus, one can further study the possibilities of the model by checking different scenarios of σ_1, σ_2 or h being functions of x or y. Nevertheless, the most natural way is to introduce a custom function $\tau(y)$ as inter-event time. In such case the switching probabilities above can be interpreted as probability fluxes per time unit. And thus, one can divide the

 ISBN: 978-1-61208-139-7

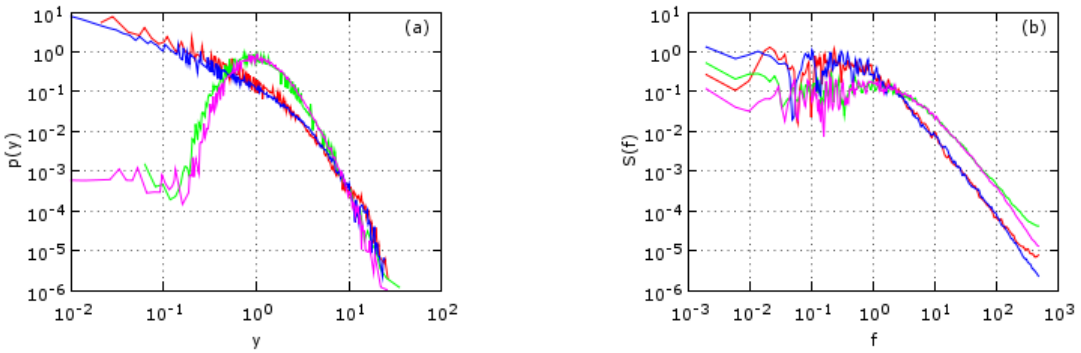


Fig. 1. Agreement between statistical properties of y, (a) probability density function and (b) power spectral density, obtained from stochastic (blue and magenta curves) and agent-based (red and green curves) models. Two qualitatively different model phases are shown: red and blue curves correspond to herding dominant model phase ($\sigma_1 = \sigma_2 = 0.2$, $h = 5$), while green and magenta curves correspond to individual behavior dominant model phase ($\sigma_1 = \sigma_2 = 16$, $h = 5$).

aforementioned constants by $\tau(y)$. We have chosen the case of

$$\mathrm{d}y = \left(\sigma_1 + y\frac{2h-\sigma_2}{\tau(y)}\right)(1+y)\mathrm{d}t + \sqrt{\frac{2hy}{\tau(y)}}(1+y)\mathrm{d}W, \quad (14)$$

as nonlinear SDE driving statistics of return in financial market. We did not devide σ_1 by $\tau(y)$ on purpose as one could argue that individual behavior of fundamentalist trader does not depend on the observed returns.

In Fig. 2, we have shown statistical properties of the stochastic model (14) with different $\tau(y)$ scenarios in use.

Note that while obtained stochastic model appears to be too crude to reproduce statistical properties of financial markets in such details as our stochastic model [2] based on empirical analyzes, it contains long range power-law statistics of return. Obtained equations are very similar to some general stochastic models of the financial markets [14], [17] and thus, in future development might be able to serve as microscopic justification for them and maybe for our more sophisticated modeling [2].

III. Two treatments of the Bass Diffusion model

The Bass model introduces a differential equation for the diffusion rate of new products or technologies [9]

$$\frac{\mathrm{d}N(t)}{\mathrm{d}t} = [M - N(t)][p + \frac{q}{M}N(t)], \quad (15)$$

$$N(0) = 0. \quad (16)$$

where $N(t)$ denotes the number of product users at time t; M is a market potential (number of potential users), p is the coefficient of innovation, the likelihood of an individual to adopt the product due to influence by the commercials or similar external sources, q is the coefficient of imitation, a measure of likelihood that an individual will adopt the product due to influence by other people who already adopted the product. This nonlinear differential equation serves as a macroscopic description of new product adoption by customers widely used in business planning [10].

Another approach to the same problem is related with agent based modeling of product adoption by individual users, or agents. The diffusion process is simulated by computers, where individual decisions of adoption occur with specific adoption probability affected by the other individuals in the neighborhood. It is easy to show that Bass diffusion process is a specific case of Kirman's herding model [3]. Indeed, lets define $x(t)$ as $x(t) = N(t)/M$ and in analogy with Kirman's model probability that new user will adopt the product as

$$\pi^+(x) = (1-x)\left(\frac{\sigma_1}{M} + \frac{h}{M}x\right). \quad (17)$$

In the case of Bass diffusion process is of one direction and $\pi^-(x) = 0$. Note that we assume an extensive herding in equation (17) as only in this case the stochastic term in corresponding Langevin equation vanishes with $M \to \infty$. Then the functions defining the macroscopic system description are as follows

$$A(x) = M\pi^+(x) = (1-x)(\sigma_1 + hx), \quad (18)$$

$$D(x) = \pi^+(x) = \frac{(1-x)}{M}(\sigma_1 + hx). \quad (19)$$

In the limit $M \to \infty$ one gets Bass diffusion equation (15) with $p = \sigma_1$ and $q = h$ instead of Langevin equation. This proofs that Bass diffusion is a special case of Kirman's herding model. Though this simple relation looks straightforward, we derive it and confirm by numerical simulations in fairly original way.

In Figure 3 we demonstrate the correspondence between macroscopic and microscopic Bass diffusion description. Agent based and continuous descriptions of product adoption ΔN per time interval Δt converge when number of potential users M or time interval Δt increase.

One of the goals of developing these models on the web site [11] was to provide theoretical background of Bass diffusion model and practical steps how such computer simulations can be created even with limited IT knowledge and then used for practical purposes. Thus, we target small and medium enter-

 ISBN: 978-1-61208-139-7

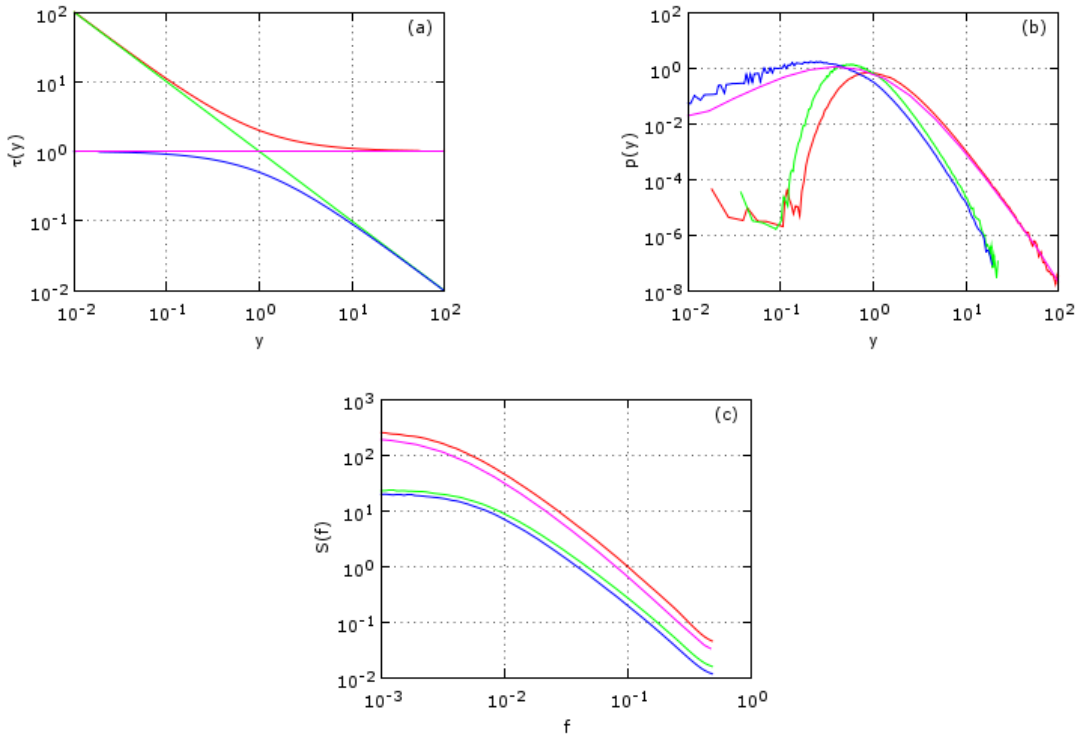


Fig. 2. Statistical properties, probability density functions (sub-figure (b)) and power spectral density (sub-figure (c)), obtained while solving (14) with different $\tau(y)$ scenarios (sub-figure (a)) being in use. Model parameters were set as follows: $\sigma_1 = \sigma_2 = 0.009$, $h_0 = 0.003$.

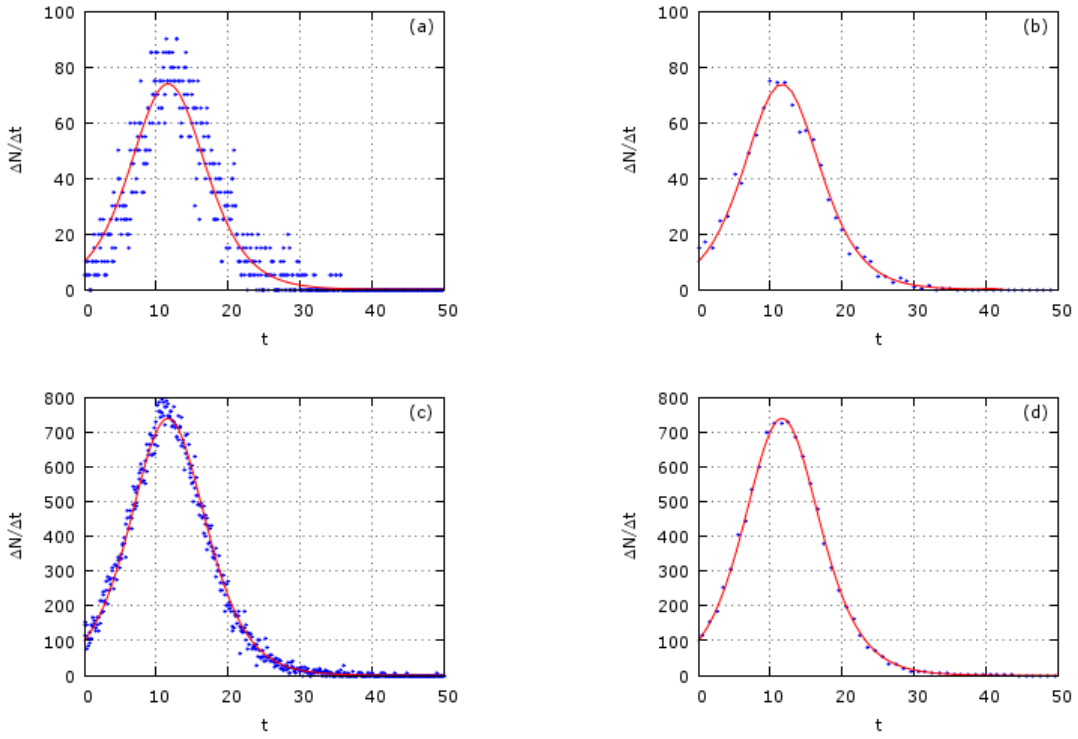


Fig. 3. Comparison of Bass diffusion, $\Delta N/\Delta t$ versus t, in macroscopic description (red line) and agent based model (blue points) shows convergence when time interval Δt or number of potential users M are increasing. (a) $M = 1000$, $\Delta t = 0.1$; (b) $M = 1000$, $\Delta t = 1$; (c) $M = 10000$, $\Delta t = 0.1$; (d) $M = 10000$, $\Delta t = 1$. Other parameters are as follows $p = 0.01$, $q = 0.275$.

 ISBN: 978-1-61208-139-7

prises to encourage them to use modern computer simulation tools for business planning and other purposes.

Computer models published at the [11] provide a relatively easy starting point to get acquainted with computer simulation and enables portal visitors to use these computer models interactively, running them directly in a window of web browser, changing parameters and observing results. This significantly increases accessibility and dissemination of these simulations.

IV. Conclusions and future work

Reasoning of stochastic models of complex systems by the microscopic interactions of agents is still a challenge for researchers. Only very general models such as Kirman's herding model in ant colony or Bass diffusion model for new product adoption have well established agent based versions and can be described by stochastic or ordinary differential equations. There are many different attempts of microscopic modeling in more sophisticated systems, such as financial markets or other social systems, intended to reproduce the same empirically defined properties. The ambiguity of microscopic description in complex systems is an objective obstacle for quantitative modeling. Simple enough agent based models with established or expected corresponding macroscopic description are indispensable in modeling of more sophisticated systems. In this contribution we discussed various extensions and applications of Kirman's herding model.

First of all, we modify Kirman's model introducing interevent time $\tau(y)$ or trading activity $1/\tau(y)$ as functions of driving return y. This produces the feedback from macroscopic variables on the rate of microscopic processes and strong nonlinearity in stochastic differential equations responsible for the long range power-law statics of financial variables. We do expect further development of this approach introducing the mood of chartists as independent agent based process.

One more outcome of Kirman's herding behavior of agents is one direction process - Bass diffusion. This simple example of correspondence between very well established microscopic and macroscopic modeling becomes valuable for further description of diffusion in social systems. Models presented on the interactive web site [6] have to facilitate further extensive use of computer modeling in economics, business and education.

Acknowledgment

Work presented in this paper is supported by EU SF Project "Science for Business and Society", project number: VP2-1.4-ŪM-03-K-01-019.

We also express deep gratitude to Lithuanian Business Support Agency.

References

[1] M. Waldrop, *Complexity: The emerging order at the edge of order and chaos*. New York: Simon & Schuster, 1992.

[2] V. Gontis, J. Ruseckas, and A. Kononovicius, "A non-linear double stochastic model of return in financial markets," in *Stochastic Control*, C. Myers, Ed. Scyio, 2010, pp. 559–580.

[3] A. Kirman, "Ants, rationality, and recruitment," *Quarterly Journal of Economics*, vol. 108, pp. 137–156, 1993.

[4] S. Alfarano, T. Lux, and F. Wagner, "Estimation of agent-based models: The case of an asymmetric herding model," *Computational Economics*, vol. 26, no. 1, pp. 19–49, 2005.

[5] ——, "Time variation of higher moments in a financial market with heterogeneous agents: An analytical approach," *Journal of Economic Dynamics and Control*, vol. 32, no. 1, pp. 101–136, 2008.

[6] V. Gontis, V. Daniūnas, and A. Kononovičius, "Physics of risk," Web site: http://mokslasplius.lt/rizikos-fizika/en.

[7] A. Kononovičius and V. Gontis, "Kirmans ant colony model," Web page: http://mokslasplius.lt/rizikos-fizika/en/agent-based-models/kirman-ants.

[8] ——, "Stochastic ant colony model," Web page: http://mokslasplius.lt/rizikos-fizika/en/stochastic-models/stochastic-ant-colony-model.

[9] F. M. Bass, "A new product growth model for consumer durables," *Management Science*, vol. 15, pp. 215–227, 1969.

[10] V. Mahajan, E. Muller, and F. M. Bass, "New-product diffusion models," in *Handbooks in Operations Research and Management Science*, G. L. L. J. Eliashberg, Ed. Amsterdam: North Holland, 1993, vol. 5: Marketing, pp. 349–408.

[11] V. Daniūnas, "Verslo modeliai," Section on web site: http://mokslasplius.lt/rizikos-fizika/business.

[12] Oracle Corporation, "java.com: Java + you," Web site: http://www.java.com/en/.

[13] X. Technologies, "Anylogic multi-paradigm simulation software," Web site: http://www.xjtek.com/anylogic, 2011.

[14] J. Ruseckas and B. Kaulakys, "1/f noise from nonlinear stochastic differential equations," *Physical Review E*, vol. 81, p. 031105, 2010.

[15] N. G. van Kampen, *Stochastic process in Physics and Chemistry*. Amsterdam: North Holland, 1992.

[16] C. W. Gardiner, *Handbook of Stochastic Methods*. Berlin: Springer, 1997.

[17] S. Reimann, V. Gontis, and M. Alaburda, "Interplay between positive feedbacks in the generalized CEV process," *Physica A*, vol. 390, no. 8, pp. 1393–1401, 2011.

 ISBN: 978-1-61208-139-7

Data Fusion Integrated Mobile Platform for Intelligent Travel Information Management

Marius Minea

University "Politehnica" of Bucharest
Transports Faculty, T.E.T. Dept.
Bucharest, Romania
Marius.Minea@upb.ro

Martin Böhm

Austria Tech – Gesellschaft des
Bundes für technologiepolitische
Maßnahmen GmbH
ITS Deployment
Vienna, Austria
Martin.Boehm@austriatech.org

Sorin Dumitrescu

Electronic Solutions Ltd.
ITS and Communications consultants
Bucharest, Romania
Sorin.Dumitrescu@elsol.ro

***Abstract* — This paper presents the concept and initial results of an European research project called *In Time,* regarding the integration via unified interfacing and the intelligent management of Multimodal Real time Travel and Traffic Information (MRTTI). The cooperation of different technologies, telematics applications and data fusion is used to produce more efficient travel management, with the final goal to reduce the energy consumption and emissions in urban areas, across the different modes of transports, by changing the mobility behaviour (modal shift) of the single traveller.**

Keywords – mobile real-time traffic and travel information; services; reliability; operability.

I. INTRODUCTION

Efficient transport and travel are two components without which a modern society cannot be conceived. Recently, the increased growth of traffic demand in major urban areas lead to increased usage of intelligent transport systems and technologies, especially those related to traffic management and information systems [2] (pp. 1-2). Different applications were developed lately for modern mobile devices, running on different operating systems (e.g., Windows Mobile, Windows Phone, Android, Symbian, Apple iOS etc.). These help people in finding their location, points of interest, routes or nearby facilities. But these applications only relate to the producer's software, quality of maps and/or static information. The idea of the *In Time* project [1] (*Description of Work*, pp. 9-38) was to develop a standardised interface, as a start point in conceiving a unified platform for the management of global mobile, real-time travel and traffic information for urban areas. The common interface will help a traveller across Europe to easier find his way and needed travel information, based on real-time information, no matter what the traffic information service provider would be. The usage of real-time information adds a lot of value to the efficiency of the information system, due to its possibility to re-configure routes, helping the user to avoid congested areas. One of the benefices is that it also helps users to change their modal transport behaviour, improving the usage of the public transport versus private cars. These actions finally produce less emissions and energy consumption in traffic, helping in obtaining greener cities. The *In Time* solution, with the commonly agreed standardized interface is now under testing in six European pilot sites to ensure the easy access of real-time multimodal traffic data for external Traffic Information Service Providers (TISPs) and to check several impacts: users' impact, traffic impact or the environmental one. This model already ensures the easy access to all urban traffic-related data within a larger region, resulting in the distribution to the end-users via several consistent information channels and in parallel enhancing the user acceptance.

The paper presents the concept and architecture of the system, the benefits, initial results of the first tests and possible future development of the applications.

II. THE NOVELTY OF THE *IN TIME* CONCEPT

The mobility is one of the attributes of a modern person. This is why the project *In Time* is focused on increasing the mobility of people in urban areas. *In Time* (acronym for Intelligent and Efficient Travel Management) is an EU funded project that aims at drastic reductions of energy consumption in urban areas, employing dynamic information for improving route guiding, transport and traffic information (using information related to multiple modes of transport). Until present, no such wide-scale system and service for collecting, integrating, converting and delivering real-time traffic information has been implemented. This system is required to ensure interfacing with different information systems, collection, processing and effective transport of data to final, either mobile or fixed users. Due to the fact that, at the present moment, there are no standardised interfaces to traffic and travel information systems or local relevant authorities (such as Traffic Control Centres, Road Police, parking management, weather information providers, port, airport or railway operators etc.), the *In Time* platform had to face a consistent challenge of how to unify the formats and contents of relevant information from these various sources. Requirements related to mobile real time travel and traffic information services [1] (*WP3.2.1*, pp. 10-38) have been identified on the basis of previous research projects (such as *eMOTION*, *i-Travel*, *LINK*, *KITE etc.*) and

 ISBN: 978-1-61208-139-7

were grouped in several categories, that helped the design of the system architecture, presented further in Figure 1.

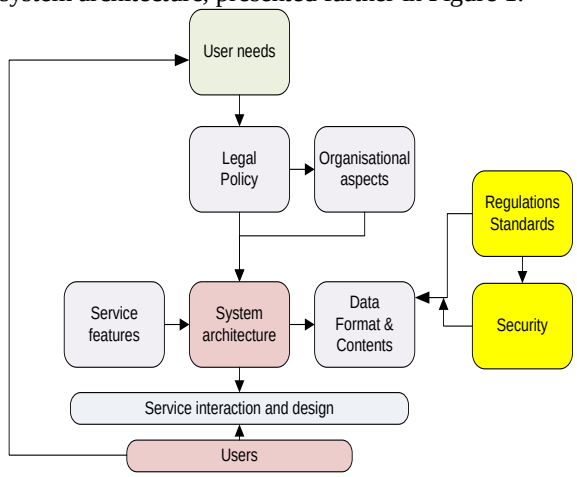


Figure 1 Factors that influenced the *In Time* system architecture design

Based on large market studies performed [1] (*WP2.2*, pp. 45-49 and *WP.3.4.1*., pp. 11-33) in the first stage of In Time project, on the interviewing with the relevant stakeholders and users, a set of services has been identified. A selective list, containing the most relevant of them, is the following:

Table 1 List of main services developed in the In Time project

Crt. No.	Services	
	Service name	***Service type***
1	Static road traffic information	Static, core
2	Dynamic road traffic information	Dynamic, core
3	Static parking information	Static, core
4	Static public transport information	Static, core
5	Walking information	Static, core
6	Dynamic road traffic routing information	Dynamic, core
7	Dynamic public transport information	Dynamic, core
8	Dynamic public transport journey routing	Dynamic, core
9	Dynamic parking information	Dynamic, core
11	Dynamic cycling planning	Dynamic, core
14	Dynamic traffic event information	Dynamic, add-on
15	Dynamic weather information	Add-on
17	Comparative pre-trip dynamic multimodal journey planning	Core, main mobile application

Taking into consideration the above services mentioned, a set of quality indicators related to the service itself and the connected meta-data have also been analysed. On its initial launch, the data management platform of the *In Time* services is able to process various information and data from various sources. Therefore, the quality of some meta-data categories is particularly important (as presented in the following Table 2):

Table 2 Meta-data categories of the unified data platform and service quality indicators

Crt. No.	**Meta-data categories**	
1	Availability of information in time	Always/seasonal/on request
2	Geo-referencing availability	Y / N
3	Regional-related information availability	Y / N
4	Content/service type	Static / dynamic
5	Max. update rate for dynamic content	Time interval
6	Content / service availability	Public / restricted
Crt. No.	**Service quality indicators**	
1	Dynamic data update time accuracy	[s]
2	Geo-referenced data quality	[lat/lon]
3	Data completeness	-
4	Matching with existing interface standard	-
5	Failure / error handling capability and reporting	-

As described above, the *In Time* data-fusion platform needs information from heterogeneous sources, with different timelines, formats, standards or interfacing requirements. Therefore, the design of the system's architecture had to take into account all these requirements. It is a difficult task to adapt into a single system so many components and to make it usable in an efficient manner. Therefore, it is expected that the services provided by this unified platform will achieve:

- Business-to-business services, enabling European-wide TISPs[1] to get access to regional traffic and travel data and services in pilot cities via a harmonised standardised open interface; this will also enable the TISPs to provide interoperable and multimodal real-time traffic and travel services (*e*-services) to their end-users; *e*-services will influence the on-trip travel behaviour by optimising journeys, taking into account the energy consumption; typical users foreseen: all persons employing mobile navigation devices;
- Web-based interoperable and intermodal pre-trip information will be provided by the pilot operators and will have the potential to influence the travel behaviour in the trip planning stage by taking into account the environmental aspects. Typical users foreseen: persons planning an urban trip for short term.

The initial research and design phase has now presently come to end and the integrated platform architecture has been successfully developed in six European pilot cities: Vienna, Munich, Florence, Bucharest, Brno and Oslo. The current phase is designed to perform wide testing of the system and its impact, both on social and environmental aspects.

[1] Traffic Information Service Providers

 ISBN: 978-1-61208-139-7

III. *IN TIME* PLATFORM ARCHITECTURE

A. *Building the In Time platform architecture*

Due to its complex tasks, the *In Time* integrated data management platform has to rely on a concept such as the multimodal Regional Data/Service Server (RDSS), which has to act as a service-oriented middleware infrastructure, providing a specific set of services and data management operations designated to cover: individual traffic information, public transport information, location based services, intermodal transport planning and/or weather information. The core of this concept is the design of a commonly agreed standardised interface that must satisfy all the requirements of standards and data formats, permitting the management of various types of information from various providers. Using such a unified interface, the users are expected to gain access to real-time multimodal data from external traffic information providers. The principles of interoperability are shown in the figure below (Figure 2):

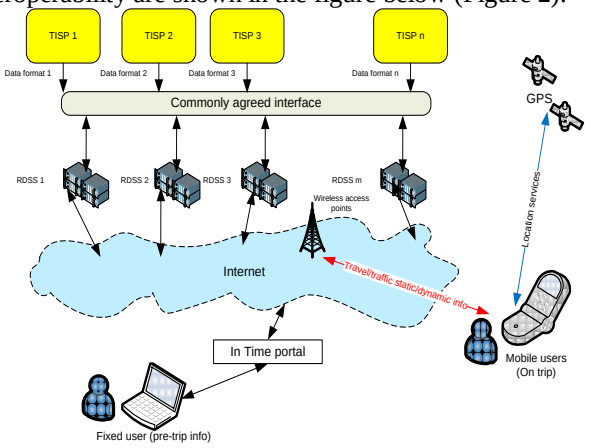


Figure 2 Principle of interoperability

The designed structure of the Commonly Agreed Interface (CAI) and the related reference standards and protocols took into account specific requirements of integration and harmonisation:

- Harmonisation of standards used by each TISP for each single domain of the integrated data management platform;
- Harmonisation across domains.

The first element means that each of the domains of *In Time* platform (traffic, public transport, parking etc.) in existing pilot systems usually make use of proprietary standardised interfaces. In the most common cases the situation is different from pilot to pilot and therefore the need for harmonisation. The second item means that, in order to achieve the goal of providing data and services really suitable for the implementation of the applications by TISPs (which integrates those data and services coming from different domains and different RDSSs) it is necessary to take into account the dependencies between the different domains.

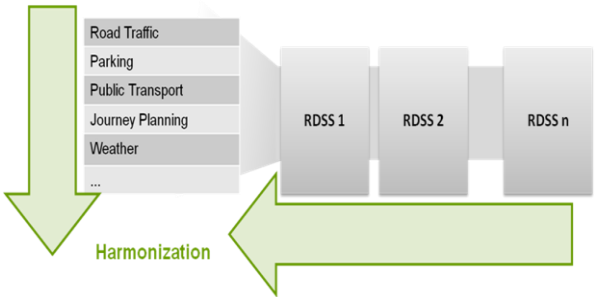


Figure 3 The "bi-dimensional" harmonisation of technical standards

Using a Model Driven Architecture (MDA) approach the specifications of the system have been designed based on the following operations:

- Modelling of data and services based on existing domain standards;
- The existing domain standards are selected and harmonised for the design of a conceptual model;
- A Geography Markup Language (GML) Application Schema and Web Service Definition Language (WSDL) were generated for obtaining exact specifications.

The CAI includes a single information space, based on a conceptual data model obtained by harmonising several international and European standards along ISO 19100 Geographic Information Standards. It includes domains such as: individual traffic (based on *DATEX II*[2]), public transport (*Transmodel*[3], *IFOPT*[4], *SIRI*[5], *TPEG*[6]), location based services, weather reporting, inter-modal transport planning.

B. *Ensuring the Path for Information Provision and Use*

The internal services of the application must ensure the correct information flowing between the different architecture entities and terminals. Amongst these, the Data Services and the Mapping Services are the most often used resources, due to the intensive employment of location based and routing information in the application. These services expose *In Time* standard interfaces for the provision of data (WFS[7]) and the maps (WMS[8]).

[2] DATEX II – a standard for the exchange of traffic related data.
[3] Transmodel – a reference data model for the public transport.
[4] IFOPT *(Identification of Fixed Objects in Public Transport)* defines a model and identification principles for the main fixed objects related to public access to public transport.
[5] SIRI *(Standard Interface for Real Time Information)* – an XML protocol which allows distributed computers to exchange real-time information about public transport vehicles and services.
[6] TPEG *(Transport Protocol Experts Group)* – a standard for delivering traffic information via digital formats such as DAB, DMB, DVB or over the Internet.
[7] WFS – Web Feature Service – a standard interface for providing requests for geographical features over the Internet
[8] WMS – Web Map Service – a standard protocol for serving georeferenced map images generated by a map server over the Internet

 ISBN: 978-1-61208-139-7

Figure 4 below shows this process.

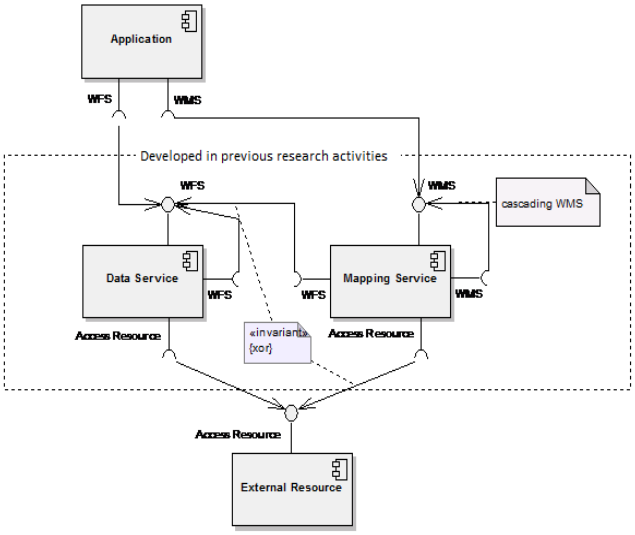


Figure 4 Integrated platform components for information provision and use

IV. LOCAL PILOT CITY IMPLEMENTATION AND INITIAL TESTS

Bucharest (Romania) is one of the pilot cities to demonstrate the *In Time* integrated platform functionalities. Here, the road traffic signalling and the public transport are managed by a common system [5] (pp.3-4), the Bucharest Traffic Management System (BTMS). This infrastructure incorporates local equipment and outstations (traffic detectors, traffic controllers, WiFi hotspots, CCTV, fibre optic communication network etc.) and central computers in the traffic control centre. The *In Time* system interfaces with BTMS and collects dynamic information from this traffic/travel information provider. As expected, the BTMS employs proprietary and in general different standards from the systems in other cities (Tables 3 and 4). The service provider (TISP) can request an *In Time* service already available at the RDSS or in case this *In Time* service is not available at the RDSS, it can request the required content from the RDSS, to allow him to create this *In Time* service himself. The quality of the services offered is one of the most important factors that define the utility of the system. This feature may be considered from a dual point of view: the quality of the information offered by local systems and the quality of service offered by the *In Time* system. Only the second one is important for this analysis. Of course, in the quality of the In Time service also has to be considered the delay in delivering real-time information from the local systems. The local tests performed in Bucharest showed the critical influence of traffic congestion in estimating the correct time of arrival at destination (ETA), especially for the "car" and "public transport" modes. Thus, the quality of the dynamic information delivered to the In Time system proved crucial. For example, Table 5 shows differences in ETAs and time lags recorded for some tests performed in Bucharest.

Table 3 Availability and interfacing with services in Bucharest, compared with other pilot cities

Service	*Bucharest*	Brno	Florence	Vienna
Routing	*Proprietary (+WDSL[9])*	TMC[10]	Proprietary	Proprietary
Location	*Proprietary*	TMC	WFS	WFS
Messaging	*Proprietary (+WDSL)*	TMC	WFS	Proprietary (XTIS[11])
Mapping	*WMS*	Proprie-tary	WMS	WMS

Table 4 Contents and standards available in BTMS compared to other pilot cities

Content	*Bucharest*	Brno	Florence	Vienna
Traffic Data	DATEX	Alert-C[12]	DATEX	TICXML[13], DATEX
Public transport	*Proprietary*	Proprie-tary	Proprietar y	Proprietary
Map data	*Images through OGC WMS[14]*	Proprie-tary	Images through OGC WMS	Images through OGC WMS

As it can be observed from the above Tables 3 and 4, different formats and standards had to be feed in a single database, via a commonly standardised interface. The Bucharest system's architecture is presented in a simplified manner in the Figure 5. The main sub-systems of the BTMS are PTM (Public Transport Management) [3] (pp. 2-5) and UTC (Urban Traffic Control). These sub-systems form an integrated component for obtaining dynamic traffic/travel information. The BTMS is connected to the RDSS via a VPN tunnel, in order to protect the information and the traffic management system. Traffic messages and other information is converted into the needed format by the CAI and then delivered to the users via RDSSs.

These tests of the In Time system include:

- Checking the existence and correct operation of all available services in the pilot city (functional testing);
- Checking the quality of the services provided (validation testing);
- Checking the impact of the services:
  - Social impact over the users;
  - Traffic impact;
  - Environmental impact.

[9] WDSL – Web Service Definition Language – a specification for describing network services as a set of endpoints operating on messages, containing either document-oriented or procedure-oriented information

[10] TMC – Traffic Message Center

[11] XTIS – Extended Tool Integrated Services

[12] Alert-C – European standard for language independent exchange of traffic information via a RDS-TMC channel

[13] TIC-XML – common XML schema containing already translated and coded data from a traffic information centre (TIC)

[14] OGC (WMS) – Open GIS (Geographical Information System) for the web

 ISBN: 978-1-61208-139-7

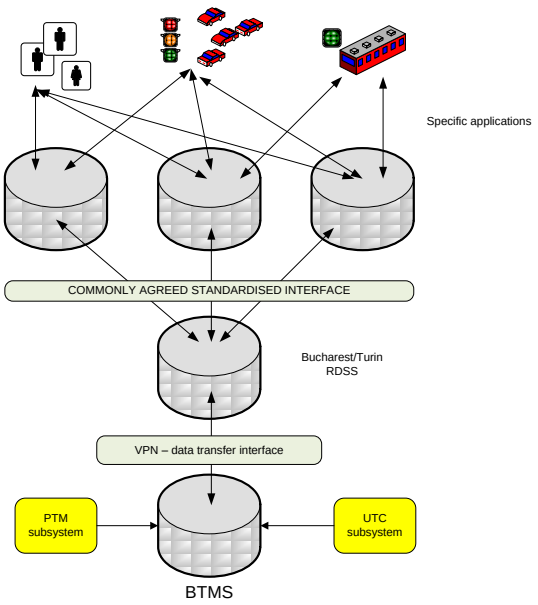


Figure 5 The Bucharest pilot *In Time* physical architecture

After the successful testing of the existence and correct operation of the *In Time* system functionalities, one of the most difficult problems is to determine the quality parameters. This is particularly difficult as the test operator has to determine if the poor response time, for example, is due to the *In Time* system or to the data service provider. Generally speaking, the existing quality concepts are close to the quality model of ISO 19113. This concept is unfortunately not exactly usable in this case, because the scope of that procedure is to use a quality model within the whole information chain and not only for a dataset from data producers. The quality characteristics *availability*, *up-to-datedness*, *completeness* and *correctness* define the connection between the geographic data and the final mobile user. The computation of *availability* can be effected by failure rate in the physical support. It indicates that the degree of probability of a failure of a dataset within an assumed period and can be measured by the function:

$$R(t) = \frac{n_{ZV}(t)}{N} \quad (1)$$

where the quotient is the number n_{zv} of missing entities at the moment t and the number N of entities in the system.

The concretion of the *"up-to-dateness"* can be effected by comparing the rate of update with the rate of change: comparison of how the dataset is changing and how the content in accordance with the universe of discourse is changing. This quality parameter can describe whether the geo-information is actually valid or not.

The omission rate is a quality parameter for completeness and can be described by the following equation:

$$CM = 1 - \frac{n_{IC}}{N} \quad (2)$$

where N is the number of entities in the modelled reality and n_{IC} the number of correspondent missing entities in the database. In a similar way can be defined the other quality parameters of the model. Beside the quality parameters of the model itself, when testing such a complex system it is important to also determine, from the users' point of view, the responsiveness of the information chain. Because the *In Time* system and services have been started very recently, only preliminary tests have been carried on in Bucharest. In the table below there are shown some of the time lags determined on different mobile devices (time measured between sending the command and receiving the answer – tested on 3 devices, information updated via GPRS/3G network) and estimations of arrival time at destination, that the mobile application delivers to the user.

Table 5 Initial results for time lag measured at the information request

Test type / mobile platform	Apple iOS	Symbian OS 5th edition	Windows Mobile 6.x
Time lag for presenting selected dest.	1-2 s	1-2 s	1-2 s
Max. time lag to determine routes to destination	14 – 80 s	20 – 82 s	17 – 55 s
Max. ETA error, "Pedestrian" mode	5 min	5 min	5 min
Max ETA error, "Car" mode	10 – 18 min	10 – 15 min	n/a
Max ETA error, "Public Transport" mode	- 5 min... + 25 min	- 5 min... +40 hours (!)	n/a

The maximum ETA error for "Public Transport" mode for the Symbian application showed strange figures in some isolated cases (see Table 5), due to possible mistakes in selecting the public transport lines, but this is probably a software issue that is to be solved in the next project steps.

V. FUTURE DEVELOPMENTS

The services that the *In Time* integrated platform is providing include a large number of physical and virtual entities: local systems/interfaces, local software applications, fixed or mobile communication networks, data processing and converting etc. Therefore, in order to ensure a good operation time and presence of these services especially to mobile users, several communication methods are to be taken into consideration. Moreover, for obtaining a good "visibility" of these services, they have to be expanded in the near future, integrating more and more applications from the transport field: expansion to other transport modes information/routing (trains, airplanes, ferries etc.) and coverage expansion (covering not only urban areas, but also interurban). In the local pilot of Bucharest city, several tests have also been performed in order to determine the usability of different WiFi hot spots for obtaining dynamic MRTTI also in bus stations, parks, street junctions etc. In this spirit, the tests took into consideration also the usability of some WiFi special hot spots, such as the ones used by the public transport. The local BTMS communication infrastructure is comprised of 150+ Access Points installed in intersections of Bucharest and WiFi clients installed on each public transport vehicle. Therefore, several tests were performed in order to

 ISBN: 978-1-61208-139-7

determine the availability and some communications' parameters, like *delay* and *throughput*, by using these WiFi Acces Points, on the 2,4 GHz and 5 GHz bands [4] (pp.3-5). The measurement area was selected in central Bucharest. There were chosen two types of intersections: busy ones (from the RF communications point of view) and free ones, with few interference and usage of WiFi spectrum. We present further the results in the junction with the most busy RF environment (many governmental buildings and public/private Access Points) – Victoriei Square. Spectrum analyses in both bands have also been performed, seeking for an observation regarding the usage of WiFi protocol in the mentioned locations. While both 2.4 GHz and 5 GHz bands are installed, the most significant activity was detected in 2.4 GHz band. This is why we considered presenting only the results regarding this band. The Victoriei Square area of study had the following characteristics (Figure 6):

- High usage of the 2.4 – 2.5 GHz spectrum, very low in 5 GHz spectrum; beside the WiFi and Bluetooth, high power CW continuous carrier transmitters have been detected;
- A large number of Access Points (over 40) have been identified in the area; WiFi channel utilization reaches in some moments 90%, especially in the zone of channels 1 to 5.
- There have been identified many interferences due to adjacent WiFi channels and also to the powerful RF continuous carrier transmitters in the area;

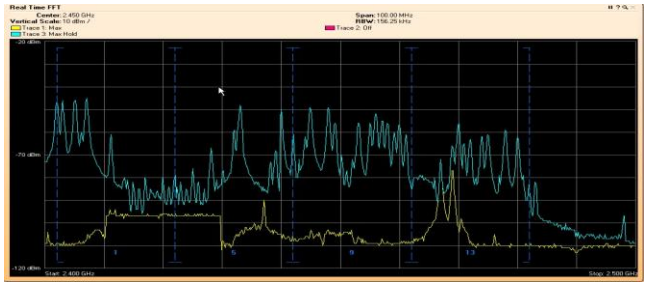


Figure 6 Spectrum analysis for the 2.4 GHz band

The next diagram presents the channel utilisation for the 2,4 GHz band in this area (Figure 7):

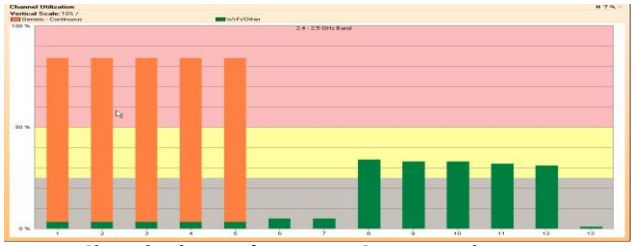


Figure 7 Channel utilisation for Victoriei Square WiFi hot spot

Regarding the WiFi protocol, the test results determined that even in this area, considered the worst case scenarios, the IP communication is possible. The maximum measured delay was about 1,2 seconds and data throughput din not decreased under 1Mbps.

In the future it will be also possible to deliver *In Time* content via any other WiFi access points in urban areas.

VI. CONCLUSIONS

The analysis of In Time system's performances performed in Bucharest showed several critical aspects:

- The errors in the ETAs depend more on the quality of the real-time traffic information (for the car and public transport modes) and less on the mobile communication network quality, although it is important to have good GPS and communication channels coverage [4] (pp. 2-3);
- The accuracy of the real-time information delivered by local traffic information providers is crucial for the precision of ETAs.
- The tests demonstrated the concept is valuable also for fuel consumption / pollutant emissions reductions, delivering better routes and therefore diminishing congestion in traffic.

It is expected that the usage of the *In Time* services will increase in the future, employing also available urban hot spots in support of the information broadcasting. The integrated services that the *In Time* project introduces will increase in usefulness and will constitute a tool to easily find traffic/transport related information anywhere. In the future, the expansion of these services will help the urban society in reducing the emissions and fuel, or energy consumption by choosing more appropriate modes for urban transport or saving time and fuel when travelling.

ACKNOWLEDGMENT

The authors would like to thank all the *In Time* project partners for their fruitful cooperation. They hope that the concept of the system and the services provided will continue to be developed via new research projects and a common, harmonised European standard in the next future.

REFERENCES

[1] * * * - CIP ICT PSP 2008-2, Research Project 238880 "Intelligent and Efficient Travel Management for European Cities (*In Time*)", 22 partners, DESCA FP 7, ICT PSP Support Programme European Union;

[2] M. Minea, Carmen Eleonora Stan, R.S. Timnea. "Integrated Platform for Road Traffic Safety Data Collection and Information Management". IARIA ICCGI Conference, IEEE, Valencia 2010;

[3] S. Dumitrescu, M. Minea. „Wireless Communication System for Public Transport Management as part of Bucharest Traffic Management System". ITS World Congress 2010, Busan, South Corea;

[4] M. Minea, S. Dumitrescu. „Vehicle to Infrastructure Communications – Technologies and EMC Problems in Public Transport Management System". IEEE 9th International Conference on Telecommunications in Modern Satellite, Cable and Broadcasting Services. TELSIKS 2009, Niš, Serbia, 7- 9 October 2009;

[5] M. Minea. „Implementation of the Bucharest Traffic Management System – solutions to problems and In-Time Project" – Invited speaker at 16th World ITS Congress, Special Interest Session SIS 33, Stockholm, Sweden, 21-25 September 2009;

 ISBN: 978-1-61208-139-7

Building a Color Recognizer System on the Smart Mobile Device for the Visually Impaired People

Hsiao Ping Lee
Department of Applied Information Sciences
Chung Shan Medical University,
Department of Medical Research
Chung Shan Medical University Hospital,
Taichung 402, Taiwan, R.O.C
ping@csmu.edu.tw

Jun-Te Huang, Chien-Hsing Chen*
Department of Applied Information Sciences
Chung Shan Medical University,
Taichung 402, Taiwan, R.O.C
cy3331@gmail.com

Tzu-Fang Sheu
Department of Computer Science and Communication Engineering
Providence University,
Taichung 433, Taiwan, R.O.C
fang@pu.edu.tw

***Abstract*—One of the most important characteristics of a developed country is to take care and promote the life quality of handicapped people. There are more than 55,000 registered visually impaired people in Taiwan, and the population is growing year by year. The amount of the assistive devices or systems for the visually impaired people, which are designed and made locally in Taiwan, are very few. Most of the assistive devices and systems available to the vision-impaired people in Taiwan are imported and expensive. It means that more and more visually impaired people in Taiwan have to face various difficulties in life, which are difficult to handle by themselves, without any help from assistive devices or systems. Therefore, it is important to develop assistive devices and systems for the visually impaired people in Taiwan. It is very difficult, or impossible, for blind persons to identify object's color. The problem might cause inconvenience or further serious danger in life, for example, wearing wrong-color cloth or taking wrong drugs. The problem on color identification can be simply solved by the assistance from sighted people. But, it is an infeasible solution because of the heavy loads and cost in human resources. In this paper, a color recognition system for blind people will be developed. The system is built on the smart phones with a camera device, and, therefore, is very portable. With an accessible user interface and the text-to-speech technology, the blind people can easily access the color information from the system. The blind people can identify object's color anywhere and anytime by using the color recognizer. The developed system provides a feasible solution to the problem of color identification that the blind people faced before. Moreover, the developed system will help the blind people to increase their life quality and decrease their need for scarce sighted assistance. The color recognizing system can help the blind people identify object's color themselves.**

Keywords- Visually Impaired People; Text-To-Speech Technology; Color Recognizer; Mobile Assistive System.

I. INTRODUCTION

Vision is an important and natural way for humans to receive information from the environment. We rely on vision to handle most things in daily life. Due to the defect in vision caused by congenital disease or accidental injuries, visually impaired people, especially totally blind people, have to face lots of inconvenience in their daily life. Therefore, it is an important issue to develop assistive devices or systems for blind people. The assistive technology for the blind has been developed for a long time in some developed countries, such as those in Europe and America. Their functions are appropriate and useful. However, the assistive devices for blind people are very few in Taiwan. In Taiwan, the assistive devices available to the blind people from Europe and America are too expensive.

A day in our life, we often faced with many choices, such as pick out an object with correct color to use from many objects. It is really common for us which have normal vision function. For visually impaired people, they face the problem again and again every day. For example, a cup, a book, maybe visually impaired people can touch their shape to know what it is and guess function. But, the blind people are still unable to know the object's color. It is too difficult to know what the object's color is it if he or she wants. To solve the problem, listen to other people is a way to help blind people know whether it is a white cup, or that a book's front cover is red. But they still can not complete the work by themselves. When blind people go to school or go to work or attend meetings, they need to use conveniently wear lounge suit, about white shirt, black suit trousers, purple tie. It is very easy for most people to pick it with correct color. But, this is very difficult for visually impaired people. As a result, family members must be a role of the eyes and stay with visually impaired people to help them. In this way, family's burden will increase. The government must to prepare a budget and train many people who to help visually impaired people. It consumes considerable money and human resources. This paper focuses on assisting visually impaired people for self-management their daily life. We develop a mobile color recognizer (MCR) for blind people. They can use it to recognize object's color correctly. To use it conveniently, MCR must have portability and practicality. Therefore, we built the system on the smart phone which has a camera device. Consider the person who is usual Chinese, we use text-to-speech (TTS) technology [1, 2, 3, 4, 5] with Chinese voice. MCR will work and output Chinese voice message, the visually impaired people can easily control the system to obtain the information which they need anywhere and anytime.

II. COLOR RECOGNIZER

Color recognizer [6] is a software which people can use to identify object's color. The recognizer is a system that

 ISBN: 978-1-61208-139-7

combines the computer, camera device and recognizer application. There are many similar products and technologies nowadays, such as face recognition system, picture recognition. Image identification applications have been used universal in daily life. However, building a color recognizer in the computer or laptop still has some problems. It is too large and weight to take along. Therefore, some products have built the color recognizer on the smart device. In this paper, we refer to the color recognizer developed by CodeFactory company [7]. Provide a new way for visually impaired people to obtain colors information around. But, this product also has some shortcomings, for example, it just work on nine kind of cell phone which use Symbian system. Moreover, the assistive system uses the English voice text-to-speech engine. It is inconvenience for many visually impaired people which use others language.

III. STATE OF ART

The color recognizer from the CodeFactory company is a mobile assistive software which developed for blind people to recognize object's color. It uses a camera device to capture images and feedback voice to user. The software's main functions have:

- It can recognize eleven kinds of colors.
- It has a simple function that determines with the light level, such as bright, dark or normal.
- Support voice feedback, the software can read the recognition results by voice.

However, the color recognition system from the CodeFactory has the following drawbacks:

- The software just can work on few phones which use Symbian OS. It is inconvenience, if I use others platform, such as Android, Windows Mobile.
- The software is too expensive, for Taiwan people, the blind people must cost 4000NTD to buy it. And, it just can use on the only one phone.
- It is only available for English voice output; for visually impaired people who usual Chinese, maybe it is unfriendly for them.

Ours system can reduce the expenses for blind people who use Chinese. Because we can product this assistive system ourselves, it is cheaper than import it from other countries.

The MCR system uses the text-to-speech engine with Chinese voice. For people who are Chinese, they can accept and control the MCR system easily.

IV. SYSTEM ARCHITECTURE

The mobile color system is a convenience assistive device for blind people. The blind people can use the camera device which is on the smart phones to take photos from the object and save it in memory. It could be a JPEG or a BMP format file. When the system gets a picture, it will analyze the picture automatically and identify the object's color. Then output the text-based messages to the TTS

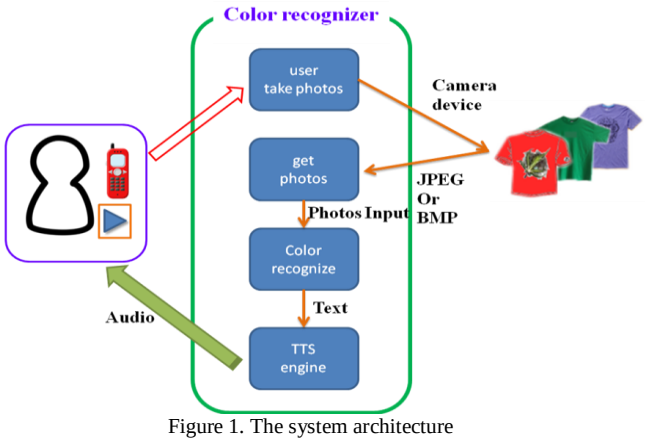


Figure 1. The system architecture

engine. TTS engine transforms the text-based message that becomes an audio-based message, and play it, providing the blind people another way to get information of colors without other people's help.

The MCR system is composed of three modules: the smart device, color recognizer and TTS. The MCR system is implemented in Java-based technology. We use the TTS with Chinese voice, which can be on the smart phones. The system architecture is presented in Figure 1.

The smart devices must have the camera because the users need to control the camera device to take photos. Considering the assistive devices' price and the available tools for blind people, we choose the smart phones from NOKIA to be the platform for our system. The smart phones use the Symbian S60 series operating system [8,9,10]. The Mobile Speak for Symbian which is a widely known screen reader software, which works on the Symbian operating system. This software can access most of information in operating system. On the other hand, the blind people can use this software to control the system easily.

We use the Java language to develop the MCR system. Java ME [11, 12, 13] is a kind of language which provides the developers with programming tools for the mobile devices. We can implement a program to drive the camera and take object's photos. When users take the photos, the system will save them in JPEG or BMP format in the memory.

The MCR system gets the images from memory and recognizes the color information automatically. Then the MCR system output the text-based message to the TTS engine with Chinese voice. TTS engine handles the text-based message and convert it to the audio-based message. As a result, the blind people can hear the audio-based message to know what the color of the objects is. In fact, they can identify the object's colors by themselves.

When this system is implemented, we hope to achieve the following results:

- Because we use Java language, the MCR system can easier transplant to other smart phones which unused Symbian operating system.

 ISBN: 978-1-61208-139-7

- Color recognition is the main function of this system, and the visually impaired people can use the system to identify their cloths color.
- Have a special and easier user interface for the blind people to control the MCR system, and they can use the system without other people's help.
- The price is cheaper, so the blind people can reduce expenses to buy the MCR assistive system.
- We hope the system is able to reduce human resources.
- The system is able to work on the mobile devices. Let the system has provided with portable property and make the life be convenience for the visually impaired people.

  With the text-to-speech function that the MCR system can output recognition result in Chinese voice message. It lets the visually impaired people who use Chinese to easily to control this system.

V. REAL IMPLEMENTATION

When we get pictures from a camera device and save them in memory, the developers are able to use the Java APIs that number JSR-135 package to drive the camera device. The MCR system will access memory to take the pictures and call the program of color identification. After that, the color recognition algorithm analyzes the pixels which a block in the pictures, and compute the RGB values. Because we hope the MCR system can work on the different platforms, we use the Java language to develop it. After obtaining the computed RGB values (which can compare with color space to know the colors information), the system provides the text-based results, and then, the Mobile Speak for Symbian screen reader software (which runs on the background) is able to access the text-based messages and provide them to TTS engine. The TTS engine converts the text-based messages to the audio-based messages, and feedback to the users.

VI. RESULTS

We are going to implement a mobile color recognizer prototype with Chinese TTS on the smart phones to support and solve the difficult problems of object's color recognition. Via the camera device, we can take the photos and save them in smart phone's memory. The color recognizer will analyze when it got the photos. Last, the text-based recognition results will be output to the TTS engine. The TTS converts the text-based message to the audio-based in Chinese voice, and play it.

We have driven the camera device successfully. Using the camera device to take photos and save them in smart phone's memory. Currently, we focus on the implementation of the color recognition algorithm. We survey some literatures about image processing or color identification. We should to know how to apply these technologies?

In the future, we face some problems when we designing the MCR system; we must to solve these problems:

- Because the assistive system is designed in Java language, we must search some APIs that the smart phones support (to read the APIs documents to know the functions and how to use them).
- We must to understand the property of the smart devices. And attempts to drive the camera device.

VII. CONCLUSIONS

In the paper, we use the NOKIA's smart phones, NOKIA 5800 to equip the MCR system. We have been completed to drive the camera device which use the JSR-135 package, and a really simple color recognition function. The color recognition algorithm analyzes the pixels and compute the RGB values to compare with the color space. At now, the color recognition function has been completed to recognize red, blue, green three basic colors. On the other hand, the screen reader software, we can download the Mobile Speak for Symbian from the Internet and use it. Currently, we survey literatures about image processing and color identification. We fix the programs and enhance the color recognition ability. When the color recognition algorithm is implemented, the MCR system is able to work well.

ACKNOWLEDGMENT

The authors would like to thank the National Science Council of the Republic of China, Taiwan R.O.C, for financially supporting this research under Grants NSC99-2218-E-040-001.

REFERENCES

[1] Prahaliad and J. Zhejiang. A Text to Speech Interface for Universal Digital Library Univ SCI, vol. 6A, no. 11, pp. 1229-1234, 2005.

[2] M.S Yu and N.H Pan. A Statistical Model with Hierarchical Structure for Predicting Prosody in a Mandarin Text-To-Speech System, Journal of the Chinese Institute of Engineers, vol. 28, no. 3, pp. 385 -399, March 2005.

[3] M. Mihkla. Modeling Speech Temporal Structure for Estonian Text-to-Speech Synthesis: Feature Selection, Institute of the Estonian Language, Tallinn, vol. 11, no. 3, pp. 284-298, 2007.

[4] A. Black and K. Lenzo. Building Voices in the Festival Speech Synthesis System. http://festvox.org/bsv/. 2000. Retrieved 2011-02-11.

[5] G. Bruce and M. Horne. Timing in speech: a multilevel process. In Prosody: theory and experiment, Kluwer Academic Publishers, pp. 281–334. 2000.

[6] Erocos International Mobile Color Recognizer, http://www.low-vision.be/HTML/mobile_Colorrecognizer.htm, Retrieved 2010-11-10.

[7] Code Factory. Making mobile phones and PDAs accessible to the blind and visually impaired, http://www.CodeFactory.es/en/products.asp?id=315, Retrieved 2011-1-25.

[8] NOKIA. S60 5th Edition SDK for Symbian OS Release Notes，

 ISBN: 978-1-61208-139-7

http://www.forum.nokia.com/info/sw.nokia.com/id/e2211487-db31-49c0-a862-a770397d7a2c/S60_5th_Edition_SDK_Release_Notes.html, Retrieved 2011-2-9.

[9] NOKIA. Symbian SDKs, http://www.forum.nokia.com/info/sw.nokia.com/id/ec866fab-4b76-49f6-b5a5-af0631419e9c/S60_All_in_One_SDKs.htm，Retrieved 2010-10-27

[10] NOKIA. ActiveState ActivePerl, http://www.forum.nokia.com/info/sw.nokia.com/id/08714ccb-f405-4ec6-b8ec-80ad83942d50/ActiveState/ActivePerl.html, Retrieved 2010-10-27.

[11] R. Ben, Mason, Sam, Rocha, Daniel, Litovski, Ivan, and Cartwright, Java Me on Symbian OS: Inside the Smartphone Model, Hayun, John Wiley & Sons Inc，April 2009.

[12] Oracle Sun Delevolpers Network. Mobile Media API - JSR135 http://java.sun.com/products/mmapi/，Retrieved 2011-5-15.

[13] Wiki Pedia. Java Mobile Media API, http://en.wikipedia.org/wiki/Java_Mobile_Media_API, Retrieved 2011.5.15.

 ISBN: 978-1-61208-139-7

Pareto Archived Simulated Annealing for Single Machine Job Shop Scheduling with Multiple Objectives

Samer Hanoun
Centre for Intelligent Systems Research
Deakin University
Geelong, Australia
samer.hanoun@deakin.edu.au

Saeid Nahavandi
Centre for Intelligent Systems Research
Deakin University
Geelong, Australia
saeid.nahavandi@deakin.edu.au

Hans Kull
Informatic Technologies
Geelong, Australia
h.kull@inmatic.com

***Abstract*—In this paper, the single machine job shop scheduling problem is studied with the objectives of minimizing the tardiness and the material cost of jobs. The simultaneous consideration of these objectives is the multi-criteria optimization problem under study. A metaheuristic procedure based on simulated annealing is proposed to find the approximate Pareto optimal (non-dominated) solutions. The two objectives are combined in one composite utility function based on the decision maker's interest in having a schedule with weighted combination. In view of the unknown nature of the weights for the defined objectives, a priori approach is applied to search for the non-dominated set of solutions based on the Pareto dominance. The obtained solutions set is presented to the decision maker to choose the best solution according to his preferences. The performance of the algorithm is evaluated in terms of the number of non-dominated schedules generated and the proximity of the obtained non-dominated front to the true Pareto front. Results show that the produced solutions do not differ significantly from the optimal solutions.**

Keywords-Multi-criteria optimization; Simulated annealing; Metaheuristic procedures; Pareto optimal; Job shop scheduling.

I. INTRODUCTION

Real industry problems require simultaneous optimization of several incomparable and conflicting criteria. Often, there is no single optimal solution; rather there is a set of alternative solutions. In joinery manufacturing, the decision maker aims at simultaneously minimizing the tardiness and the material cost for the produced jobs. Jobs with similar materials have a savings factor when scheduled together. On the other hand, the customer requires fast delivery once the order is confirmed. Therein lies a dilemma: scheduling jobs with similar materials would help control the material cost, but this would definitely increase the tardiness. Minimizing the tardiness will meet the customer's requirements, but does not generate higher revenue. A proper balance would minimize the material cost while simultaneously finishing all the jobs in a timely manner. In other words, a trade-off must be made between the material cost and a timely completion of all the jobs. Hence, in most real industry scheduling problems, we encounter the multi-objective optimization.

A general multi-objective optimization problem can be formulated in the following way. Given an n-dimensional solution space S of decision variables vector $X = \{x_1, \ldots, x_n\}$, it is required to find a vector X^* that satisfies a given set of criteria depending on K objective functions $Z(X) = \{Z_1(X), \ldots, Z_K(X)\}$. Finding the ideal vector X^* that minimizes all objective functions simultaneously is usually unfeasible. The solution space S is generally restricted by a series of constraints, such as $g_j(X^*) = b_j$ for $j = 1, \ldots, m$, and bounds on the decision variables. Objectives under consideration always conflict with each other, hence, optimizing vector X with respect to a single objective often results in unacceptable results with respect to the other objectives. Therefore, a perfect multi-objective solution that simultaneously optimizes each objective function is almost impossible. A reasonable solution to a multi-objective problem is to investigate a set of solutions, each of, which satisfies the objectives at an acceptable level, and without being dominated by any other solution. Marler and Arora [1] summarize the multi-objective optimization area within the following definitions:

1) *Dominant solution*: If all objective functions are used for minimization, a feasible solution X is said to dominate another feasible solution Y ($X \succ Y$), if $Z_i(X) \geq Z_i(Y)$ for $i = 1, \ldots, K$ and $Z_i(X) < Zi(Y)$ for at least one objective function.
2) *Pareto optimal (Efficient) solution*: A solution is said to be Pareto optimal if it is not dominated by any other solution in the solution space. A Pareto optimal solution cannot be improved with respect to any objective without worsening at least one of the other objectives.
3) *Pareto optimal set*: The set of all feasible non-dominated solutions in S is referred to as the Pareto optimal set. For many problems, the number of Pareto optimal solutions is enormous (perhaps infinite). Therefore, the problem of reducing Pareto optimal sets by obtaining the additional information is very important.
4) *Pareto front*: For a given Pareto optimal set, the corresponding objective function vector values in the objec-

 ISBN: 978-1-61208-139-7

tive space are called the Pareto front.

Scheduling problems are combinatorial optimization problems. In most cases, they are NP hard for even a single criterion optimization and are therefore unlikely to be solvable in polynomial time. The approaches are classified, Nagar et al. [2], into two groups: (1) finding the exact optimal solution using implicit enumeration methods based on either branch-and-bound or dynamic programming techniques; (2) finding a near optimal solution using heuristic methods. Heuristics are either constructive (e.g., Panneerselvam [3]) or improvement derived from metaheuristic approaches, such as genetic algorithm (GA) and simulated annealing (SA) (e.g., Sridhar and Rajendran [4], Suman [5]).

SA has become very popular for solving multicriterion optimization problems [6][7][8]. The increasing acceptance of this technique is due to its ability to: (1) find multiple solutions in a single run; (2) work without derivatives; (3) converge speedily to Pareto-optimal solutions with a high degree of accuracy; and (4) handle both continuous function and combinatorial optimization problems with ease. There have been a few techniques that incorporate the concept of Pareto-dominance. Some such methods are proposed in [9][10][11] and [12], which use Pareto-domination based acceptance criterion.

In this paper, the concept of Pareto-dominance is incorporated into the SA procedure to find the non-dominated set of solutions required by the decision maker. We start by briefly discussing the problem and the methodology of combining the objectives into a single weighted composite function in Section II. In Section III, we describe the Pareto archived simulated annealing (PASA) algorithm, and in Section IV, the computational study carried out to show the performance of the algorithm. Finally, we draw conclusions in Section V.

II. The Problem under Study

In the joinery domain, the cost of products, such as kitchens, is largely determined by the number of material sheets used in manufacturing. Jobs with similar materials can be scheduled together to decrease the amount of material waste. This leads to minimizing the production cost and therefore increase in the profit; however, not without affecting the tardiness of the jobs. The goal is to find the proper balance between these objectives.

The deterministic job shop scheduling problem considered, in this paper, consists of a finite set J of n jobs to be processed on a single machine. It is desired to find the order (schedule) in, which these n jobs should be processed to maximize the total cost savings C and minimize the total tardiness time T.

Every two jobs, j and k, with the same material have a savings factor S_{jk}, which shows the reduction in material that can be achieved when producing the two jobs in sequence ($S_{jk} = S_{kj}$). Given the number of material sheets N and the cost of a material sheet M, the cost savings CS_{jk} is calculated as:

$$CS_{jk} = CS_{kj} = M_j * (N_j + N_k) * S_{jk} \quad (1)$$

where $j = 1, \ldots, n, \quad k = 1, \ldots, n$

The total cost savings C is defined by:

$$C = \frac{1}{2} \sum_{j=1}^{n} \sum_{k=1}^{n} CS_{jk} \quad (2)$$

Each job is to be processed for an uninterrupted processing period of p_j. The process time p_j is assumed to be known in advance, and necessary setup times are included in the processing times. The tardiness T of job j is determined by the completion time c_j. It is calculated as:

$$T = \sum_{j=1}^{n} \max(0, c_j - d_j) \quad (3)$$

where d_j is the due date and c_j is the completion time of job j.

It is worth to note that in general, we may have to minimize all the objective functions, maximize them all, or minimize some functions and maximize others. However, any objective function can always be converted from the minimization form to the maximization form, and vice versa since:

$$\max(f(a)) = -\min(f(a)) \text{ and}$$
$$\min(f(a)) = -\max(f(a))$$

This conversion is applied to the total cost savings C objective to transforming it to a minimization objective.

An attractive approach adopted by several investigators [13][14][15][16] is to combine the objectives into a weighted sum:

$$E(x) = \sum_{i=1}^{K} w_i f_i(x) \quad (4)$$

The composite objective is used as the energy to be minimized in a scalar form. Therefore, the two objectives (1) the cost saving C and (2) the tardiness T are combined in one energy function as:

$$E = w * T - (1 - w) * C \quad (5)$$

where w ($0 \leq w \leq 1$) is the weight assigned to each objective during the search process.

It is clear that SA with a composite energy as in (4) and (5) will converge to points on the Pareto optimal front where the objectives have ratios given by w_i^{-1}, if such points exist. However, it is unclear how to choose the weights in advance. Recognizing this, w is initialized to 0 and is increased by 0.1 at each search process in order to realize various search directions to uncover more non-dominated solutions in the solution space.

The notations used throughout this paper are given below.
n: Number of jobs;
p_j: The processing time of job j;
d_j: The due date of job j;
σ: The current schedule;
T: Tardiness of the schedule σ;
C: Total cost savings of the schedule σ;
σ': The candidate schedule;

 ISBN: 978-1-61208-139-7

T': Tardiness of the schedule σ';
C': Total cost savings of the schedule σ';
σ_b: The best solution obtained during the search;
T_b: The best tardiness obtained during the search;
C_b: The best total cost savings obtained during the search;
w: The non-negative weight of the objectives;
Z: The weighted sum of the objectives for the schedule according to (5);

III. The Pareto Archived Simulated Annealing Algorithm (PASA)

SA is a metaheuristic algorithm based on the basic idea of neighborhoods. It was derived from the analogy between the simulation of the annealing of solid and the strategy of solving combinatorial optimization problems [17]. A neighboring solution is derived from its originator solution by a random move, which results a new slightly different solution. This increases the chance of finding an improved solution within a neighborhood more than in less correlated areas of the search space. Also, SA overcomes the problem of getting stuck in local minima, by allowing worse solutions (lesser quality) to be taken some of the time (i.e., allowing some uphill steps). The simplicity of the approach and its substantial reduction in computation time [18][19] has made it a valuable tool for solving multi-objective optimization problems [13][15][16].

In this section the main components of the PASA algorithm are presented. The implementation of the algorithm is described in Figure 1. To preserve the non-dominated solutions obtained during the search process, an archive is maintained for storage. The Pareto search and archiving procedure, as well as the procedures followed for setting the parameters are explained below.

A. Pareto Search and Archiving

The PASA algorithm starts its search with a randomly generated solution σ. This solution is added to the Pareto archive and the objectives T and C and the weighted sum, based on w, of the two objectives are calculated. A neighbour solution σ' is generated from the current solution σ using the Randomly Pairwise Interchange mechanism. The candidate solution σ' is then compared to σ for non-domination. In case of the two objectives T and C, a solution σ' is said to dominate a solution σ, if the following condition is satisfied:

$$\begin{aligned}&[((T' \leq T) \quad AND \quad (C' \geq C))\\ AND \quad &((T' < T) \quad OR \quad (C' > C))]\end{aligned} \tag{6}$$

If the candidate solution σ' dominates σ, then σ' becomes the current solution. Otherwise, the dominated candidate solution is accepted with the acceptance probability P_{accept} as given in (7).

$$P_{accept} = \exp^{-(\triangle Z/T)}, \quad \triangle Z = Z' - Z \tag{7}$$

Whenever a candidate solution σ' is accepted, it is compared with every member of the archive. Once any solution in the archive is identified as a dominated solution, it is removed from the archive. If σ' is dominated by any existing solution, then it is discarded and comparison is terminated. After all comparisons, non-dominated solutions will be left in the archive and σ' is added to the archive, if those within the archive and σ' are not dominating each other. Irrespective of whether the candidate solution is added into the archive or not, the search process is continued with the current solution.

B. Parameter Settings

The value of the initial temperature is chosen by experimentation. The range of change ΔZ in the value of the objective function with different moves is determined. The initial value of temperature t_o is calculated based on the initial acceptance ratio λ_o, and the average increase in the objective function, ΔZ_0:

$$t_o = -\frac{\Delta Z_0}{\ln(\lambda_o)} \tag{8}$$

The following steps describe the method used to calculate the value of t_o. Non-improver solutions are accepted with a probability of about 95 percent in the primary iterations (i.e., $\lambda_o = 0.95$).

Step 1:
/* Q represents the number of samples */
for $q = 1$ to Q **do**
  repeat
    Generate two solutions X_1 and X_2 at random
  until $Z(X_1) \neq Z(X_2)$
  $t_o^q = -\frac{|Z(X_1) - Z(X_2)|}{\ln(0.95)}$
end for

Step 2:
$t_o = \frac{1}{Q} \sum_{q=1}^{Q} t_o^q$

Enough number of iterations at each temperature are carried out to ensure that all represented states are searched and to enable reaching the global optimum. For our problem, a 150 non-improving iterations are used to terminate the current temperature level. The temperature is decremented in a proportional manner using the relationship $t_{i+1} = \alpha * t_i$, where α is the cooling factor constant and chosen to be 0.98. A final temperature value t_f equals to 5 percent of the initial temperature t_o is used for stopping the algorithm (i.e., $t_f = 0.05 * t_o$).

The re-annealing procedure restarts the SA process with the best solution obtained during the previous run as the seed solution. The search direction is changed by changing the weight coefficient w to uncover more non-dominated solutions. Initially, w is set to 0 and is changed with increments of 0.1 for every search process. During, the re-annealing, the temperature and other parameters are re-set to their initial values. The re-annealing process is carried until w reaches the value 1.0.

 ISBN: 978-1-61208-139-7

Algorithm PASA
Calculate the initial temperature t_0.
Initialize the Archive.
Initialize w = 0, non-improving iterations at each temperature (nt = 150), cooling factor $\alpha = 0.98$ and final temperature $t_f = 0.05 * t_0$.
Generate a random solution (σ_{seed}), add σ_{seed} to the Archive, and let $\sigma_b = \sigma_{seed}$.
while ($w \leq 1.0$)
  $t = t_0$.
  $\sigma = \sigma_{seed}$.
  Calculate T, C, and Z.
  Let $T_b = T$, $C_b = C$, and $Z_b = Z$.
  while ($t \geq t_f$)
    k = 1
    while (k $\geq nt$)
      Generate a neighbour solution σ' from σ.
      Calculate T', C', and Z'.
      if (σ' dominates σ) OR (σ' and σ are non-dominating to each other)
        $\sigma = \sigma'$, $T = T'$, $C = C'$, and $Z = Z'$.
        Check dominance of σ' w.r.t all solutions in the Archive and update the Archive.
        if (σ' dominates σ_b)
          $\sigma_b = \sigma'$, $T_b = T$, $C_b = C$, and $Z_b = Z$.
        End if
      else
        Generate a random number U.
        if ($U < e^{-\Delta Z/T}$)
          $\sigma = \sigma'$, $T = T'$, $C = C'$, and $Z = Z'$.
          Check dominance of σ' w.r.t all solutions in the Archive and update the Archive.
        End if
      End if
      k = k + 1
    End while
    $t = \alpha * t$
  End while
  $w = w + 0.1$
  $\sigma_{seed} = \sigma_b$.
End while
Return the Archive containing the generated non-dominated solutions.

Fig. 1. The PASA algorithm

IV. Computational Results

In this section, effectiveness of the proposed algorithm in obtaining the Pareto front is measured by considering the extreme solutions, i.e., the best tardiness and the best total cost savings, of the Pareto optimal solution set as the reference. The performance is verified using a number of numerical examples, inspired by the real data and generated randomly with pre-defined parameters. The problem sets used for testing consist of 5, 6, 7, 8, 9 and 10 jobs. Processing times for jobs are generated based on the job size, while the due dates are generated with different levels of tightness as proposed in [20]. The total processing time $P = \sum_{i=1}^{n} p_i$ is computed first, then the due date for each job is generated from the uniform distribution:

$$[P(1 - TF - \frac{RDD}{2})), P(1 - TF + \frac{RDD}{2}))] \quad (9)$$

where TF is the average tardiness factor and RDD is the range of due dates. The settings of $TF = 0.6$ and $RDD = 0.4$ are used.

The relative percentage deviation (RPD), defined by (10), in the objective value of the obtained non-dominated front with respect to the objective value of the extreme solution is used

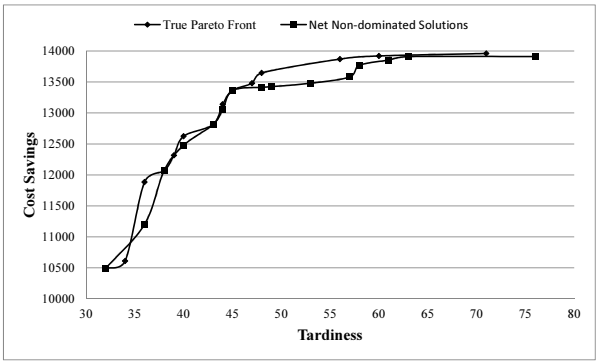


Fig. 2. True Pareto front and net non-dominated solutions for problem 22

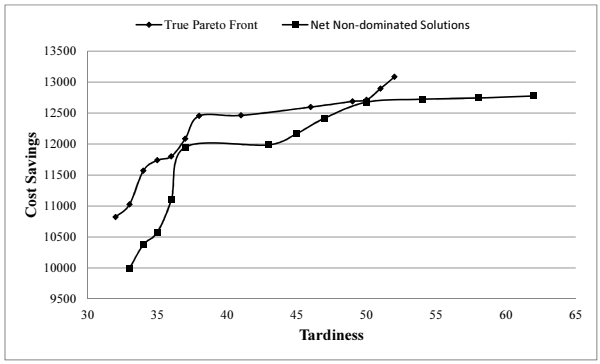


Fig. 3. True Pareto front and net non-dominated solutions for problem 27

as the main quality metric. Additionally, the mean relative percentage deviation (MRPD) is calculated for each problem set.

$$RPD = \frac{O_{obtained} - O_{extreme}}{O_{extreme}} * 100 \qquad (10)$$

Table I shows the performance results of the algorithm for the generated problem sets. The true extreme solutions are obtained by enumerating all possible schedules to find the optimal values for T and C. The best values for T and C generated by the PASA are compared to the true extreme solutions. It is observed that the extreme solutions of the non-dominated front generated by PASA are very close to extreme solutions of the corresponding Pareto front. The non-dominated solutions generated are within 2.87% in T and 1.59% in C of the true extreme Pareto solutions on the average with a maximum deviation of 5.56% in T and 2.36% in C. Table II presents the net non-dominated solutions obtained for some problem instances. Figure 2 and Figure 3 show the net non-dominated front relative to the true Pareto front for sample of the problems (problem no. 22 and problem no. 27). Given the experimental results, the PASA produced very high quality solutions with low computational complexity based on the combinatorial nature of the problem.

V. Conclusion and Future Work

In this paper, a SA algorithm is presented to find Pareto solutions for the minimization of tardiness and the maximization of material cost savings for the single machine job shop scheduling problem in the joinery manufacturing domain. Different problem sets are solved with the proposed algorithm and the approximate Pareto optimal solutions are found. These solutions are compared with the true Pareto optimal front obtained by enumeration. Results show that the proposed method generates very close solutions to the optimal solutions for some problems and the true extreme solutions for other problems. Archiving the non-dominated solutions during the search process enables the decision maker to choose the best solution according to the conditions and constraints present at the time of decision making. For future work, more than two criteria will be considered as well as problems with dynamic and stochastic data.

Acknowledgment

This work is supported under the Australian Research Council (ARC) grant number LP0991175.

References

[1] R. T. Marler and J. S. Arora, "Survey of multi-objective optimization methods for engineering," *Structural Multidisciplinary Optimization*, vol. 26, pp. 369–395, 2004.

[2] A. Nagar, J. Haddock, and S. Heragu, "Multiple and bicriteria scheduling: A literature survey," *Eur J Oper Res*, vol. 81, p. 88104, 1995.

[3] R. Panneerselvam, "Simple heuristic to minimize total tardiness in a single machine scheduling problem," *Int J Adv Manuf Tech*, vol. 30, pp. 722–726, 2006.

[4] J. Sridhar and C. Rajendran, "Scheduling in flowshop and cellular manufacturing systems with multiple objectives a genetic algorithmic approach," *Prod Plan Control*, vol. 7, pp. 374–382, 1996.

[5] B. Suman, "Multiobjective simulated annealing a metaheuristic technique for multiobjective optimization of a constrained problem," *Found Comput Decis Soc*, vol. 27, pp. 171–191, 2002.

[6] R. W. Eglese, "Simulated annealing: A tool for operational research," *Eur J Oper Res*, vol. 46, pp. 271–281, 1990.

[7] P. Serafini, "Simulated annealing for multiobjective optimization problems," in *10th International Conference on Multiple Criteria Decision Making, Taipei Taiwan*, 1992, pp. 87–96.

[8] B. Suman and P. Kumar, "A survey of simulated annealing as a tool for single and multiobjective optimization," *J Oper Res Soc*, vol. 57, pp. 1143–1160, 2006.

[9] K. Smith, R. Everson, and J. Fieldsend, "Dominance measures for multi-objective simulated annealing," in *IEEE Congr Evol Comput*, 2004, pp. 23–30.

[10] R. K. Suresh and K. M. Mohanasundaram, "Pareto archived simulated annealing for job shop scheduling with multiple objectives," *Int J Adv Manuf Technol*, vol. 29, p. 184196, 2006.

[11] A. Haidine and R. Lehnert, "Multi-case multi-objective simulated annealing (mc-mosa): New approach to adapt simulated annealing to multi-objective optimization," *International Journal of Information and Communication Engineering*, vol. 4, no. 7, pp. 507–515, 2008.

[12] S. Bandyopadhyay, S. Saha, U. Maulik, and K. Deb, "A simulated annealing-based multiobjective optimization algorithm: Amosa," *IEEE Tansactions on Evolutionary Computation*, vol. 12, pp. 269–238, 2008.

[13] P. Czyzak and A. Jaszkiewicz, "Pareto simulated annealing - a metaheuristic for multiple-objective combinatorial optimization," *Journal of Multi-Criteria Decision Analysis*, vol. 7, no. 1, pp. 34–37, 1998.

[14] A. Suppapitnarm, K. A. Seffen, G. T. Parks, and P. J. Clarkson, "A simulated annealing algorithm for multiobjective optimization," *Engineering Optimization*, vol. 33, pp. 59–85, 2000.

[15] P. Serafini, "Simulated annealing for multiobjective optimization problems," in *Multiple CriteriaDecision Making. Expand and Enrich the Domains of Thinking and Application*, G. H. Tzeng, H. F. Wang, V. P. Wen, and P. L. Yu, Eds. Springer-Verlag, Berlin, 1994, pp. 283–292.

 ISBN: 978-1-61208-139-7

TABLE I
PERFORMANCE OF THE PASA ALGORITHM, COMPARED TO THE OPTIMAL SOLUTION OBTAINED BY ENUMERATION, MEASURED IN TERMS OF THE RELATIVE PERCENTAGE DEVIATION (RPD) IN TARDINESS T AND TOTAL COST SAVINGS C

Problem no.	n	Optimal T	PASA T	RPD T	MRPD in T	Optimal C	PASA C	RPD C	MRPD in C
1	5	5	5	0.0		9523.04	9523.04	0.0	
2	5	4	4	0.0		4702.88	4702.88	0.0	
3	5	5	5	0.0	0.0	1850.37	1850.37	0.0	0.0
4	5	1	1	0.0		8106	8106	0.0	
5	5	2	2	0.0		8633.12	8633.12	0.0	
6	6	4	4	0.0		15691.5	15691.5	0.0	
7	6	15	15	0.0		5512.56	5512.56	0.0	
8	6	11	11	0.0	0.0	5687.64	5687.64	0.0	0.0
9	6	12	12	0.0		15420.24	15420.24	0.0	
10	6	13	13	0.0		14578.83	14578.83	0.0	
11	7	25	25	0.0		6203.12	6203.12	0.0	
12	7	18	18	0.0		10756.8	10756.8	0.0	
13	7	21	21	0.0	0.0	17698.68	17698.68	0.0	0.0
14	7	24	24	0.0		7284.69	7284.69	0.0	
15	7	22	22	0.0		18127.56	18127.56	0.0	
16	8	15	15	0.0		20439.54	20439.54	0.0	
17	8	18	18	0.0		23523.98	23523.98	0.0	
18	8	13	13	0.0	0.0	7512.15	7512.15	0.0	
19	8	49	49	0.0		10909.92	10909.92	0.0	
20	8	22	22	0.0		8081.01	8081.01	0.0	
21	9	33	33	0.0		28511.82	28511.82	0.0	
22	9	32	32	0.0		13960.54	13909.94	0.36	
23	9	36	36	0.0	0.0	24160.38	23958.06	0.84	0.24
24	9	25	25	0.0		23490	23490	0.0	
25	9	26	26	0.0		20104.14	20104.14	0.0	
26	10	62	62	0.0		17372.08	17330.28	0.24	
27	10	32	33	3.13		13083.84	12774.96	2.36	
28	10	33	34	3.03	2.87	10876.74	10681.86	1.79	1.59
28	10	36	38	5.56		5484.2	5365.2	2.17	
30	10	38	39	2.63		11345.43	11190.1	1.37	

TABLE II
THE NET NON-DOMINATED FRONT OBTAINED BY THE PASA ALGORITHM FOR INSTANCES OF THE PROBLEM SETS

	Prob. 1		Prob. 6		Prob. 11		Prob. 16		Prob. 22		Prob. 27	
	T	C	T	C	T	C	T	C	T	C	T	C
1	5	7979.68	4	13988.7	25	5430.04	15	16117.2	32	10489.38	33	9995.04
2	6	9336.48	5	14176.8	26	5436.2	16	17609.04	36	11202.84	34	10375.2
3	11	9523.04	7	14731.2	27	5855.08	18	18541.44	38	12068.1	35	10577.16
4			8	15691.5	28	5861.24	22	18714.6	40	12477.96	36	11099.88
5					29	6024.48	25	19174.14	43	12811.92	37	11951.28
6					32	6113.8	26	19467.18	44	13054.8	43	11986.92
7					42	6203.12	27	20439.54	45	13363.46	45	12169.08
8							34	20439.54	48	13414.06	47	12418.56
9									49	13424.18	50	12675.96
10									53	13479.84	54	12723.48
11									57	13581.04	58	12743.28
12									58	13768.26	62	12774.96
13									61	13854.28		
14									63	13909.94		
15									76	13909.94		

[16] E. L. Ulungu, J. Teghaem, P. Fortemps, and D. Tuyttens, "Mosa method: a tool for solving multiobjective combinatorial decision problems," *Journal of multi-criteria decision analysis*, vol. 8, pp. 221–236, 1999.

[17] S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi, "Optimization by simulated annealing," *Science*, vol. 20, pp. 671–680, 1983.

[18] S. Rajasekaran, "On the convergence time of simulated annealing," University of Pennsylvania Department of Computer and Information Science, Tech. Rep. MS-CIS-90-89, 1990.

[19] D. Bertsimas and J. Tsitsiklis, "Simulated annealing," *Statistical Science*, vol. 8, no. 1, pp. 10–15, 1993.

[20] C. N. Potts, "Single machine tardiness sequencing heuristics," *IIE Transactions*, vol. 23, no. 4, pp. 346–354., 1991.

 ISBN: 978-1-61208-139-7

Compiler-based Differentiation of Numerical Simulation Codes

Michel Schanen, Michael Förster, Boris Gendler, Uwe Naumann
LuFG Informatik 12: Software and Tools for Computational Engineering
RWTH Aachen University
Aachen, Germany
{schanen,foerster,bgendler,naumann}@stce.rwth-aachen.de

***Abstract*—Based on algorithmic differentiation, we present a derivative code compiler capable of transforming implementations of multivariate vector functions into a program for computing derivatives. The resulting values are accurate up to machine precision compared to the common numerical approximation by finite differences. This paper gives a short mathematical background of algorithmic differentiation while focusing on the user's perspective of applying derivative generation tools on an already implemented code. This process is illustrated by a one dimensional implementation of Burgers' equation in a generic optimization setting using for example Newton's method. In this implementation, finite differences are replaced by the computation of adjoints, thus saving an order of magnitude in terms of computational complexity.**

***Keywords*-Algorithmic Differentiation; Source Transformation; C/C++; Optimization; Numerical Simulation;**

I. Introduction

A typical problem in fluid dynamics is given by the continuous Burgers equation [1]

$$\frac{\partial u}{\partial t} + u\frac{\partial u}{\partial x} = \nu\frac{\partial^2 u}{\partial x^2} \quad , \tag{1}$$

describing shock waves moving through gases. u denotes the velocity field of the fluid with viscosity ν. Similar governing equations represent the core of many numerical simulations. Such simulations are often subject to various optimization techniques involving derivatives. Thus, Burgers' equation will serve as a case study for a compiler-based approach to the accumulation of the required derivatives.

Suppose we solve the differential equation in (1) by discretization using finite differences on a equidistant one-dimensional grid with n_x points. For given initial conditions $u_{i,0}$ with $0 < i \leq n_x$ we simulate a physical process by integrating over n_t time steps according to the leapfrog/DuFort-Frankel scheme presented in [2]. At time step j we compute $u_{i,j+1}$ for time step $j+1$ according to

$$\begin{aligned} u_{i,j+1} = u_{i,j-1} &- \frac{\Delta t}{\Delta x}\left(u_{i,j}\left(u_{i+1,j} - u_{i-1,j}\right)\right) \\ &+ \frac{2\Delta t}{\Delta x^2}\left(u_{i+1,j} - \left(u_{i,j+1} + u_{i,j-1}\right) + u_{i-1,j}\right), \end{aligned} \tag{2}$$

where Δt is the time interval and Δx is the distance between two grid points. In general, if the initial conditions $u_{i,0}$ cannot be accurately measured, they are essentially replaced by approximated values. To improve their accuracy additional observed values $u^{ob} \in \mathbb{R}^{n_x \times n_t}$ are taken into account. The discrepancy between observed values $u^{ob}_{i,j}$ and simulated values $u_{i,j}$ are evaluated by the cost function

$$y = \frac{1}{2}\sum_{i=1}^{n_x}\sum_{j=1}^{n_t}(u_{i,j} - u^{ob}_{i,j})^2 \quad , \tag{3}$$

allows us to obtain improved estimations for the initial conditions by applying, for example, Newton's method [3] to solve the data assimilation problem with Burgers' equation as constraints [4]. The single Newton steps are repeated until the residual cost y undercuts a certain threshold.

In Section II, we introduce algorithmic differentiation as implemented by our derivative code compiler `dcc` covering both the tangent-linear as well as the adjoint model. Section III provides a user's perspective on the application of `dcc`. Higher-order differentiation models are discussed in Section IV. Finally, the results of our case study are discussed in Section V.

II. Algorithmic Differentiation

The minimization of the residual is implemented by resorting to Newton's second-order method for minimization. In general, Newton's method may be applied to arbitrary differentiable multivariate vector functions $\mathbf{y} = F(\mathbf{x}) : \mathbb{R}^n \to \mathbb{R}^m$. This algorithm heavily depends on the accurate and fast computation of Jacobian and Hessian values, since one iterative step $\mathbf{x}_i \to \mathbf{x}_{i+1}$ is computed by

$$\mathbf{x}_{i+1} = \mathbf{x}_i - \nabla^2 F(\mathbf{x}_i)^{-1} \cdot \nabla F(\mathbf{x}_i) \quad . \tag{4}$$

The easiest method of approximating partial derivatives $\nabla_{x_i} F$ uses the finite difference quotient

$$\nabla_{x_i} F(\mathbf{x}) \approx \frac{F(\mathbf{x} + h \cdot \mathbf{e_i}) - F(\mathbf{x})}{h} \quad , \tag{5}$$

for the Cartesian basis vector $\mathbf{e_i} \in \mathbb{R}^n$ and with $\mathbf{x} \in \mathbb{R}^n$, $h \to 0$. In order to accumulate the Jacobian of a multivariate function the method is rerun n times to perturb each component of the input vector $\mathbf{x}$. The main advantage of this method resides in its straightforward implementation; no additional changes to the code of the function F are necessary. However, the derivatives accumulated through finite differences are only approximations. This represents a major

 ISBN: 978-1-61208-139-7

drawback for codes that simulate highly nonlinear systems, resulting in truncation and cancellation errors or simply providing wrong results. In particular by applying the Taylor expansion to the second-order centered difference quotient we derive a machine precision induced approximation error of $\frac{\epsilon}{h^2}$, with ϵ being the rounding error.

Algorithmic differentiation (AD) [5] solves this problem analytically, changing the underlying code to compute derivatives by applying symbolic differentiation rules to individual assignments and using the chain rule to propagate derivatives along the flow of control. The achieved accuracy only depends on the machine's precision ϵ. There exist two distinct derivative models, differing in the order of application of the associative chain rule. Let ∇F be the Jacobian of F. The *tangent-linear* code

$$\begin{aligned} &F\big(\overset{\downarrow}{\mathbf{x}}, \underset{\downarrow}{\mathbf{y}}\big) \xrightarrow{\texttt{dcc}} \dot{F}\big(\overset{\downarrow}{\mathbf{x}}, \overset{\downarrow}{\dot{\mathbf{x}}}, \underset{\downarrow}{\mathbf{y}}, \underset{\downarrow}{\dot{\mathbf{y}}}\big) \quad , \\ &\text{where} \\ &\qquad \dot{\mathbf{y}} = \nabla F\left(\mathbf{x}\right) \cdot \dot{\mathbf{x}} \\ &\text{and} \quad \mathbf{y} = F(\mathbf{x}) \quad , \end{aligned} \tag{6}$$

of F computes the directional derivative $\dot{\mathbf{y}}$ of the outputs $\mathbf{y}$ with respect to the inputs $\mathbf{x}$ for a given direction $\dot{\mathbf{x}} \in \mathbb{R}^n$, while arrows designate inputs and outputs. By iteratively setting $\dot{\mathbf{x}}$ equal to each of the n Cartesian basis vectors in $\mathbb{R}^n$, we accumulate the entire Jacobian. This leads to a runtime complexity identical to finite differences of $\mathcal{O}(n) \cdot cost(F)$, where $cost(F)$ denotes the computational cost of a single function evaluation.

By exploiting the associativity of the chain rule, the *adjoint* code

$$\begin{aligned} &F\big(\overset{\downarrow}{\mathbf{x}}, \underset{\downarrow}{\mathbf{y}}\big) \xrightarrow{\texttt{dcc}} \bar{F}\big(\overset{\downarrow}{\mathbf{x}}, \overset{\downarrow}{\bar{\mathbf{x}}}, \underset{\downarrow}{\mathbf{y}}, \underset{\downarrow}{\bar{\mathbf{y}}}\big) \quad , \\ &\text{where} \\ &\qquad \mathbf{y} = F(\mathbf{x}) \\ &\text{and} \quad \bar{\mathbf{x}} = \bar{\mathbf{x}} + \nabla F\left(\mathbf{x}\right)^{\intercal} \cdot \bar{\mathbf{y}} \quad , \end{aligned} \tag{7}$$

of F computes *adjoints* $\bar{\mathbf{x}} \in \mathbb{R}^n$ of the inputs $\mathbf{x}$ for given adjoints $\bar{\mathbf{y}} \in \mathbb{R}^m$ of the outputs. To accumulate the entire Jacobian we have to iteratively set $\bar{\mathbf{y}}$ equal to each Cartesian basis vector of $\mathbb{R}^m$ yielding a runtime complexity of $\mathcal{O}(m) \cdot cost(F)$. Note that for scalar functions with $m = 1$ the accumulation of the Jacobian amounts to the computation of one gradient yielding a runtime cost of $\mathcal{O}(1) \cdot cost(F)$ for the adjoint model compared to $\mathcal{O}(n) \cdot cost(F)$ for the tangent-linear model. In this particular case, we are able to compute gradients at a small constant multiple of the cost of a single function evaluation. The reduction of this factor down toward the theoretical minimum of three [5] is one of the major challenges addressed by ongoing research and development in the field of AD [6], [7], [8].

III. dcc - A Derivative Code Compiler

Numerical optimization problems are commonly implemented as multivariate scalar functions $y = F(\mathbf{x}) : \mathbb{R}^n \to \mathbb{R}$, describing some residual y of a numerical model. We assume that the goal is to minimize a norm of this residual y by adapting the inputs $\mathbf{x}$. Therefore, for better readability and without the loss of generality, in this paper, we will only cover multivariate scalar functions.

The main link between dcc and the mathematical models of AD is the ability to decompose each function implementation into single assignment code (SAC) as follows:

$$\begin{aligned} &\text{for } j = n, \ldots, n+p \\ &\qquad v_j = \varphi_j(v_i)_{i \prec j} \quad . \end{aligned} \tag{8}$$

The entire program is regarded as a sequence of $p+1$ elemental statements. In each statement an elemental function φ_j is applied to a set of variables $(v_i)_{i \prec j}$ yielding the unique *intermediate* variable v_j with $i \prec j$ denoting a dependence of v_j on v_i. The *independent* inputs are given by $v_i = x_i$ for $i = 0, \ldots, n-1$ while the *dependent* output of F is the final value $y = v_{n+p}$. When dcc applies the tangent-linear model to each of the $p+1$ assignments, we obtain

$$\begin{aligned} &\text{for } j = n, \ldots, n+p \\ &\qquad \dot{v}_j = \sum_{i \prec j} \frac{\partial \varphi_j}{\partial v_i} \cdot \dot{v}_i \\ &\qquad v_j = \varphi_j(v_i)_{i \prec j} \quad . \end{aligned} \tag{9}$$

Considering the j-th assignment in (9), the local k-th entry of the gradient $(\frac{\partial \varphi_j}{\partial v_k})_{k \prec j}$ is provided in $\dot{v}_j$ by setting $\dot{v}_k$ to one and all $(\dot{v}_i)_{k \neq i \prec j}$ to zero. The gradient component $(\frac{\partial y}{\partial x_k})_{k \in \{0,\ldots,n-1\}}$ is obtained by evaluating (9) and setting $\dot{x}_k$ to one and all other $(\dot{x}_i)_{k \neq i \in \{0,\ldots,n-1\}}$ to zero. To get the whole gradient we have to evaluate (9) n times letting $\dot{\mathbf{x}}$ range over the Cartesian basis vectors in $\mathbb{R}^n$. The adjoint model is acquired by transforming (8) into:

$$\begin{aligned} &\text{for } j = n, \ldots, n+p \\ &\qquad v_j = \varphi_j(v_i)_{i \prec j} \\ &\qquad \text{for } i \prec j \text{ and } j = n+p, \ldots, n \\ &\qquad\qquad \bar{v}_i = \bar{v}_i + \frac{\partial \varphi_j}{\partial v_i}(v_k)_{k \prec j} \cdot \bar{v}_j \quad . \end{aligned} \tag{10}$$

The first part consists of the original assignments $j = n, \ldots, n+p$ and is called *forward section*. The *reverse section* follows with the computation of the adjoint variables in the order $j = n+p, \ldots, n$. Note the reversed order of the assignments as well as the changed data flow of the left and right-hand sides compared with the original assignments. To compute the local gradient $(\frac{\partial \varphi_j}{\partial v_k})_{k \prec j}$ we have to initialize $(\bar{v}_i)_{i \prec j}$ with zero and $\bar{v}_j$ with one. The initialization with zero is mandatory because $(\bar{v}_i)_{i \prec j}$ occurs in (10) on both sides of the adjoint assignment. According to (7), the adjoint variable $\bar{v}_j$ is an input variable. Therefore it is initialized with the "Cartesian basis vector" in $\mathbb{R}$.

 ISBN: 978-1-61208-139-7

The important advantage of the adjoint model is that by evaluating (10) only once we obtain the full gradient $\frac{\partial y}{\partial \mathbf{x}}$ in $\bar{x}_i = \bar{v}_i$ for $i = 0, \ldots, n-1$. To achieve this we have to initialize $(\bar{x}_i)_{i=0,\ldots,n-1}$ with zero and $\bar{y}$ with one. As mentioned above $\bar{\mathbf{x}}$ must be zero because it occurs not only on the left-hand side in (7) and y is initialized with the value of the Cartesian basis vector in $\mathbb{R}$.

In (8), we assumed that the input code is given as a SAC. This is an oversimplification in terms of real codes. The adjoint code has to deal with the fact that real code variables are overwritten frequently. One way to simulate the predicate of unique intermediate variables is to store certain left-hand side variables on a stack during the augmented forward section. Candidates for storing on the stack are those variables that are being overwritten and are required for later use during the computation of the local gradients and associated adjoints. Before evaluating the corresponding adjoint assignment in the reverse section the values are restored from the stack.

For illustration purposes we consider Listing 1 showing an implementation of the non-linear reduction $y(\mathbf{x}) = \prod_{i=0}^{n-1} \sin(x_i)$. dcc parses only functions with **void** as a return type (line 1). All inputs and return values are passed through the arguments, which in turn only consist of arrays (called by pointers) and scalar values (called by reference). Additionally we may pass an arbitrary number of integer arguments by value or by reference. We assume that all differentiable functions are implemented using values of type **double**. Therefore, only variables of type **double** are directly affected by the differentiation process.

```
1  void f(int n, double *x, double &y)
2  {
3    int i=0;
4    y=0;
5    for(i=0;i<n;i++) {
6      y=y*sin(x[i]);
7    }
8  }
```

Listing 1: dcc input code.

Using the command line `dcc f.c -t`, we instruct the compiler to use the tangent-linear (`-t`) mode in order to generate the function t1_f (**t**angent-linear, **1**st-order version of f) presented in Listing 2. The original function arguments x and y are augmented with their associated tangent-linear variables t1_x and t1_y. Inside a driver program this code has to be rerun n times letting the input vector t1_x range over the Cartesian basis vectors in $\mathbb{R}^n$ to accumulate the entire gradient. Listing 3 shows how to use the generated code of Listing 2 in a driver program. Lines 2 and 5 let input variable t1_x range over the Cartesian basis vectors. By setting t1_x[i] to 1 the function t1_f (line 3) computes the partial derivative of y with respect to x[i].

The command line `dcc f.c -a` tells dcc to apply the adjoint mode (`-a`) to `f.c`. The result is the function

```
1  void t1_f(int n, double* x, double* t1_x
2                 , double& y, double& t1_y)
3  {
4    ...
5    for(int i=0;i<n;i++) {
6      y=y*sin(x[i]);
7      t1_y=t1_y*sin(x[i])+y*cos(x[i])*t1_x[i];
8    }
9    ...
10 }
```

Listing 2: Tangent-linear version of f as generated by dcc

```
1  for(int i=0; i<n;i++) {
2    t1_x[i]=1;
3    t1_f(n, x, t1_x, y, t1_y);
4    gradient[i]=t1_y;
5    t1_x[i]=0;
6  }
```

Listing 3: Driver for t1_f

a1_f (**a**djoint, **1**st-order version of f) shown in Listing 4. As in the tangent-linear case each function argument is augmented by an associated adjoint component, here a1_x and a1_y. As mentioned above we need a stack in the adjoint code for storing data during the forward section. The *augmented forward section* uses stacks to store values that are being overwritten and to store the control flow. The actual implementation of the stack is not under consideration here; therefore we replaced the calls to the stacks with macro definitions for better readability. By default, dcc generates code that uses static arrays which ensures high runtime performance. There are three different stacks used in the adjoint code. The stack called CS is for storing the control flow, FDS takes floating point values and IDS keeps integer values. The unique identifier of the two basic blocks [9] in the forward section are stored in lines 6 and 9. For example, after evaluating the augmented forward section of Listing 4, the stack CS contains the following sequence

$$0, \underbrace{1, \ldots, 1}_{\text{n times}} \tag{11}$$

In line 10, variable y is stored onto the stack because it is overwritten in each iteration although needed in line 21. Hence, we restore the value of y in line 20. For the same reason we store and restore the value of i in line 11 and 19. The reverse section consist of a loop that processes the control flow stack CS. The basic block identifiers are restored from the stack and depending on the value, the corresponding adjoint basic block is executed. For example, the sequence given in (11) as content in the CS stack leads to a n-times evaluation of the adjoint basic block one and afterward one evaluation of the adjoint basic block zero. The basic block one in line 9 to 11 has the corresponding adjoint basic block in line 19 to 22. In contrast to (7), in

 ISBN: 978-1-61208-139-7

```
1  void a1_f(int n, double* x, double* a1_x,
2                   double& y, double& a1_y)
3  {
4    int i=0;
5    // augmented forward section
6    CS_PUSH(0);
7    y=0;
8    for ( i=0; i<n; i++) {
9           CS_PUSH(1);
10          FDS_PUSH(y); y=y*sin(x[i]);
11          IDS_PUSH(i);
12   }
13   // reverse section
14   while (CS_NON_EMPTY) {
15          if (CS_TOP==0) {
16            a1_y=0;
17          }
18          if (CS_TOP==1) {
19            IDS_POP(i);
20            FDS_POP(y);
21            a1_x[i]+=y*cos(x[i])*a1_y;
22            a1_y=sin(x[i])*a1_y;
23          }
24          CS_POP;
25   }
26 }
```

Listing 4: Adjoint dcc output

line 22 the adjoint a1_y is not incremented but assigned. This is due to the fact that y is on both hand sides of the original assignment in line 10. This brings an aliasing effect into play. This effect can be avoided with help of intermediate variables; making this code difficult to read. For that reason we show the adjoint assignment without intermediate variables. dcc generates adjoint assignments with intermediate variables and incrementation of the left-hand side as shown in (7). The dcc-generated code and the one shown here are semantically equivalent. To accumulate the gradient using the function a1_f, we again have to write a driver, presented in Listing 5. It is sufficient to initialize the adjoint variable a1_y and call the adjoint function a1_f only once to get the whole gradient (line 2), illustrating the reduced runtime complexity of the adjoint mode.

```
1  a1_y=1;
2  a1_f(n, x, a1_x, y, a1_y);
3  for(int j=0; j<n;j++)
4     gradient[j]=a1_x[j];
```

Listing 5: Driver for a1_f

IV. Higher Order Differentiation

Numerical optimization algorithms often involve higher-order derivative models. Thus, the need for Hessians is imminent. With this in mind, dcc was designed to generate higher-order derivative codes effortlessly using its *reapplication feature*. dcc is able to generate jth-order derivative code by reading $(j-1)$th-order derivative code as the input. In this section we will focus on second-order models.

The tangent-linear mode reapplied to the first-order tangent-linear code (6) with $m = 1$ for scalar functions yields the second-order tangent-linear code

$$\dot{F}\big(\overset{\downarrow}{\mathbf{x}},\overset{\downarrow}{\dot{\mathbf{x}}},y,\dot{y}\big) \xrightarrow{\texttt{dcc}} \tilde{\dot{F}}\big(\overset{\downarrow}{\mathbf{x}},\overset{\downarrow}{\tilde{\mathbf{x}}},\overset{\downarrow}{\dot{\mathbf{x}}},\overset{\downarrow}{\dot{\tilde{\mathbf{x}}}},y,\tilde{y},\dot{y},\dot{\tilde{y}}\big) \quad ,$$

where

$$\begin{aligned}
\tilde{y} &= \left(\nabla^2 F(\mathbf{x})\cdot\dot{\mathbf{x}}\right)^{\intercal}\cdot\tilde{\mathbf{x}} + \nabla F(\mathbf{x})\cdot\dot{\tilde{\mathbf{x}}} \quad , \\
\dot{y} &= \nabla F(\mathbf{x})\cdot\dot{\mathbf{x}} \quad , \\
\tilde{y} &= \nabla F(\mathbf{x})\cdot\tilde{\mathbf{x}} \quad \text{and} \\
y &= F(\mathbf{x}) \quad .
\end{aligned} \tag{12}$$

Again, dcc generates exactly the implementation of the mathematical model. As we see in (12), the term $\nabla F(\mathbf{x})\cdot\tilde{\dot{\mathbf{x}}}$ must be equal to 0 in order to accumulate the entries of the Hessian $\nabla^2 F$. As a consequence, $\tilde{\dot{\mathbf{x}}}$ must be set to 0 on input. The product $\left(\nabla^2 F(\mathbf{x})\cdot\dot{\mathbf{x}}\right)^{\intercal}\cdot\tilde{\mathbf{x}}$ represents a projection of the Hessian, determined by the vectors $\dot{\mathbf{x}}$ and $\tilde{\mathbf{x}}$. In our case with $m = 1$ the Hessian $\nabla^2 F \in \mathbb{R}^{n\times n}$ has n^2 entries.

To compute the entry $\nabla F_{i,j}$ of the Hessian the vectors $\tilde{\mathbf{x}}$ and $\dot{\mathbf{x}}$ have to be set to the i-th and j-th Cartesian basis vectors, respectively. In order to accumulate the whole Hessian this step has to be repeated for each entry, yielding a computational complexity of $\mathcal{O}(n^2)\cdot cost(F)$. Taking either adjoint or tangent-linear first-order input code, we reapply dcc by invoking dcc -t -d 2 t1_foo.cpp. This tells dcc to generate second-order (-d 2) tangent-linear (-t) derivative code while avoiding internal namespace clashes.

Looking at the possible combinations of the two differentiation models, there exist another three second-order models. We may either apply the adjoint model to the tangent-linear code or apply the adjoint mode to the adjoint code. We will focus on the model where tangent-linear mode is applied to the adjoint code, called *tangent-linear over adjoint* mode.

This time the adjoint code (7) is taken as the input for the reapplication of the tangent-linear mode, obtaining

$$\bar{F}\big(\overset{\downarrow}{\mathbf{x}},\overset{\downarrow}{\bar{\mathbf{x}}},y,\overset{\downarrow}{\bar{y}}\big) \xrightarrow{\texttt{dcc}} \dot{\bar{F}}\big(\overset{\downarrow}{\mathbf{x}},\overset{\downarrow}{\dot{\mathbf{x}}},\overset{\downarrow}{\bar{\mathbf{x}}},\overset{\downarrow}{\dot{\bar{\mathbf{x}}}},y,\dot{y},\overset{\downarrow}{\bar{y}},\overset{\downarrow}{\dot{\bar{y}}}\big) \quad ,$$

where

$$\begin{aligned}
\dot{y} &= \nabla F(\mathbf{x})\cdot\dot{\mathbf{x}} \quad , \\
y &= F(\mathbf{x}) \quad , \\
\dot{\bar{\mathbf{x}}} &= \dot{\bar{\mathbf{x}}} + \dot{\mathbf{x}}^{\intercal}\cdot\nabla^2 F(\mathbf{x})\cdot\bar{y} + \nabla F(\mathbf{x})^{\intercal}\cdot\dot{\bar{y}} \quad \text{and} \\
\bar{\mathbf{x}} &= \bar{\mathbf{x}} + \nabla F(\mathbf{x})^{\intercal}\cdot\bar{y} \quad .
\end{aligned} \tag{13}$$

The generated implementation computes the term $\dot{\mathbf{x}}^{\intercal}\cdot\nabla^2 F(\mathbf{x})\cdot\bar{y}$. This time we do not end up with one single entry, but we are able to harvest one complete row $\nabla^2 F_i$ of the Hessian in $\dot{\mathbf{x}}$. To achieve this, the term $\nabla F(\mathbf{x})^{\intercal}\cdot\dot{\bar{y}}$ and thus $\dot{\bar{y}}$ must be set to 0 on input. The scalar $\bar{y}$ must be set to 1. Finally to compute a row of the

 ISBN: 978-1-61208-139-7

n	250	500	1000	2000
f (s)	0.03	0.08	0.15	0.32
TLM (s)	33	109	457	1615
ADJ (s)	0.21	0.43	0.85	1.82
TLM-ADJ (s)	150	587	2286	8559
IDS size	7500502	15001002	30002002	60004002
FDS size	5000002	10000002	20000002	40000002
CS size	7500503	15001003	30002003	60004003

Table I: Time and memory requirements for gradient computation

Hessian $\nabla^2 F_i$, $\dot{x}$ must be set to the i-th Cartesian basis vector. As such, we have to rerun this model n times in order to accumulate the whole Hessian, yielding only a linear increase in runtime complexity of $\mathcal{O}(n) \cdot cost\,(F)$.

The desired `dcc` command is `dcc -a -d 2 t1_foo.cpp` resulting in the file `a2_t1_foo.cpp`. The option `-a` instructs `dcc` to generate adjoint code.

V. Case Study

As discussed in Section I, we run a test case on an inverse problem based on Burgers' equation (1). As a start we take the code presented in [2] implementing the original function with the signature of

```
1 void f(int n, int nt, double& cost, double**
      u, double* ui...)
2 {
3 ...
4 }
```

Listing 6: Signature of Burgers' function

Taking `n` grid points of `ui` as the initial conditions we integrate over `nt` timesteps. The values are saved in the two dimensional array `u` for each grid point `i` and time step `j`.

To solve the inverse problem we need the derivatives of `cost` with respect to the initial conditions `ui`.

The results in Table I represent the runtime of one full gradient accumulation as well as the memory requirements in adjoint and tangent-linear mode. Additionally one Hessian accumulation is performed using the tangent-linear over adjoint model (13). Different problem sizes are simulated with varying `n`. We also mention the different stack size shown in Section III.

If we assume four bytes per integer and control stack element plus eight bytes for a floating data stack element we end up with a memory requirement of $\approx$ 610 MB for the Hessian accumulation. The tests were running on a GenuineIntel computer with Intel(R) Core(TM)2 Duo CPU and with 2000.000 MHz cpu.

The execution time of the tangent-linear gradient computation is growing proportionally to the problem size `nx` and the execution time of `f`:

$$FM : \frac{cost(F')}{cost(F)} \sim \mathcal{O}(n).$$

The single execuiton of *t1_f* takes approximately twice so much as the execution of *f*.

The execution time of the adjoint gradient computation is growing only proportional to the execution time of `f`:

$$AM : \frac{cost(F')}{cost(F)} \sim \mathcal{O}(1).$$

Finally we accumulate the Hessian using tangent-linear over adjoint mode. Here, the runtime is growing linearly with respect to `n` as well as `f` since the dimension of the dependent `cost` is equal to 1.

$$FM - AM : \frac{cost(F'')}{cost(F)} \sim \mathcal{O}(n).$$

For scalar functions in particular, the runtime complexity for accumulating the Hessian using AD is the same as the runtime complexity of the gradient accumulation using finite difference. This enables developers to implement a second-order model where a first-order model has been used so far.

VI. Outlook & Conclusion

We have presented a source transformation compiler for a restricted subset of C/C++. As such, `dcc` runs on any system with a valid C/C++ compiler making it a very portable tool. Its unique reapplication feature allows code to be transformed up to any order of differentiation.

Additionally, several extensions were implemented. As these programs run on cluster systems, they often rely on parallelization techniques. The most widely used parallelization method is MPI. Hence, there is a need for adjoint MPI enabled code [10]. This feature has been integrated into `dcc` using an adjoint MPI library [11]. Additionally there are attempts to achieve the same goal with OpenMP [12]. For the sake of brevity we did not mention the program analysis `dcc` performs. For better efficiency, `dcc` uses *activity* and *TBR* analyses [13].

References

[1] D. Zwillinger, *Handbook of Differential Equations, 3rd ed.* Boston, MA: Academic Press, 1997.

[2] E. Kalnay, "Atmospheric modeling, data assimilation and predictability," 2003.

[3] T. Kelley, *Solving nonlinear equations with Newton's method*, ser. Fundamentals of Algorithms. Philadelphia, PA: Society for Industrial and Applied Mathematics (SIAM), 2003.

[4] A. Tikhonov, "On the stability of inverse problems," *Dokl. Akad. Nauk SSSR*, vol. 39, no. 5, pp. 195–198, 1943.

[5] A. Griewank and A. Walter, *Evaluating Derivatives. Principles and Techniques of Algorithmic Differentiation (2nd Edition)*. Philadelphia: SIAM, 2008.

[6] G. Corliss and A. Griewank, Eds., *Automatic Differentiation: Theory, Implementation, and Application*, ser. Proceedings Series. Philadelphia: SIAM, 1991.

 ISBN: 978-1-61208-139-7

[7] G. Corliss, C. Faure, A. Griewank, L. Hascoët, and U. Naumann, Eds., *Automatic Differentiation of Algorithms – From Simulation to Optimization*. New York: Springer, 2002.

[8] U. Naumann, "Dag Reversal is NP-complete," *Journal of Discrete Algorithms*, vol. 7, no. 4, pp. 40–410, 2009. [Online]. Available: http://www.sciencedirect.com/science/article/B758J-4THC1FD-2/2/7ddfc2eab484bbe184d4dcdf16d8e58a

[9] A. Aho, M. Lam, R. Sethi, and J. Ullman, *Compilers. Principles, Techniques, and Tools (Second Edition)*. Reading, MA: Addison-Wesley, 2007.

[10] P. Hovland and C. Bischof, "Automatic differentiation for message-passing parallel programs," in *Parallel Processing Symposium, 1998. IPPS/SPDP 1998. Proceedings of the First Merged International ... and Symposium on Parallel and Distributed Processing 1998*, mar-3 apr 1998, pp. 98 –104.

[11] M. Schanen, U. Naumann, and M. Förster, "Second-order adjoint algorithmic differentiation by source transformation of mpi code," in *Recent Advances in the Message Passing Interface, Lecture Notes in Computer Science*. Springer, 2010, pp. 257–264.

[12] OpenMP Architecture Review Board, "OpenMP Application Program Interface," Specification, 2008. [Online]. Available: http://www.openmp.org/mp-documents/spec30.pdf

[13] L. Hascoët, U. Naumann, and V. Pascual, "To-be-recorded analysis in reverse mode automatic differentiation," *Future Generation Computer Systems*, vol. 21, pp. 1401–1417, 2005.

 ISBN: 978-1-61208-139-7

Analysis of BitTorrent Networks

Daniel Kowalczyk, Leszek Koszalka
Department of Systems and Computer Networks,
Wroclaw University of Technology,
Wroclaw, Poland
E-mail: {daniel.kowalczyk, leszek.koszalka}@pwr.wroc.pl

***Abstract*—Peer-to-peer systems have emerged as an attractive alternative to client/server approaches. By efficiently leveraging the upload bandwidth of the end users, BitTorrent becomes a standard for scalable content distribution. In this paper, we concentrate on the overall performance of BitTorrent, in particular on impact of physical media distribution, such as fluctuations in cable and inside protocol parameter configuration. It is concluded that the decrease of the default optimistic un-choking time has a highly positive impact on the protocol performance. Moreover, it is shown that the delay of the client's network connection has also remarkable impact on the performance of the BitTorrent protocol.**

Keywords- network traffic; peer-to-peer network; network protocol; simulation experiment; efficiency.

I. INTRODUCTION

In the past few years we can observe growing popularity of peer-to-peer (P2P) networks. One of the peer-to-peer protocols is the BitTorrent (BT) [1], released ten years ago [2]. A recent analysis of the latest P2P trends word-wide shows that BT is still the most popular file-sharing protocol being recognized as the king of P2P traffic, because of generating approximately 45-78% of all P2P traffic, and 27-55% of all Internet traffic [3]. In our opinion, such a big popularity of this protocol is caused mainly by relatively simple architecture and so-called tit-for-tat mechanism. Tit-for-tat (TFT) policy [4] is used in BT to encourage each peer to upload to other peers while downloading [5]. Dongyu et al. [6] showed that with TFT a peer with a smaller upload rate will get slower download speed.

In this paper, we do make an attempt of some-aspects analysis of the BT protocol on the basis of results of simulation experiments. The simulation environment may have some advantages in comparison to the real network. Firstly, we can manage the peers in swarm parameters that in the real environment are random, e.g., the number of peers in swarm. Secondly, the whole experiment in the real environment would last days, weeks or even months [7], when using BIT-SIM it takes up to couple of minutes. Thirdly, in simulation environment we have an opportunity to create a proper scenario, proper bandwidth and type of internet connection that would be in real world also very time-consuming activity.

Our experimentation system was implemented in OMNeT++ simulation environment. This implementation was based on De Vogeleer, Erman, and Popescu ideas described initially in [8] and next, developed as BIT-SIM simulator [9]. After analysis of the obtained results of preliminary experiments, we stated the following research theses: (i) that optimistic un-choking session times have a big impact on overall BitTorrent performance, (ii) that link delay fluctuations may have a negative influence on BitTorrent downloading mechanism of efficiency.

The rest of the paper is organized as follows. Section II provides a review of simulators used for modeling and P2P network and BitTorrent file-sharing system. This review is based on the related works. In Section III the simulation environment is described, including models of network processes. In Section IV, we present the concept of research and an experiment design. The results of investigations are presented and discussed in Section V in correspondence to our research theses. Final remarks and some ideas for future research appear in Section VI.

II. RELATED WORKS

There are number of existing P2P simulators. In this section, we will overview some of these simulators. Traditional packet-level network simulators provide accurate low-level models of the network hardware and protocols but are too detailed to be effective in analysis of large scale P2P networks. For example, the ns-family environment, including the most widely used ns-2 simulator [10]. However, ns-2 and ns-3 simulators have weaknesses: (i) they are too detailed to be effective in analysis of large scale P2P networks, and (ii) they are very troublesome in adapting to P2P simulation because of the complexity and interdependency between simulation modules (scheduler, core simulator models, protocol models, and application level models). These disadvantages cause difficulties in attempt to extend the functionality of simulator with new models.

Many research teams have created their own overlay simulators. Some of them are used by experimenters, including PeerSim [11], P2PSim [12], OverSim, TOSim (Trust Overlay Simulator) [13]. Some of them are for only specific purposes and, thus, they are not efficient for general P2P protocol evaluation. NeuroGrid Simulator [14] is focused on simulating searches over content distribution network. Query-Cycle Simulator [15] is a cycle-based simulation framework for file-sharing P2P network. However, CANSimulator [16], FreeNet Simulator [17] may support network protocols, including BT protocol, either.

 ISBN: 978-1-61208-139-7

There is an existing implementation of BT in GPS (General P2P Simulator) environment [18]. Moreover, validation of this BT model was made by comparison it to a small scale network, also performance of the simulator is not so good, even the current version [19]. The other known implementation of BT protocol is in OMNET++ simulator [9]. This BT model was composed from three basic modules, including Tracker, Tracker Client, and Peer-wire. These modules correspond to the principles of BT actions.

The performances of many various P2P file sharing systems and BT system have been modeled in various ways. Qiu and Srikant [6] constructed a simple fluid model based on the Markov chains [20] describing the dynamics of the BT system. In [21], a statistical mathematical model is presented, which describes the evolution of BitTorrent.

In [22], Jun and Ahamad discussed the properties of the incentive mechanism of BT. Their analysis, based on the experimental results, showed that the original incentive mechanism of BT can induce free riding. They proposed a game theoretic framework that is more robust against free riders than the original mechanism.

Also several analytical studies of BT incentive mechanisms are presented in [23][24][25]. It was shown in [23] that BT mechanisms cannot prevent a systematic fairness through a set of simulations. Tian, Wu, and Ng [25] found that the standard tit-for-tat strategy cannot improve file availability. They proposed an innovative tit-for-tat strategy.

This brief review might show that various aspects of P2P networks and properties of BT protocol have been just discussed, analyzed and described in literature. The objective of this paper is checking the impact of some mechanisms on efficiency of BT, and making an attempt in finding some improvement of tit-for-tat mechanism.

III. SIMULATION ENVIRONMENT

The implementation of the BT protocol in OMNeT++ environment was based upon the mainline client version 4.0.2 presented in [26]. The mainline client is considered as the reference implementation of the BT protocol. The choking and rarest first algorithms are implemented just as they have been presented in [27].

The algorithms associated with BT, e.g., the peer selection and the piece selection can be implemented in different ways. Creating simulation environment, we take into consideration the following assumptions:

- All messages are responded immediately. Processing time for a message is zero (in simulation time), except for piece requests, which have a configurable response delay.
- Leecher starts downloading from another peer at the moment it is un-choked by that specific peer.
- A new block is requested immediately after a block has arrived from a peer, provided that the client is not choked in the meantime by that peer.
- Handshakes and bit-fields are exchanged without processing delay.
- Response delay is used only when handling piece requests.
- Piece selection algorithm can be configured from the default *.ini file.
- Swarm sizes are easy configurable by changing only one parameter in the default *.ini file.
- BitTorrent clients are created dynamically upon start-up of the simulation.
- Maximum swarm size of a simulation is not explicitly defined - it can be altered by configuring two parameters: the amount of clients connecting during the simulation run and the session inter-arrival time.

Many of the parameters of a BitTorrent swarm have not been previously considered that is why some assumptions regarding the input distributions of these parameters were necessary, including the exponential seeding time which was taken as proposed in [19].

IV. EXPERIMENTAL SETUP

The main part of experiment design – parameters used for the simulation scenarios - is presented in Table I.

TABLE I. SIMULATION PARAMETERS FOR SCENARIOS

Parameter	Value
Number of seeds	1
Number of peers	200
Max number of connections	200
Number of peers requested	50
Session inter-arrival time	Exponential, μ = 12,395s
Link delay	Uniform, [50, 400]ms
Bandwidth throughput	Asymmetric - 4Mbps/1Mbps
Asymmetric link	Yes
Initial piece distribution	0%
Request waiting time	Exponential, μ = 100ms
Block size	215 (32768)bytes
Simulation runs	20

TABLE II. PARAMETERS OF TORRENT FILES

Parameter	Value
Number of pieces	1205
Piece size	62144 bytes
Download piece size	74825472 bytes

Choosing the number of peers equal to 200 and the number of repeating simulations per scenario equal to 20 represent an acceptable trade-off between simulation time and the number of resulting data to be obtained. The reason of taking a single seed is giving opportunities for starting to upload a new content in the network. The maximum number of connections equal to 200 is the sum of 199 peer connections and the single tracker connection. The value of the swarm inter-arrival time was selected at

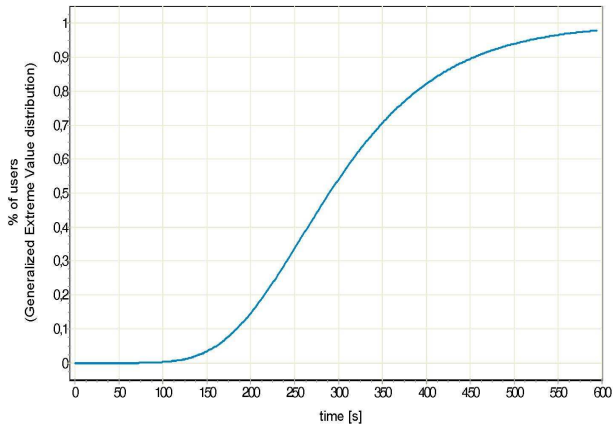
 ISBN: 978-1-61208-139-7

random from the results reported in [28]. The distribution parameters for link delay and bandwidth were selected to cover a large spread of possible link types. However, it may be observed that with this chosen data the session and message dynamics correspond to real-world environment. The initial piece distribution of each peer was chosen to be uniformly distributed. The parameter represents the ratio of pieces available at a joining peer when it enters the swarm (the initial seed always has all pieces). Which specific pieces are available is selected randomly. This was done to reflect the fact that the measurements reported in [28] were performed on swarms that already contained active peers already in possession of pieces of the content. Request waiting time, generated by different delay created processes, was taken as of exponential distribution with mean value equal to 100 [ms]. The block size of 32768 bytes was selected, following default size recommended in [26]. Remark: The experiment design does not take into consideration any information below the application layer, such as host names, IP addresses, or port numbers, thus, meta data required to join a swarm were proposed (Table II) for characteristic of the torrent file.

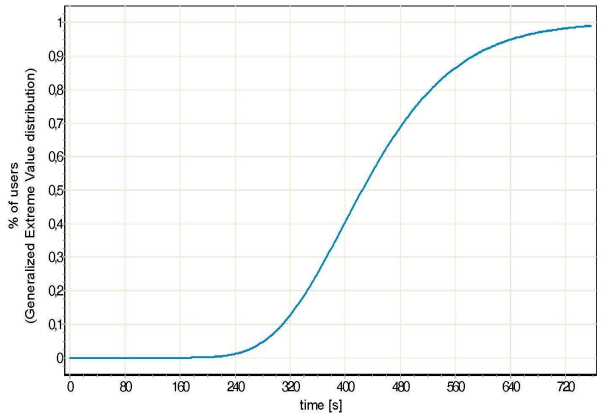
V. INVESTIGATIONS

A. *Link delay*

It is well-known that the throughput in TCP protocol depends strictly on the RTT - the elapsed time for transit of a signal over a closed circuit, or time elapsed for a message. Thus, if two TCP flows compete for the resources of the same bottleneck link, the connection with a smaller RTT can receive a higher bandwidth share than the other. Since a BT peer uploads to those peers from whom it downloads with high rates, peers on links with large delays may be characterized with the worse performance. To confirm this thesis we made the simulation experiment. The cumulative distribution functions for 20 runs are shown in Figure 1.



(a)



(b)

Figure 1. CDF for download time: (a) 10[ms] delay, (b) 100[ms] delay.

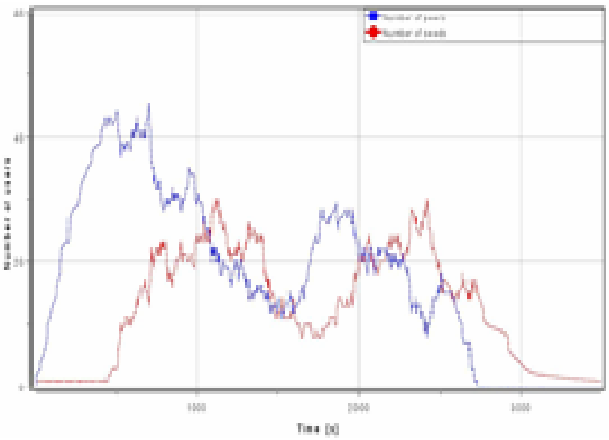
It may be observed that the download time for the case with 10 [ms] delay (Figure 1a) is of 290 [sec] and is less than 430 [sec] in the case with 100 [ms] delay (Figure 1b). Therefore, the mean download performance deteriorates by 33% for peers with greater delays.

It may be also observed, that the seeds and peers behaviors (see Figure 2) are different in relation to the delay. For delay of 10 [ms], the number of peers is changing in dynamical way. This is the result of the fast message exchange between users. In seed case two plots are almost identical and no major differences are observed.

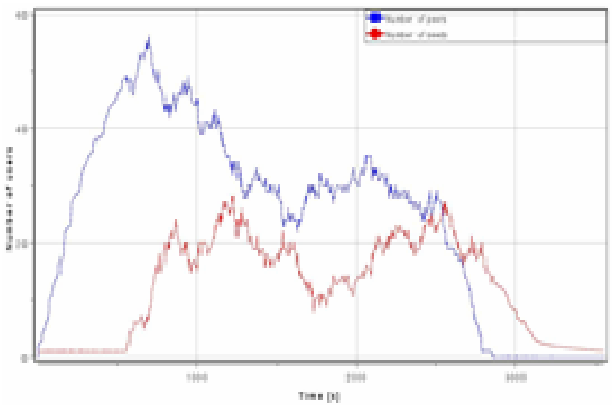
B. *Modified default tit-for-tat mechanism*

The second scenario provides some minor changes in default tit-for-tat mechanism, exactly in optimistic un-choking. The default un-choking times in optimistic un-choking mechanism is 30 [sec].

In order to check the relation of optimistic un-choking time to other parameters, the value of optimistic un-choking time was lowered to 10 [sec].



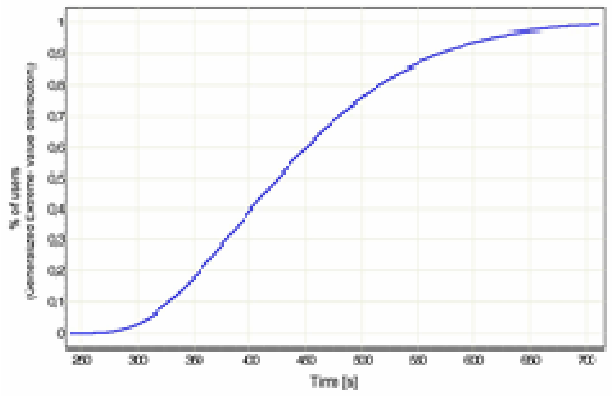
(a)



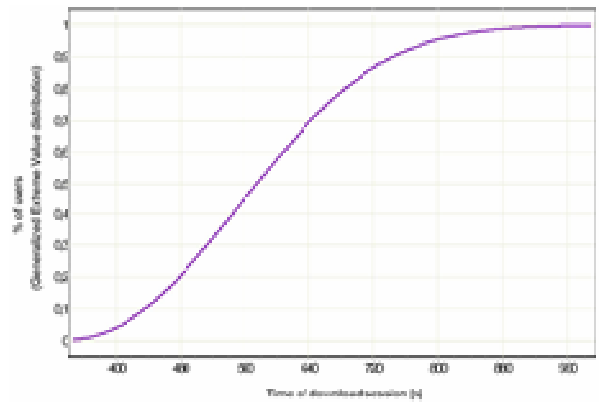
(b)

Figure 2. Seed vs peer session: (a) 10[ms] delay, (b) 100[ms] delay.

This parameter was not configurable from the default *.ini file in BIT-SIM simulator (some minor interference in source code was necessary).



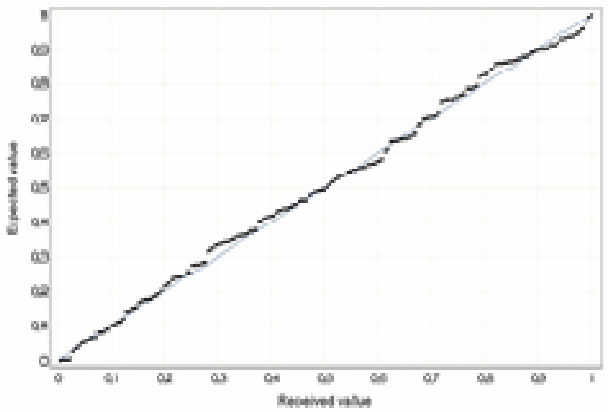
(a)



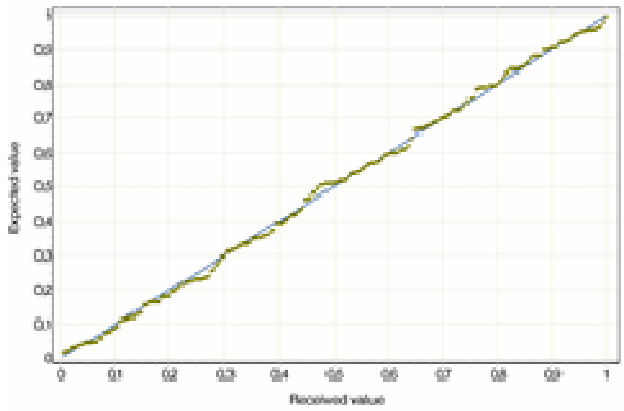
(b)

Figure 3. CDF for download time: (a) with tit-for-tat change, (b) without tit-for-tat change.

Measurement results are presented in Figure 3 and Figure 4. When we compare the plots from Figure 3(a) and 3(b), we can observe some minor improvement. The mean value of 50% for Figure 3(b) is equal to 580[sec] in turn in the optimistic un-choking mechanism with 10 [sec] un-choking times.



(a)



(b)

Figure 4. P-P plot for share ratio: (a) with tit-for-tat change, (b) without tit-for-tat change.

In Figure 3(a), the considered value is equal to 425 [sec], resulting in profit (improvement) of 27%. It may be observed that the share ratio value shown in Figure 4 is almost the same in two plots presented in 4(a) and in 4(b). Thanks to that, the change does not affect tit-for-tat in overall negative way.

C. Overall conclusion

In Table III, the main results of experiments have been gathered. In the table, in the fourth column, the 'change' is referring to percentage gain which resulted after changing the values of the considered parameters: *link delay* and *optimistic un-choking time* from the 'old' one to the 'new' one.

 ISBN: 978-1-61208-139-7

TABLE III. MAIN RESULTS OF EXPERIMENT

Parameter	Value		Change
	Old	New	
Link delay	10 [ms]	100 [ms]	33%
Optimistic unchoking time	30 [s]	10 [s]	27%

VI. FINAL REMARKS

In this paper, we consider the two theses formulated in Section I, concerned performance and behavior of BitTorrent protocol according to (i) its internal tit-for-tat mechanism changes and (ii) its external one like link delay (independent of protocol).

In the case (i), we confirmed that even small appropriate changes in the internal protocol (like change of optimistic un-choking time) can ensure more efficient data exchange. That is why some further investigations should be made with adjusted parameters inside BitTorrent mechanism configurations and using the idea of multistage experiments as proposed in [30].

In the case (ii) concerning link delay, we have shown that TCP protocol (which is used in BitTorrent) is not a most efficient which can be used. The link delay fluctuations can cause a visible protocol performance decrease. Improvement can be made by possible using the UDP protocol - to which the BitTorrent is migrating right now. The prototype of that can be treated as a kind of protocol already created in the µTorrent client and from the name of the client can be named µTP. Although all the BitTorrent protocol codes are open to the Internet community, the µTP protocol is closed for now.

It also should be mentioned that the simulator used in this paper, may be regarded as a useful tool for conducting experiments, however, in limited range because there are some functionalities not available such as modular peer selection and peer snubbing, (i.e., dropping peers that do not respond quickly enough), trackerless Distributed Hash Table (DHT) protocol, encryption and super seeding. Moreover, torrent file used in simulation does not use additional information like current download status, connected peers and QoS information. The investigations in the nearer future should take into consideration these aspects to allow making detailed analysis of BitTorrent efficiency.

REFERENCES

[1] B. Cohen, ”Incentives build robustness in BitTorrent”, Proceedings to 1st Workshop on Economics of Peer-to-Peer Systems, 2003.

[2] B. Cohen, ”A new P2P application”, Yahoo e-Groups, Posted February 2001, Retrieved 2007

[3] TorrentFreak. http://torrentfreak.com/bittorrent-still-king-of-p2p-traffic- 090218, February 2009.

[4] R. Thommes and M. Coates, “Deterministic packet marking for time-varying congestion price estimation”, IEEE/ACM Trans. Networking, vol. 14, 2006, pp. 592-602.

[5] D. Guan, J. Wang, Y. Zhang, and J. Dong, “Understanding BitTorrent download performance”, Proceedings to Int. Conf. on Networking, ICN’08, pp. 330-335.

[6] D. Qiu and R. Srikant, “ Modeling and performance analysis of BitTorrent like peer-to-peer networks”, Proceedings to IEEE SIGCOMM ‘04, doi:10.1145/1030194.1015508, USA, 2004.

[7] M. Izal, G. Urvoy-Keller, E. W. Biersack, P. Felber, A Al Hamra, and L. Garces-Erice, “Dissecting BitTorrent: five months in a Torrent's lifetime**”,** , PAM 2004, Lecture Notes in Computer Science, vol. 3015, Springer 2004, pp. 1-11.

[8] K. De Vogeleer, D. Erman, and A. Popescu. ”Simulating BitTorrent”, Proceedings to International Conference SimuTools’08, Marseille, France, ISBN 978-963-9799-20-2, March 2008.

[9] BIT-SIM simulator. http://www.its.bth.se/staff/kdv/BIT-SIM/, June 2009.

[10] Berkeley/LNBL/ISI, “The ns-2 simulator”, http://www.isi.edu/nsnam/ns/, 2007.

[11] A. Montresor and M. Jelasity, “PeerSim: A scalable P2P simulator”, Peer-to-Peer Computing, 2009, pp. 99-100.

[12] P2Psim. http://pdos.csail.mit.edu/p2psim/, 2009.

[13] W. Wang and G. Zeng,”A Generic trust overlay simulator for P2P networks”, Proceedings to PRDC’06, December 2006, pp. 401-402.

[14] NeuroGrid. http://www.neurogrid.net/, 2009.

[15] T. Condie, S. D. Kamvar, H. Garcia-Molina, “Adaptive peer-to-peer topologies”, Peer-to-Peer Computing, 2004, pp. 53-62.

[16] CANSimulator. http://sourceforge.net/projects/cansimulator.

[17] J. Pfeifer,”Freenet casching algorithms under high load”, http://www.cs.usask.ca/classes/498/t1/898/W7/P2/freenet, 2009.

[18] W. Yang and N. B. Abu-Ghazaleh, “GPS: A general Peer-to-Peer simulator and its use for modeling BitTorrent.” Proceedings to MASCOTS'2005, 2005, pp. 425- 434.

[19] GPS simulator. http://www.cs.binghamton.edu/, 2009.

[20] X. Yang and G. de Veciana, “Performance of peer-to-peer networks: service capacity and role of resources sharing policies”, Journal Performance Evaluation, vol. 63, March 2006.

[21] Z. Liu and C. Chen, “Modeling fetch-at-most-once behavior in peer-to-peer file-sharing systems. Proceedings to APWeb Workshop, 2006, pp. 717-724.

[22] S. Jun and M. Ahamad, “ Incentives in BitTorrent induce free riding”, Proceedings to ACM SIGCOMM Workshop on Economics of Peer-to-Peer Systems, USA, ACM Press, August 2005, pp. 116-121.

[23] A. R. Bharambe, C. Herley, and V. N. Padmanabhan, “Analyzing and improving BitTorrent performance”, Technical Report MSR-TR-2005-03, Microsoft Research, Redmond, WA, February 2005.

[24] K. Csernai, M. Jelasity, J. A. Pouwelse,, and T. Vinko, “:Modeling un-connectable peers in private BitTorrent communities”, Proceedings to PDP’11, 2011, pp. 582-589.

[25] Y. Tian, D. Wu, and K.-W. Ng, “Modeling, analysis and improvement for BitTorrent-like file sharing networks”, Proceedings IEEE INFOCOM 2006, Barcelona, Spain, 2006.

[26] BitTorrent, Inc. Bittorrent. http:www.bittorrent.com.BitTorrent, 2009.

[27] A. Legout, G. Urvoy-Keller, and P. Michiardi. ”Rarest first and choke algorithms are enough”, Proceedings to ACM SIGCOMM/USENIX ICM, October 2006.

[28] D. Erman, D. Saavedra, J. A. Sanchez-Gonzales, and A. Popescu, “Validating BitTorrent models”, Telecommunications Systems, vol. 39, 2008, pp. 103-116.

[29] S. Tewari, L. Kleinrock, “Optimal search performance in unstructured peer-to-peer networks with clustered demands”, IEEE Journal on Selected Areas in Communications, vol. 25, 2007, pp. 84-95.

[30] D. Ohia, L. Koszalka, and A. Kasprzak, “Evolutionary algorithm for congestion problem in computer networks”, Springer, Lecture Notes in Artificial Intelligence, vol. 5711, 2009, pp. 113-122.

 ISBN: 978-1-61208-139-7

An Autonomic Framework for Service Configuration

Patcharee Thongtra
Department of Telematics
Norwegian University of Science and Technology
N-7491 Trondheim, Norway
patt@item.ntnu.no

Finn Arve Aagesen
Department of Telematics
Norwegian University of Science and Technology
N-7491 Trondheim, Norway
finnarve@item.ntnu.no

***Abstract* — An autonomic framework for service configuration functionality is proposed. The framework has goals and policies. *Goals* express required performance and income measures. *Policies* define actions in states with unwanted performance and income measures. All functionality is executed by *autonomic elements* (AEs) that have ability to download and execute behavior specifications during run-time. An AE has several generic functionality components. Two important generic components of an AE are *Judge* and *Strategist*. A *Strategist* selects actions in a state with unwanted performance and income measures to reach a state defined by goal performance and income measures. A *Judge* gives rewards to actions based on the ability to move towards a state with goal performance and income measures. The *Strategist*'s selection of actions is based on the rewards given by the *Judge*. AE functionality is realized by the combination of Extended Finite State Machines (EFSM), a Reasoning Machine (RM) and a Learning Machine (LM). A case study of an adaptable streaming system is presented. Using the proposed model, the streaming system can select actions for capability allocation adaptation more appropriately as evaluated by the performance and income measure results.**

Keywords-Autonomic; Service configuration; Policy; Autonomic Elements.

I. INTRODUCTION

Networked service systems are considered. Services are realized by service components, which by their inter-working provide a service in the role of a service provider to service users [1]. Service components are executed as software components in nodes, which are physical processing units such as servers, routers, switches, PCs and mobile phones. A service framework is here defined as the overall structural and behavior framework for the specification and execution of services. *Service configuration* comprises capability configuration, capability allocation, service deployment and instantiation, system performance diagnosis, fault diagnosis and service adaptation. *A capability* is an inherent property of a node required as a basis to implement services [1]. Capabilities can be classified into resources, functions and data. Examples are CPU, memory, transmission capacity of connected transmission links, available special hardware, and available programs and data.

The service configuration is done with respect to required capabilities and capability performances as well as required service performances. In this paper, we focus on service configuration of *adaptable service systems*, here defined as a service system that can adapt by itself related to changes by users, nodes, capabilities, system performances and service functionalities.

In this paper, an *autonomic* approach to adaptable service systems is proposed. *Autonomic systems* have ability to manage themselves and to adapt dynamically to changes in accordance with given objectives [2, 3]. The autonomic system is *constituted* by distributed components denoted as *autonomic elements (AEs)*. An AE is the smallest entity that can manage its internal behaviors and relationships with other entities in accordance with its defined behavior. *A service component as already defined is realized by one AE*. An AE is constituted by several generic functionality components. Two important components are *Judge* and *Strategist*. The *Judge* and *Strategist* apply defined *goals* and *policies*. Goals express required performance and income measures. A policy is defined by conditions, constraints and actions, and defines accordingly actions to adapt the system in states with unwanted performance and income measures. The *Judge* gives rewards to actions based on the ability to move towards a state with goal performance and income measures. The *Strategist* selects actions based on the rewards given by the *Judge*.

The reasons behind the Autonomic Element model and the AE functionality components' specifications (see Section III) are service components based on the classical Extended Finite State Machine (EFSM) approach can provides the software update flexibility [4], and Reasoning Machine (RM) using the policies can add the ability to cope with various situations more flexible [5, 6].

This paper is organized as follows. Section II defines autonomic properties. Section III defines the main concepts of what is denoted as the *Goal-based Policy Ontology*. The details of the *Autonomic Element Model* are described in Section IV. Section V describes how AEs are used to realize service functionalities that are necessary for the service configuration. Section VI presents a case study, related works are presented in Section VII, and finally, summary and conclusions are presented in Section VIII.

II. PROPERTIES OF AUTONOMIC ELEMENTS

An autonomic system consists of a set of decentralized autonomic elements (AEs), as defined in Section I. AE functionality is realized by the combination of Extended Finite State Machines (EFSM), a Reasoning Machine (RM) and a Learning Machine (LM). An AE is a generic software

 ISBN: 978-1-61208-139-7

component that can dynamically download and execute EFSM, RM and LM specifications. Properties can be classified as *individual* AE properties and *shared* AE properties, i.e., properties of the AEs constituted by the cooperation of AEs. AEs have the following *individual* properties:

- *Automaticity*: An AE can manage its EFSM states, variables, actions and policies.
- *Awareness*: An AE is able to monitor its own EFSM states and performances.
- *Goal-driven:* An AE operates and/or controls its functions towards goals.

AEs have the following *shared* properties:

- *Automaticity*: AEs can manage capabilities and nodes, i.e., capabilities and nodes can be added and removed.
- *Adaptability*: The goals, policies and EFSM behavior of an AE can be changed.
- *Awareness*: An AE is able to monitor available nodes and capabilities. Information about EFSM states and performance measures can be made available to other AEs.
- *Mobility:* An AE can move to a new node and resume its operation.

III. GOAL-BASED POLICY ONTOLOGY

An ontology is a formal and explicit specification of a *shared* conceptualization [7] containing both objects and functions operating on instances of objects. We can define independent concepts and relational concepts defined by mathematical logics, e.g., if-then-else. In applications with reasoning capability, the logic concepts can be represented and processed flexibly as rules [1].

Figure 1 presents a simplified diagram of the concepts in the Goal-based Policy Ontology. At the top level we have *goal*, *policy* and *inherent state*, which all are related to service and capability as defined in Section I. The instantiated AEs have inherent states that can comprise *measures* related to functionality and performance of services and capabilities as well as income. *System performance* is defined as the sum of capability performance and service performance.

As a basis for the optimal adaptation, service level agreements (SLA) are needed between the service users and the service provider. An SLA class defines service functionalities, capabilities, QoS levels, prices and penalty. *Service income* includes the estimated income paid by the users for using services in normal QoS conditions and the penalty cost paid back to the users when the service qualities and functionalities are lower than defined by SLA. In general, goal, policy and inherent state concepts have the SLA class as a parameter.

The goal is defined by a *goal expression* and a *weight*. The goal *expression* defines a required system performance or service income measure. A goal example is: "*Service response time of premium service SLA class < 2 secs*". The goal *weight* identifies a goal's importance. A goal can be associated with a set of policies.

A policy is defined by conditions, constraints and actions. The condition defines the activation of the policy execution. The constraint restricts the usage of the policy, and is described by an expression of required and inherent functionality and performance of services and capabilities, required and inherent service incomes, available nodes and their capabilities, as well as system time. An action has an estimated operation cost and accumulated reward.

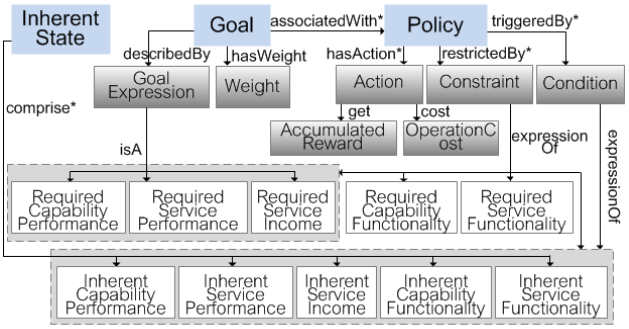


Figure 1. Goal-based Policy Ontology.

A policy example related to the goal example given above is: "*If CPU utilization > 95% and the time is between 18:00-24:00, ignore new service requests of users of ordinary SLA classes that request service time > 2 mins*". It is expressed with Conditions: CPU utilization > 95%, Constraints: system time between 18:00-24:00 and service time request > 2 mins, and Actions: ignore new service requests of users of ordinary SLA classes.

Table I lists notations used for capability, service and income concepts.

TABLE I. THE CAPABILITY, SERVICE AND INCOME CONCEPT NOTATION

$\hat{C}_R$	Required capability performance set
$\hat{C}_I$	Inherent capability performance set
$\overline{C_R}$	Required capability functionality set
$\overline{C_I}$	Inherent capability functionality set
$\hat{C}_{A,n}$	Set of available capabilities in node n; n=[1, N]
$\hat{S}_R$	Required service performance set
$\hat{S}_I$	Inherent service performance set
$\overline{S_R}$	Required service functionality set
$\overline{S_I}$	Inherent service functionality set
I_R	Required service income
I_I	Inherent service income

IV. AUTONOMIC ELEMENT MODEL

An AE is composed of four functional modules: i) *Main Function*, ii) *Strategist*, iii) *Judge* and iv) *Communicator*, as illustrated by Figure 2. The behaviors of the various modules are explained in the following subsections IV.A-IV.D. The life-cycle of an AE is described in Section IV.E.

 ISBN: 978-1-61208-139-7

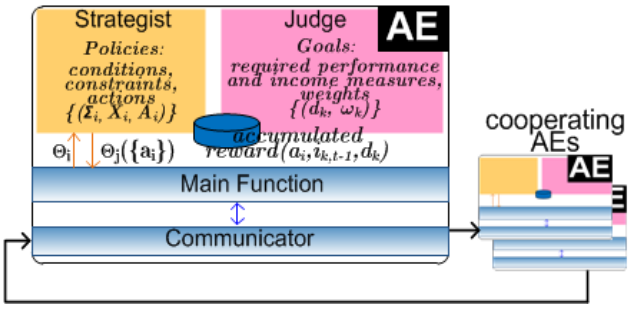


Figure 2. Autonomic Element Model.

A. Main Function

Main Function coordinates the functionality of an AE. An AE has some general behavior which is common for all AEs, and some behavior depending on the specific role of the AE (see Section V). In general, an AE will have requirements with respect to capabilities and capability functionalities and performances. The specific need depends on the specific functionality of the AE. The *Main Function* behavior is based on an *Extended Finite State Machine (EFSM)* model E defined (≡) as:

$$E \equiv \{ S_M, S_I, S_S, V, M, O, Q, F_S, F_O, F_V \} \quad (1)$$

where S_M is a set of all states, S_I is an initial state and S_S is a set of *stable* states. V is a set of variables including the inherent state variables. M is a set of input messages, O is a set of output messages and Q is a message input queue. F_S is a state transition function (F_S: S x M x V -> S), F_O is an output function (F_O: S x M x V -> O) and F_V is a set of actions performed during a specific state transition.

An AE can move to a new node. A *stable state* is a state of the *Main Function* where an AE's functionality can move safely and be re-instantiated in a new node based on the restoration of EFSM state, variables, and queued messages. *Strategist* is used by the *Main Function* to select appropriate actions. The *Main Function* will regularly

- Compare the condition part of the policies with inherent state variables, and will
- Activate the *Strategist* if a condition is met, which returns an action to be used by the *Main Function*

B. Strategist

Strategist selects appropriate actions to be used by the *Main Function*. The *Strategist* behavior is based on a *Reasoning Machine (RM)* model, extended from [5, 6]. It can be triggered by one condition at a time. It will execute all policies related to a condition. The RM model R is defined as:

$$R \equiv \{ \Theta, \Phi, \Pi, \xi \} \quad (2)$$

where Θ is a set of query expressions containing variables, Φ is a generic *reasoning procedure,* Π is a set of policies, and ξ is the strategist data including the inherent states from the *Main Function* and from other AEs, and available nodes and their capabilities.

$$\xi \equiv (\overline{S_I}, \hat{S}_I, \overline{C_I}, \hat{C}_I, I_I, \hat{C}_{A,n}; n=[1, N]) \quad (3)$$
$$\Pi \equiv \{ p_i \} \quad (4)$$
$$p_i \equiv (\Sigma_i, X_i, A_i) \quad (5)$$
$$\Sigma_i \equiv \text{Expression}(\overline{S_I}, \hat{S}_I, \overline{C_I}, \hat{C}_I, I_I) \quad (6)$$
$$X_i \equiv \text{Expression}(\overline{S_R}, \hat{S}_R, \overline{C_R}, \hat{C}_R, I_R, \overline{S_I}, \hat{S}_I, \overline{C_I}, \hat{C}_I, I_I, \hat{C}_{A,n}; n=[1, N], \Gamma) \quad (7)$$

A policy p_i has conditions Σ_i, constraints X_i and actions A_i. The condition is an expression of the inherent states from the *Main Function* and from other AEs. The constraint is an expression of required functionality and performance of services and capabilities, required service incomes, the inherent states from the *Main Function* and from other AEs, available nodes and their capabilities, as well as system time (Γ).

The reasoning procedure is applied to select appropriate actions with maximum accumulated rewards. It is based on *Equivalent transformation (ET)* [8], which solves a given problem by finding values for the variables of the queries. The conditions, constraints and actions can have variables. The result of the reasoning procedure can, in addition to actions, give instantiated variables.

C. Judge

Judge gives rewards to actions to be selected by the *Strategist*. The reward is a numeric value based on the ability to move towards a state with goal performance and income measures. The rewards will be accumulated over a period of time. The *Judge* behavior is based on a *Learning Machine (LM)* model L defined as:

$$L \equiv \{ \Omega, \Lambda, \Psi, \zeta \} \quad (8)$$

where Ω is a set of *goals*, Λ is a generic *rewarding procedure*, Ψ is a *reward database* storing the accumulated rewards of actions, and ζ is the judge data including the inherent states from the *Main Function* and from other AEs. We further have:

$$\zeta \equiv (\overline{S_I}, \hat{S}_I, \overline{C_I}, \hat{C}_I, I_I) \quad (9)$$
$$\Omega \equiv \{ g_k \} \quad (10)$$
$$g_k \equiv (d_k, \omega_k) \quad (11)$$

A goal g_k has goal expression d_k and weight ω_k. The sum of the goal weights is equal to 1. At time t, the rewarding procedure will calculate the reward of an action a_i, which was applied at time t-1 as:

$$\text{reward}(a_i, i_{k,t-1}, d_k) = (\Delta(i_{k,t}, i_{k,t-1}) / \Delta(d_k, i_{k,t-1})) * \omega_k - \text{cost}(a_i) \quad (12)$$

 ISBN: 978-1-61208-139-7

where $i_{k,t-1}$ and $i_{k,t}$ are an inherent state measure before and after applying the action for an monitoring interval [t-1, t], $i_k \in \zeta$ and d_k is an associated goal required measures. $\Delta(i_{k,t},i_{k,t-1})$ is the difference between $i_{k,t}$ and $i_{k,t-1}$. $\Delta(d_k,i_{k,t-1})$ is the difference between d_k and $i_{k,t-1}$. ω_k is the goal weight and $\mathrm{cost}(a_i)$ is the operation cost of a_i.

The accumulated reward of an action a_i, $\mathrm{accumulated\_reward}(a_i,i_{k,t-1},d_k)$, is then the sum of the rewards of a_i, for an inherent state measure $i_{k,t-1}$ and a goal required measure d_k.

D. Communicator

Communicator handles message sending and receiving on behalf of the *Main Function*. The *Communicator* behavior is based on the EFSM model in (1). Other AEs can subscribe to the inherent state variables of an AE. The *Communicator* will manage *subscription* messages and will send inherent state variables to other AEs on behalf of the *Main Function*.

The *Communicator* also handles the registration function on behalf of the *Main Function*. Registration message is sent to *Registry (REG)* (see Section V) that is an important AE that records the life-cycle state of AEs. The registration message contains IP address of the AE.

The *Communicator* will regularly broadcast *heartbeat message*, which is used to indicate that an AE is alive. The heartbeat messages are monitored by *Life Monitor AE (LMO)* (see Section V). In addition, the *Communicator* will inform REG about changes in the life-cycle state of the AE (see Section IV.E). REG will broadcast the changes to other AEs that subscribe to such updates.

E. Autonomic Element Life Cycle

The combined states of an AE during its life-cycle are defined follows:

- *Initial state*: An AE is instantiated in a node where there are capabilities and capability functionalities and performances as required.
- *Registering state*: An AE registers to REG.
- *Normal-Active state*: An AE provides services with normal functionality and QoS level.
- *Degraded-Active state*: If in the Normal-Active-State the capabilities are less than required and results in degraded functionality and QoS, the life-cycle state will change to Degraded-Active state. In this state, some actions selected by the *Strategist* can be taken to upgrade the capabilities and performances. From both the Normal-Active state and the Degraded-Active state, the life-cycle state can change to Moving, Suspended or Terminated.
- *Moving state*: An AE's functionality is being moved and re-instantiated in a new node. A move can only take place if the EFSM states are stable (see Section IV.A).
- *Suspended state*: An AE is suspended, i.e., by an action selected by the *Strategist*, which means that it stops executing current behavior specifications. An AE will release its allocated capabilities. At this state, an AE can start executing new behavior specifications. This makes an AE goes back to the Normal-Active state.
- *Terminated state*: Other AEs detect that an AE's heartbeat message is lost or an AE could not be reached because of some unintentional reasons, e.g., the hardware failure. REG is informed about this by LMO or the other AEs, then REG records that an AE is terminated.

V. AUTONOMIC ELEMENT-BASED SERVICE FUNCTIONALITY ARCHITECTURE

The service functionalities required for service configuration are constituted by *AEs* and repositories, as illustrated in Figure 3. The AE responsibilities and the repositories are described below.

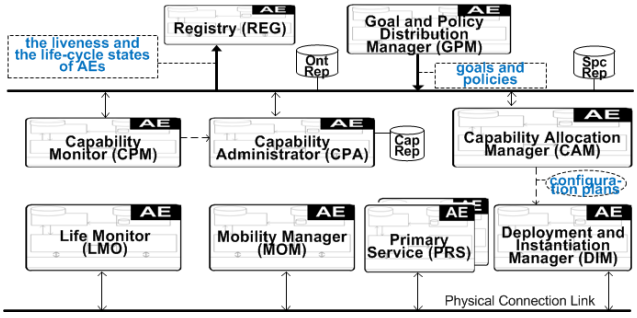


Figure 3. Autonomic Element-Based Service Functionality Architecture. Solid arrows indicate the physical connections of AEs and dashed arrows represent the message flows between AEs.

- *Primary Service (PRS)* provides ordinary user services.
- *Registry (REG)*, as already mentioned in Section IV.D, is responsible for AE registration.
- *Goal and Policy Distribution Manager (GPM)* distributes goals and policies to corresponding AEs.
- *Life Monitor (LMO)* observes the liveness of AEs by listening to heartbeat messages from AEs. LMO regularly updates the liveness of AEs to REG.
- *Capability Administrator (CPA)* maintains and provides data about capabilities and their functionalities and performances in available nodes.
- *Capability Monitor (CPM)* monitors capabilities and sends updates to CPA.
- *Capability Allocation Manager (CAM)* generates (re-) configuration plans for AEs to be instantiated in nodes. CAM fetches the capability requirements and retrieves the capabilities from CPA. A configuration plan defines in which node an AE should execute. Configuration plans are generated based on capability requirements and policies. In addition, CAM allocates capabilities to AEs. The allocation depends on the capability structure

 ISBN: 978-1-61208-139-7

and optimization criteria which can be specified in the policies.

- *Deployment and Instantiation Manager* (*DIM*) executes the configuration plan. It creates AEs in the defined nodes and assigns the behavior specifications.
- *Mobility Manager (MOM)* supports an AE when it is moved and re-instantiated in a new node. The move can be related to failures or insufficient capability performances. MOM broadcasts messages to inform other AEs when an AE's functionality is suspended or is resumed. MOM also handles an AE's connections by getting input messages on behalf of an AE and forwarding them to such AE when it is already re-instantiated.
- *Ontology Repository (OntRep)* stores the goal and policy concepts as well as the related capability and service concepts.
- *Service Specification Repository (SpcRep)* stores the AE behavior specifications and the capability requirements.
- *Capability Repository (CapRep)* stores data about available nodes and their capabilities.

VI. CASE STUDY

A music video streaming system is presented with the intention to demonstrate the *Strategist* and *Judge* solution in the proposed autonomic framework. The system is constituted by the AEs as defined in Section V. The Primary Service AEs are *Streaming Manager (STM)* and *Streaming Client (STC)*. An STM, executing on a *media streaming server (MS)*, streams the music video files to STCs. An STC is associated with an SLA class, which defines *required streaming throughput*, *price for the service* and *service provider penalties* if the agreed QoS cannot be met. Two SLA classes are applied: *premium (P)* and *ordinary (O)*. An STC is denoted by its SLA class as STC_P or STC_O. Each SLA class has different required throughput (X); the STC_P required throughput (X_P) can be 1Mbps or 600Kbps for high-resolution and degraded fair-resolution videos, while the STC_O required throughput (X_O) is 500Kbps for low-resolution videos. Prices and penalties will be defined later.

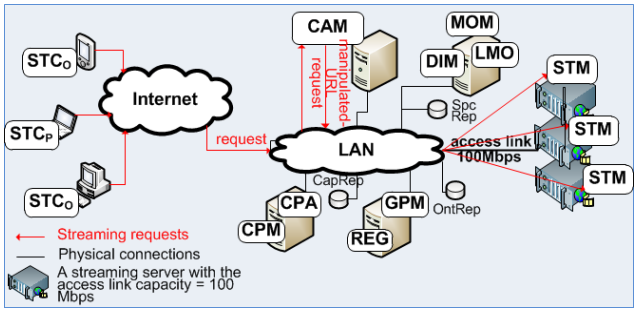


Figure 4. Streaming system example.

Figure 4 illustrates the streaming system. In this case study, CAM will accept the streaming requests on behalf of STMs. CAM will decide which an STM can serve the requests, or CAM may put them in waiting queues. CAM can also instantiate a new STM in an available MS that there is no executing STM.

The MS's required access link capacity ($C_{R,AL}$) is set to 100 Mbps. The number of STCs that can use the service at a time is limited by the MS access link capacity. When the required streaming throughput cannot be provided, a STC needs to wait until some connected requests have finished using the service. An STC_O can be disconnected, while an STC_P may have to degrade the video resolution. The service provider will pay penalties in case of waiting and disconnection of the STC. These penalty and price functions are given in Table II. A cost unit is the price paid by an ordinary client for one second streaming of the rate 500Kpbs. The price function for using the service is M(SLA_Class,X) (cost units/second). The penalty function for waiting is P_{WAIT}(SLA_Class) (cost units/second), and the penalty function for disconnection is P_{DISC}(SLA_Class) (cost units/connection).

Note that, the case study and all values, set in Table II, are same as our previous work [5, 6] in order to compare between the proposed and the previous model. The comparison results are in subsection VI.B.1.

TABLE II. THE PRICE AND PENALTY FUNCTIONS

	STC_O (X_O=500Kbps)	**STC_P (X_P=600Kbps)**	**STC_P (X_P=1Mbps)**
M(SLA_Class,X)/s	1	1.875	2
P_{WAIT}(SLA_Class)/s	5	10	10
P_{DISC}(SLA_Class)/ Connection	10	-	-

The complete set of actions A in this case study is:

$$A = \{a_D, a_B, a_N, a_I, a_R, a_T, a_M\} \quad (13)$$

A subset of A, Á, is defined as: $Á = A - \{a_M\}$. a_D is to disconnect the ordinary clients, a_B is to decrease the throughput of the premium clients, a_N is to instantiate a MS, a_I is to instantiate a new STM, a_R is to disconnect a MS, a_T is to terminate an STM and a_M is to move connected client sessions from an STM to another STM. These actions are selected by the *Strategist* of CAM. CAM executes a_N, a_I, a_R and a_T, while CAM suggests a_D, a_B and a_M to STMs.

In this case study, the considered capability is the MS access link. The required and inherent capability performance sets are denoted as $\hat{C}_R \equiv \{C_{R,AL}\}$ and $\hat{C}_I \equiv \{C_{I,AL}\}$, where $C_{R,AL}$ is the required access link capacity, and

 ISBN: 978-1-61208-139-7

$C_{I,AL}$ is the available access link capacity. The inherent service performance set $\hat{S}_I$ consists of the number of connected and waiting premium and ordinary clients ($N_{Con,P}$, $N_{Con,O}$, $N_{Wait,P}$, $N_{Wait,O}$), the number of disconnected ordinary clients ($N_{Disc,O}$), the number of MS (N_{Node}), the service time and waiting time of premium and ordinary clients ($T_{Serv,P}$, $T_{Serv,O}$, $T_{Wait,P}$, $T_{Wait,O}$). These values as well as the inherent service income (I_I) are observed per a monitoring interval Δ. The service income is defined as:

$$I_I = M(STC_O,X_O)*T_{Serv,O} + M(STC_P,X_P)*T_{Serv,P} - P_{WAIT}(STC_O)*T_{Wait,O} - P_{WAIT}(STC_P)*T_{Wait,P} - P_{DISC}(STC_O)*N_{Disc,O} - P_{Ser}*N_{Node}*\Delta \quad (14)$$

where P_{Ser} is the cost function for adding a new MS which is 150 units/second per node, while M(SLA_Class,X), P_{WAIT}(SLA_Class) and P_{DISC}(SLA_Class) are as already defined in Table II.

A. RM and LM Specification

In this case study, CAM plays an important role. Its RM specification is defined as follows:

$$R_{CAM} \equiv \{ \Theta_{CAM}, \Phi, \Pi_{CAM}, \xi_{CAM} \} \quad (15)$$

Π_{CAM} consists of five policies (p_1-p_5) as presented in Appendix. A policy defines some actions in the set A in (13).

The LM specification of CAM is defined as follows:

$$L_{CAM} \equiv \{ \Omega_{CAM}, \Lambda, \Psi_{CAM}, \zeta_{CAM} \} \quad (16)$$
$$\Omega_{CAM} \equiv \{ g_1, g_2 \} \quad (17)$$
$$g_1 \equiv (d_1: I_R > 0, \omega_1: 0.8) \quad (18)$$
$$g_2 \equiv (d_2: T_{Wait} < \Delta, \omega_2: 0.2) \quad (19)$$

where I_R is the required service income, and T_{Wait} is the sum of the waiting time of premium and ordinary clients. These goals are set in order to gain high income and to avoid high waiting time. The policy p_1-p_5 can be used when the required service income is not met, while the policy p_1-p_3 are used when the waiting time is higher than expected.

B. Experiments and Results

Two set of experiments are presented. In B.1) Á is used for the comparison between the proposed and a previous model. In B.2) different action sets A and Á are used to study the proposed model. The accumulated service income and the accumulated waiting time results are illustrated in both experiment sets.

The request arrivals are modeled as a Poisson process with an arrival intensity parameter λ_{SLA_Class}. The duration of streaming connections d_{SLA_Class} is constant and is set to 10 minutes. The traffic per MS access link ρ is defined as:

$$\rho = ((\lambda_P*d_P*X_P) + (\lambda_O*d_O*X_O))/ (N_{Node}*C_{I,AL}) \quad (20)$$

The monitoring interval Δ is 1 minute. The STCs will stop waiting and there is no penalty for waiting after 10 minutes. The number of available MS = 3. Initially, only one STM is instantiated.

1) Comparison between the proposed and a previous model

Our previous work [5, 6] presented an adaptation mechanism executed by a Reasoning Machine, which uses policies and goal to manage the adaptable systems. In the previous model, a policy consists of constraints and actions, and it is not associated with any specific conditions. So all defined policies will be executed when the systems are entering a reasoning condition. There is only one reasoning condition defined, i.e., the number of waiting clients > 0. In addition, only one goal based on the service income is used. The action is then rewarded by *the goodness score (QoX)* that is calculated by the percentage of the increased or decreased service income.

In this section, three different cases of ρ are illustrated: *a) ρ<1*, *b) ρ=1* and *c) ρ=1.15 (ρ>1)*. The STC_P request arrivals intensity (λ_P) is set to 25%, 50% and 80% of the total arrival intensity.

For case *a) ρ<1*, both models have the same behaviors. If ρ>0.5, they used $\{a_N, a_I\}$ to instantiate a MS and to instantiate a new STM, otherwise they just disconnected the ordinary clients or decreased the throughput of the premium clients. The accumulated service income and waiting time of both models are almost the same.

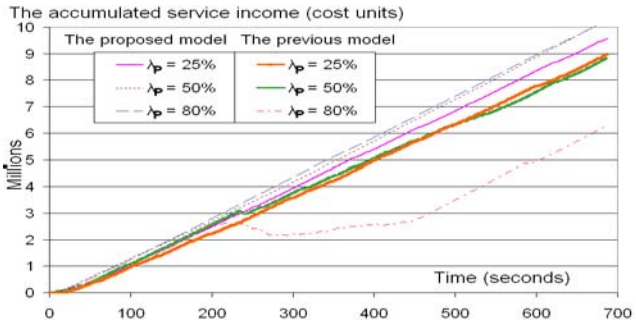


Figure 5. The accumulated service income when ρ = 1.

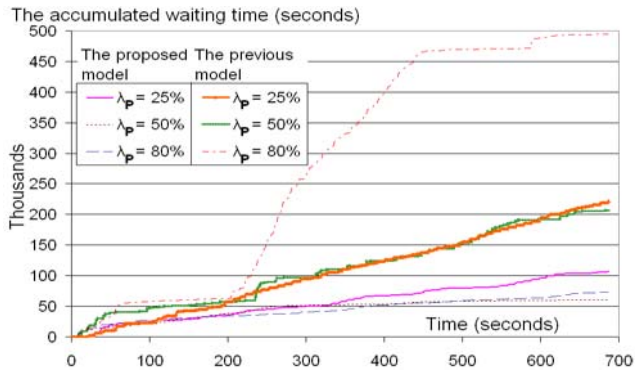


Figure 6. The accumulated waiting time when ρ = 1.

For case *b) ρ=1,* the proposed model produced higher accumulated service income and lower accumulated waiting time independent of λ_P, as depicted in Figure 5 and 6. This is because in the previous model $\{a_T, a_R\}$, which terminate an STM and disconnect a MS consecutively, was used when the number of waiting clients was little more than zero. So that, the number of MS was lower than proper required amount. It results in decreasing service income and increasing waiting time. In the proposed model, $\{a_T, a_R\}$ might be used only if the inherent service income <= 0; however, it does not happen when ρ = 1.

For case *c) ρ=1.15 (ρ>1)*, the proposed model also produced higher accumulated service income and lower accumulated waiting time. Figure 7 and 8 illustrates the accumulated service income and the accumulated waiting time for this case. The proposed model produced better results, because it took the actions to adapt the system more often than the previous model. When ρ>1, the service income could be less than 0 because of the waiting penalty. So that, in the proposed model the actions were applied both when the service income < 0 and when the waiting time > Δ, while the previous model the actions were applied only when the number of waiting clients > 0. However, when λ_P = 80% the traffic was too overloaded, and the accumulated service income was less than zero in both models.

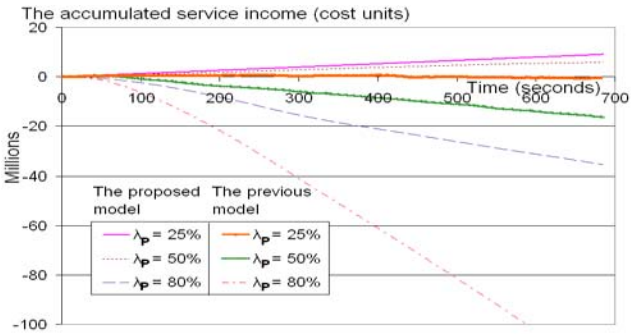


Figure 7. The accumulated service income when ρ = 1.15.

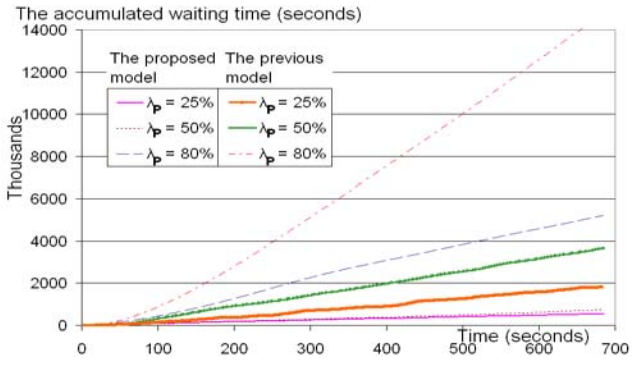


Figure 8. The accumulated waiting time when ρ = 1.15.

2) Comparison between different action sets

In this section, we compare three cases (I-III) of the new proposed model. In Case I the complete set of actions A is used, while in Case II the subset Á is used. For the last case, the complete set of actions A is also used, but there is no *Judge* component in the AEs so the actions are not rewarded. The traffics that were simulated for this scenario are relative to a function of time. The time with ρ at a fixed level, denoted as the *ρ period*, is set to 30 minutes. ρ varies from 0.2 to 1.2. λ_P is set to 50% of the total arrival intensity.

Figure 9 and 10 shows the accumulated service income and the accumulated waiting time of three cases. The brown line in these figures shows the variation of ρ. In Case I, the system learned that $\{a_M, a_T$ and $a_R\}$, which move connected STC sessions, terminate an STM and disconnect a MS consecutively, is efficient to adapt the system when ρ drops and then the required service income is not met. As a result, Case I could produce the highest accumulated service income and the lowest accumulated waiting time. For the last case, the actions were selected randomly and they were not appropriate to the states of unwanted service income and the waiting time. So, the accumulated service income of Case III was the lowest, while the accumulated waiting time was the highest.

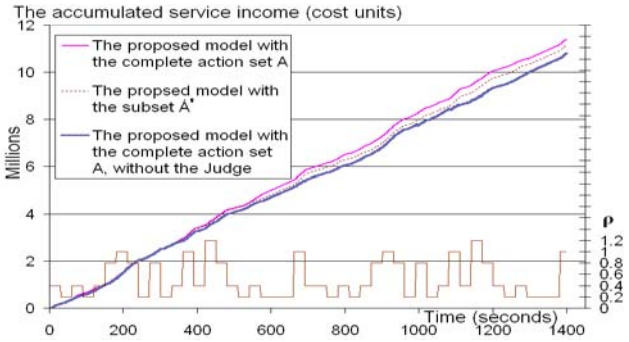


Figure 9. The accumulated service income for various ρ.

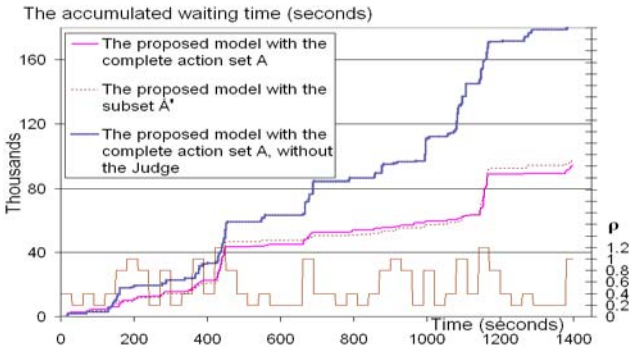


Figure 10. The accumulated waiting time for various ρ.

 ISBN: 978-1-61208-139-7

VII. Related Work

Existing service system frameworks that support run-time self-management and adaptation can be classified based on the way in which the management and adaptation functionalities are specified. The functionalities can be *statically* or *dynamically* specified. Some works propose to use templates [9] or adaptation classes [10] to statically specify these functionalities. However, the static approach lacks flexibility. All the possible adaptation must be known a beginning, and if new adaptations are required, the systems must be re-complied. Our work expresses the service management functionality for the adaptable service systems in the form of the EFSM, RM and LM specification, to be dynamically modified, added and removed at run-time. When using the EFSM specification, an *update* of changes is done by deployment of the whole specification. When using the RM and LM specification, only incremental changes of the *policies* and *goals* are deployed. However, the complete policy and goal based functionality need to be validated off-line before the deployment of the incremental changes.

There are several works that use the policies to specify the adaptation, such as [11], [5, 6], [12-15]. Accord [11] is a framework that can formulate autonomic applications as dynamic composition of AEs, with the use of policies to describe the adaptation of functional behaviors of AEs and interactions between them. However, our approach and the rest go beyond the use of policy for the specification by adding mechanisms to adapt policies or the way of using policies. Such policy adaptation can be grouped into three categories: 1) changing the policy parameters, considered in [5, 6, 12, 13]; 2) enabling/disabling a policy, found in [5, 6, 12]; 3) using techniques to select the most suitable policy and action; for instance, rewarding policies and their actions, presented in [14, 15].

Our approach is an instance of the first as well as the third category as [14, 15]. Tesauro et al. [14] presented a hybrid reinforcement technique used for resource allocation in multi-application data centers. This technique is to select optimal policies that can maximize rewards. Mesnier et al. [15] used decision trees to select accurate policies in storage systems. These policy adaptation techniques have only been applied to a single element, while our approach is potentially used in multi-autonomic elements.

VIII. Conclusions

This paper proposed an autonomic framework for adaptable service systems. The framework solution consists of *Goal-based Policy Ontology* and *Autonomic Element (AE) Model*. The Ontology defines common concepts of *goal*, *policy* and *inherent state*. AEs are generic component that can be used to realize any functionality. An AE is constituted by *Main Function*, *Strategist*, *Judge* and *Communicator* modules. The functionality of an AE is realized by two Extended Finite State Machines (EFSM), one Reasoning Machine (RM) and one Learning Machine (LM). EFSM behavior, as well as goals and policies can be modified flexibly during run-time. For attaining a specific functionality, specific EFSM, RM and LM functionality must be defined. In this paper, specific AEs handling service management functionality is proposed.

A case study is presented with focus on the Capability Allocation Manager (CAM). The experimental results show that the proposed model can produce higher service income and less waiting time than a previous model. In the proposed model, the actions are used appropriately under the associated goals and required goal measures. Moreover, it is possible to apply several goals, which each are weighed differently, depending on its importance. New actions can be added, and when there are more actions the system may reach the goals quicker.

References

[1] P. Thongtra and F. A. Aagesen. Capability Ontology in Adaptable Service System Framework. In Proc. of 5th Int. Multi-Conference on Computing in the Global Information Technology, Spain, Sep 2010.

[2] J. O. Kephart and D. M. Chess. The Vision of Autonomic Computing. IEEE Computer Society, January 2003, pp. 41-47.

[3] S. White, J. Hanson, I. Whalley, D. Chess, and J. Kephart. An architectural approach to autonomic computing. In Proc. of 1st IEEE Int. Conf. on autonomic computing, New York, May 2004, pp. 2–9.

[4] P. Thongtra and F. A. Aagesen. An Adaptable Capability Monitoring System. In Proc. of 6th Int. Conference on Networking and Services (ICNS 2010), Mexico, March, 2010.

[5] P. Supadulchai and F. A. Aagesen. Policy-based Adaptable Service Systems Architecture. In Proc. of 21st IEEE Int. Conf. on Advanced Information Networking and Applications (AINA'07), Canada, 2007.

[6] P. Supadulchai, F. A. Aagesen and P. Thongtra. Towards Policy-Supported Adaptable Service Systems. EUNICE 13th EUNICE Open European Summer School and IFIP TC6.6 Workshop on Dependable and Adaptable Networks and Services. Lecture Notes in Computer Science (LCNS) 4606, pp 128-140.

[7] R. Studer, V. R. Benjamins, and D. Fensel. Knowledge Engineering: princicples and methods. Data & Knowledge Engineering, vol. 25, pp. 161-197, 1998.

[8] K. Akama, T. Shimitsu, and E. Miyamoto. Solving Problems by Equivalent Transformation of Declarative Programs. In Journal of the Japanese Society of Artificial Intelligence, vol. 13, pp. 944-952, 1998.

[9] F. Berman, R. Wolski, H. Casanova, et al. Adaptive computing on the grid using AppLeS. In IEEE Trans. Parallel Distrib. Syst., vol. 14, no. 4, pp. 369–382, Apr. 2003.

[10] P. Boinot, R. Marlet, J. Noy´e, G. Muller, and C. Cosell. A declarative approach for designing and developing adaptive components. In Proc. of the 15th IEEE Int. Conf. on Automated Software Engineering, 2000.

[11] H. Liu and M. Parashar. Accord: a programming framework for autonomic applications. In IEEE Trans. on System, Man, and Cybernetics, vol. 36, pp. 341–352, 2006.

[12] L. Lymberopoulos, E.C. Lupu and M.S. Sloman. An Adaptive Policy-Based Framework for Network Services Management. In Journal of Networks and Systems Management, vol. 11, pp. 277–303, 2003.

[13] K. Yoshihara, M. Isomura, and H. Horiuchi. Distributed Policy-based Management Enabling Policy Adaptation on Monitoring using Active Network Technology. In Proc. of 12th IFIP/IEEE Int. Workshop on Distributed Systems: Operations and Management, France, Oct 2001.

 ISBN: 978-1-61208-139-7

[14] G. Tesauro, R. Das, N.K. Jong, and M.N. Bennani. A Hybrid Reinforcement Learning Approach to Autonomic Resource Allocation. In Proc. of 3rd IEEE Int. Conf. on Autonomic Computing (ICAC'06), Ireland, Jun 2006, pp. 65–73.

[15] M. Mesnier, E. Thereska, D. Ellard, G.R. Ganger, G.R., and M. Seltzer. File classification in self-* storage systems. In Proc. of Int. Conf. on Autonomic Computing (ICAC-04), pp. 44–51.

[16] W3C, "OWL Web Ontology Language Overview," 2004 Available at: http://www.w3.org/TR/owl-features/

[17] V. Wuwonse and M. Yoshikawa. Towards a language for metadata schemas for interoperability. In Proc. of 4th Int. Conf. on Dublin Core and Metadata Applications, China, 2004.

APPENDIX

The appendix includes the policy specifications and a list of mathematical expressions found in this paper.

A: POLICY SPECIFICATIONS

The policies as well as goals are expressed in OWL (Web Ontology Language) [16] and OWL/XDD (XML Declarative Description) [17], where the variables can be integrated with ordinary OWL elements. The variables are prefixed with the $ sign. In this paper, the policy is written in the form:

Conditions: Expressions_for_conditions,
Constraints: Expression_for_constraints,
Actions: {Action_ID},
Operation cost: Expression_for_operation_cost

Five policies (p_1-p_5) used in the case study are listed in Table III. The conditions can be the inherent service income $\$I_I <= 0$ and the waiting time $\$T_{Wait} >= \Delta$, where $\$T_{Wait} = \$T_{Wait,P} + \$T_{Wait,O}$.

TABLE III. THE POLICY SET

p_1	**Conditions:** $\$I_I <= 0$ or $\$T_{Wait} >= \Delta$, **Constraints:** $P_{WAIT}(STC_O) < P_{WAIT}(STC_P)$, **Actions:** $\{a_D\}$, **Operation Cost:** $P_{DISC}(STC_O)$ This policy can be read as: a_D should be used to disconnect a list of STC_O when $P_{WAIT}(STC_O) < P_{WAIT}(STC_P)$, and the number of STC_O being disconnected is calculated from $X_{P,1Mbps} * \$N_{Wait,P} / X_O$. a_D costs $P_{DISC}(STC_O)$ units.
p_2	**Conditions:** $\$I_I <= 0$ or $\$T_{Wait} >= \Delta$, **Constraints:** $P_{WAIT}(STC_O) > M(STC_P,X_{P,1Mbps})-M(STC_P,X_{P,600Kbps})$, **Actions:** $\{a_B\}$, **Operation Cost:** $M(STC_P,X_{P,1Mbps}) - M(STC_P,X_{P,600Kbps})$ This policy can be read as: a_B should be used to decrease the throughput of a list of STC_P when $P_{WAIT}(STC_O) > M(STC_P,X_{P,1Mbps}) - M(STC_P,X_{P,600Kbps})$, and the number of STC_P to decrease the throughput is calculated from $X_O * \$N_{Wait,O} / (X_{P,1Mbps} - X_{P,600Kbps})$. a_B costs $M(STC_P,X_{P,1Mbps}) - M(STC_P,X_{P,600Kbps})$.
p_3	**Conditions:** $\$I_I <= 0$ or $\$T_{Wait} >= \Delta$, **Constraints:** $(X_{P,1Mbps} * \$N_{Wait,P} + X_O * \$N_{Wait,O}) / C_{R,AL} > 0.1$, **Actions:** $\{a_N, a_I\}$, **Operation Cost:** $P_{Ser} * \Delta$ This policy can be read as: a_N and a_I should be used to instantiate a MS and to instantiate a new STM consecutively, when $(X_{P,1Mbps} * \$N_{Wait,P} + X_O * \$N_{Wait,O}) / C_{R,AL} > 0.1$. These actions $\{a_N, a_I\}$ cost $P_{Ser} * \Delta$.
p_4	**Conditions:** $\$I_I <= 0$, **Constraints:** $(X_{P,1Mbps} * \$N_{Wait,P} + X_O * \$N_{Wait,O}) / C_{R,AL} < 0.1$, **Actions:** $\{a_T, a_R\}$, **Operation Cost:** $P_{DISC}(STC_O) + P_{WAIT}(STC_P) – P_{Ser} * \Delta$ This policy can be read as: a_T and a_R should be used to terminate an STM and to disconnect a MS consecutively, when $(X_{P,1Mbps} * \$N_{Wait,P} + X_O * \$N_{Wait,O}) / C_{R,AL} < 0.1$. These actions $\{a_T, a_R\}$ cost $P_{DISC}(STC_O) + P_{WAIT}(STC_P) – P_{Ser} * \Delta$.
p_5	**Conditions:** $\$I_I <= 0$, **Constraints:** $(X_{P,1Mbps} * \$N_{Wait,P} + X_O * \$N_{Wait,O}) / C_{R,AL} < 0.1$, **Actions:** $\{a_M, a_T, a_R\}$, **Operation Cost:** $- P_{Ser} * \Delta$ This policy can be read as: a_M, a_T and a_R should be used to move connected STC sessions, to terminate an STM and to disconnect a MS consecutively, when $(X_{P,1Mbps} * \$N_{Wait,P} + X_O * \$N_{Wait,O}) / C_{R,AL} < 0.1$. These actions $\{a_M, a_T, a_R\}$ make profit = $P_{Ser} * \Delta$.

B: MATHEMATICAL EXPRESSIONS

Table IV lists all mathematical expressions. This table also expresses the relations to others expressions (Rel. to exp.) as well as the references to table (Ref. to tab.), where the notification used in the expression are defined.

TABLE IV. MATHMATICAL EXPRESSIONS.

No.	Mathematical expressions	Rel. to exp.	Ref. to tab.
1	$E \equiv \{ S_M, S_I, S_S, V, M, O, Q, F_S, F_O, F_V \}$	-	-
2	$R \equiv \{ \Theta, \Phi, \Pi, \xi \}$	-	-
3	$\xi \equiv (\overline{S_I}, \hat{S}_I, \overline{C_I}, \hat{C}_I, I_I, \hat{C}_{A,n}; n=[1, N])$	2	I
4	$\Pi \equiv \{ p_i \}$	2	-
5	$p_i \equiv (\Sigma_i, X_i, A_i)$	4	-
6	$\Sigma_i \equiv \text{Expression}(\overline{S_I}, \hat{S}_I, \overline{C_I}, \hat{C}_I, I_I)$	5	I
7	$X_i \equiv \text{Expression}(\overline{S_R}, \hat{S}_R, \overline{C_R}, \hat{C}_R, I_R, \overline{S_I}, \hat{S}_I, \overline{C_I}, \hat{C}_I, I_I, \hat{C}_{A,n}; n=[1, N], \Gamma)$	5	I
8	$L \equiv \{ \Omega, \Lambda, \Psi, \zeta \}$	-	-
9	$\zeta \equiv (\overline{S_I}, \hat{S}_I, \overline{C_I}, \hat{C}_I, I_I)$	8	I
10	$\Omega \equiv \{ g_k \}$	8	-
11	$g_k \equiv (d_k, \omega_k)$	10	-
12	$\text{reward}(a_i, i_{k,t-1}, d_k) = (\Delta(i_{k,t}, i_{k,t-1})/\Delta(d_k, i_{k,t-1})) * \omega_k - \text{cost}(a_i)$	8	
13	$A = \{a_D, a_B, a_N, a_I, a_R, a_T, a_M\}$	5	-
14	$I_I = M(STC_O,X_O) * T_{Serv,O} + M(STC_P,X_P) * T_{Serv,P} - P_{WAIT}(STC_O) * T_{Wait,O} - P_{WAIT}(STC_P) * T_{Wait,P} - P_{DISC}(STC_O) * N_{Disc,O} - P_{Ser} * N_{Node} * \Delta$	3, 6, 7, 9	II
15	$R_{CAM} \equiv \{ \Theta_{CAM}, \Phi, \Pi_{CAM}, \xi_{CAM} \}$	2	-
16	$L_{CAM} \equiv \{ \Omega_{CAM}, \Lambda, \Psi_{CAM}, \zeta_{CAM} \}$	8	-
17	$\Omega_{CAM} \equiv \{ g_1, g_2 \}$	10, 16	-
18	$g_1 \equiv (d_1: I_R > 0, \omega_1: 0.8)$	11, 17	-
19	$g_2 \equiv (d_2: T_{Wait} < \Delta, \omega_2: 0.2)$	11, 17	-
20	$\rho = ((\lambda_P * d_P * X_P) + (\lambda_O * d_O * X_O)) / (N_{Node} * C_{I,AL})$	-	-

Porting of C library

Testing of generated compiler

Ludek Dolihal
Department of Information systems
Faculty of information technology, BUT
Brno, Czech Republic
idolihal@fit.vutbr.cz

Tomas Hruska
Department of Information systems
Faculty of information technology, BUT
Brno, Czech Republic
hruska@fit.vutbr.cz

***Abstract*— For testing the automatically generated C compiler for embedded systems on the simulator, it is necessary to have corresponding support in the simulator itself. Testing programs written in C very often use I/O operations. This can not be done without support of C library. Hence the simulator must provide an interface for calling the functions of the operation system it runs on. In this paper we provide a method that enables programs to run, which use functions from the standard C library. After the implementation of this approach we are able to use the function provided by the C library with limitations given by the hardware.**

Keywords - Porting of a library; C library; compiler testing; simulation.

I. INTRODUCTION

One of the goals in our research group is an automatic generation of C compilers for various architectures. Currently, we are working on Microprocessor without Interlocked Pipeline Stages (MIPS). To minimize the number of errors in the automatically generated compilers, it is necessary to put the generated compilers under test. Because the whole process of compiler generation is highly automatic and we do not have all the platforms, for which we develop, available for testing, we use simulators for compiler testing instead of the chips or development kits. If one wants to test the C compiler within any simulator, it is necessary to add the support for the C library functions into the simulator, which is used for testing.

The support of the library is crucial in our project. We need to use tests written in C for the compiler testing and the tests commonly use I/O functions, functions for memory management, etc. This paper presents the idea of fitting the simulator, where the testing is performed, with support of the C library and later on the implementation of this method.

The paper is organized in the following way; the second section provides the position of the testing in the Lissom project, then a short overview of related work is given, section four discusses the reasons for choosing the library. Sections five and six discuss the theoretical and practical side of adding the library support into the simulator. Section seven concludes the paper.

II. POSITION IN LISSOM PROJECT

In the Lissom project [1], we focus mainly on hardware software codesign. In order to deliver the best possible services we want to provide the C compiler for a given platform as C is one of the main development languages for embedded systems. The C compiler is automatically generated from the description file. Besides the C compiler there are a lot of tools that are also generated from the description file. The tools include mainly:

- simulators,
- assembler,
- disassembler,
- profiler,
- hardware description.

Simulators can be either cycle accurate or instruction accurate. The profiler was thoroughly described in this article [2].

The description file is written in ISAC [3] language. The ISAC language is an architecture description language (ADL). It falls into the category of mixed architecture description languages.

We would like to produce the whole integrated development environment (IDE) for hardware software codesign. This IDE should provide all the necessary tools for developers when designing embedded systems from the scratch. The simulators are part of the IDE and the C library is part of the simulators.

The tool for generating compilers is called backendgen and is also embedded in the IDE. The backendgen was programmed manually; it is not generated. The quality of a compiler is crucial for the quality of software that is compiled by the compiler. Hence it is very important to test the compiler that is generated by the backendgen. Through locating the errors in the compiler itself we can afterwards identify and fix problems in the generation tools and in the whole development process.

The primary role of the C library is to enlarge the range of constructions that can be used during the process of testing. Without all doubts it is important to test the basic constructions such as if statement loops, function calls, etc. On the other hand it is highly desirable to have a possibility of printing outputs or the exiting program with different exit values and this can not be done without C library support. Exit values are the basic notification of program evaluation and debugging dumps are also one of the core methods of debugging. Note

 ISBN: 978-1-61208-139-7

that all the tests are designed for the given embedded system, and the tests are run on the simulator.

Secondary role of the library in the whole process of development is providing additional functions for writing programs. One of the most used groups of functions are functions for allocating memory, string comparison and parsing, input/output methods, etc.

As it is possible to generate several types of simulators in Lissom project, it will be necessary to add the library support into all types of simulators. It should not include any substantial changes to the process of generation.

III. RELATED WORK

Simulators in general are one of the most popular solutions as far as embedded system development is concerned. They are very often used for testing. We tried to pick up several examples that are connected to embedded systems development, and were published in the form of an article. The Unisim project is not aimed at embedded systems but provides an interesting idea.

In [4], a system very similar to the one that is developed within our project is suggested. It is called Upfast. The article describes system that generates different tools from a description file such as we do. The article mentions that C libraries were developed, but no closer information is given. It seems that in the simulator of the Unisim project the support for C language library has been right from the beginning. Unfortunately, this is not our case. Porting of the library is critical for us, because without the support it is very difficult to test and evaluate the results of any tests.

Another interesting system including a simulator is described in [5]. The project is called Rsim and is focused on the simulation of shared memory multiprocessors. The Rsim project works under Solaris. The Rsim simulator can not use standard system libraries. Unfortunately, it is not explained why. Instead the Rsim provides commonly used libraries and functions. The Rsim simulator was tested for support of a C library. All system calls in the Rsim are only emulated, no simulation is performed. In our system we will simulate the calls when necessary. The Rsim does not support dynamically linked libraries and our system also does not consider dynamic linking at the current state. Unfortunately, in the article [5] is not mentioned how the support for C library functions was added into the simulator.

The Unisim project [6] was developed as an open simulation environment which should deal with several crucial problems of today simulators. One of the problems is a lack of interoperability. This could be solved, according to the article, by a library of compatible modules and also by the ability to inter-operate with other simulators by wrapping them into modules. Though this may seem to be a little out of our concern, the idea of an interface within the simulator that allows adding any library, is quite interesting. In our case we will have the possibility to add or remove modules from the library in a simple way. But the idea from the Unisim project would make the import of any other library far easier than it is now.

The articles above are all related to simulations. The C programming language is not a new one and it is not possible to list all the articles that are in any way related to any library of C language. The simulator is either created in a way that it already contains the library or it has at least some interface which makes it easier to import the library in case it is wrapped in a module. Unfortunately, our simulator does not contain such an interface.

IV. CHOOSING THE LIBRARY

As we are focused mainly on embedded systems and we design the whole process of compiler development for them we dedicated quite a lot a time to choosing the correct library. It was clear right from the beginning that glibc is needlessly large and therefore not suitable for use in embedded systems. We need a library that satisfies the following criteria:

- minimalism,
- support for porting on different architectures,
- well-documented,
- new release at least once a year,
- compatibility with glibc,
- modularity.

All these conditions were satisfied by very few libraries. Amongst those we chose uClibc [7]. This library is largely minimalistic. It does not contain certain modules, because, according to the authors, it would be against the principle of minimalism. In certain areas it sacrifices better performance in favor of minimalism. For example, functions for I/O could be optimized for different platforms, but there is just one version for all platforms written in portable C that is optimized for space.

V. THEORY OF PORTING

The main reason for porting the library into a simulator is the fact that we need to add the support for C functions into the simulator itself. To be precise, we want to use the libc functions such as printf, malloc, free, etc. in the programs that will be used for testing of the compiler. And because we do not posses the development kits for all the platforms on which we run our tests we use simulators instead.

If one does not grant libc library support in the simulated environment, the number of constructions we can use and test is very limited.

Consider the following simple example written in C:

```
int main(int argc, char **argv)
{
   if(strcmp(“alpha”,”beta”)==0)
{ return 1;}
   else
{ return 0;}
```

 ISBN: 978-1-61208-139-7

}

Even this simple program can not be executed, because it uses function strcmp that is part of the C library. This program can not be compiled unless the inclusion of string.h and possibly some other header files is performed.

On the contrary the main aim of testing is to cover as wide an area as possible and also try as many different combinations of functions as we can. However, this goes against the idea of embedded solutions. And because we focus especially on embedded systems, we do not even try to cover all the functions provided by glibc, or in our case, uClibc. In fact we will use and hence test only functions that can run under the simulated environment and are useful for the programs that will be executed on the given platform. Moreover embedded systems are not designed for use of vast numbers of constructions that programming languages offer. Typically there is just one task, usually quite complicated, that is launched repeatedly. The functions we will use forms just a small part of uClibc. The functions that are not important to us can be easily removed via a configuration interface or manually. The following categories are examples of unimportant functions:

- threads, we assume that in simple programs for embedded systems one will not use threads,
- locales, all the locales were removed from the library,
- math, functions for computing sin, cos, etc.
- inet module, even though networking plays an important part in modern embedded systems the whole module was removed,
- files and operations with files, our application does not need an interface for working with files.

Now we come to the important parts of the library. Simply spoken all that really has to remain from the library are the sysdeps, this is the core of the whole system (how to allocate more memory, etc.), then important modules such as stdio (for outputs, inputs) and other modules we wish to preserve. In our case we wished to preserve the following parts of the uClibc library:

- stdio, this was the main reason for porting the library, to get in human readable form output from the simulator,
- a module for working with strings and memory, in our applications we would like to use functions such as memcpy, strcpy, strcat, etc.,
- memory functions, for example malloc, free, realloc,
- abort, exit,
- support for wchar, but without support of different encodings.

Some parts of the library could not be removed because of the dependencies. According to our estimations nearly 40 percent of the library was disabled or removed, measured by the size of the library.

There are several ways of building the library and also different methods of using it. There is a possibility of building a position independent code (PIC). Even though this is an interesting solution we decided against it. Instead of PIC we are going to compile the library into a single object and then link it to the program statically. The position of the library in the whole process of testing is shown in figure 1.

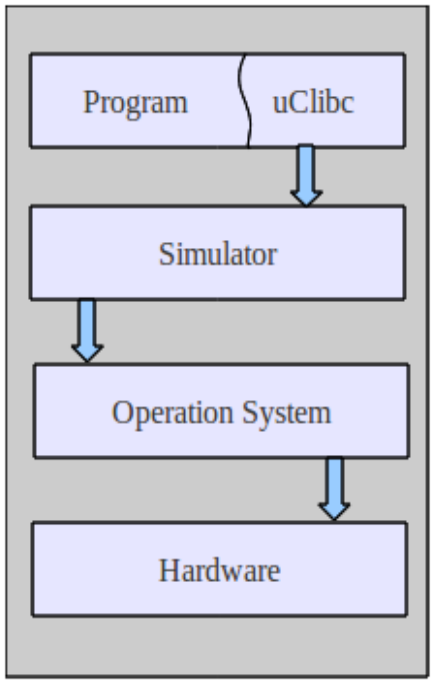


Figure 1. Position of the library in testing system.

Let's return to the functions that remain in the library. They can be divided into two groups. The first group consists of functions that are completely serviced within the simulated environment. For example, function strcmp falls into this category. This function and its declaration remain unchanged within the simulator if it is written in portable C. These functions are not tied with kernel header files so there is no need to change them.

The second group of functions consists of functions that are translated to the call of system function. Function printf can be used as an example of this group of functions. The call of printf function can be divided into three phrases that are illustrated at the following picture.

In the beginning the call of printf function is translated into the call of the system function, with the highest probability it is going to be the call of function write. Write, being the POSIX function is offered by the operation system. But as we want to use the simulator on Unix platform as well as on Windows systems we have to get rid of these dependencies. To do so we will use the special instruction principle.

A. Use of ported library of Unix and Windows systems

Before we get to the principle of a special instruction method we should explain why we need to use this method.

 ISBN: 978-1-61208-139-7

The main reason why we should oust the dependencies on the kernel header files is the fact, that we must be able to use the library under Unix like, and also under Windows like operation systems.

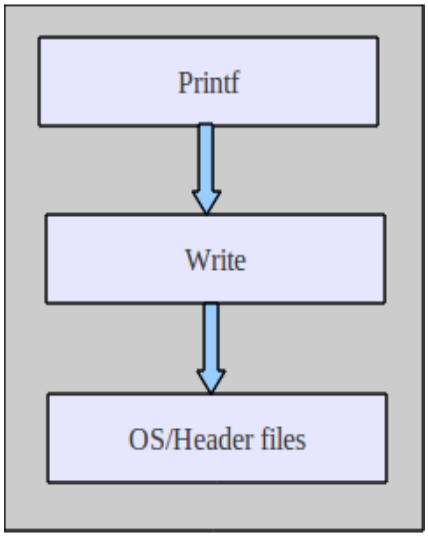


Figure 2. Scheme of calling the printf function

As long as we use the library under Unix systems everything should be all right. Though even on Unix systems there might be differences amongst the different versions of the header files. But once we use the Windows based system we can not use header file functions any more. It would almost certainly result in a crash of the system.

In our project we currently support several Unix distributions as well as Windows. Use of other operating systems is not considered.

B. Special instruction principle

The special instruction principle means, that we will use instruction with the operation code (opcode) that is not used within the instruction set of given microcontroler for the special purpose. We can do so, because we design the chip from the scratch. Usually the microcontroler has a given set of instructions. There is a defined bahaviour for each instruction. We describe the instruction set by our language ISAC. In ISAC, every instruction has an opcode and defined behaviour. So if there is any spare opcode we can model a new instruction with behaviour that suits our needs.

So far, all architectures that were modeled within the Lissom project had several free opcodes. It is typical that the instruction sets do not use all operation codes that are provided. But in case of no free opcode this method can not be used. The special instruction principle will be used for ousting the dependencies on kernel header files.

Functions provided by operation system are called by the system call (syscall) mechanism. The system calls can be quite easily detected. Each library should have defined the syscall mechanism in special source file. This syscall mechanism differs, as they usually are platform dependent. So i386 architecture will have different syscall mechanism than arm.

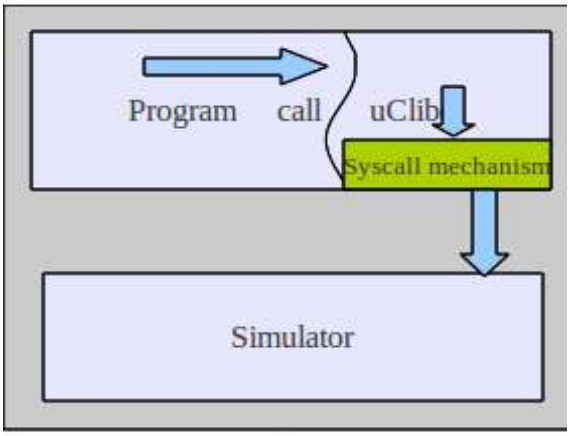


Figure 3. Scheme of calling the simulator via uClibc layer

We wish to preserve the mechanism. The syscalls will remain in the library, but with a different meaning. The file containing syscall will be changed in the following way: in the beginning the parameters of the syscall will be placed at the given addresses in the memory and we will also define where the syscall return value will be placed. Afterwards the call of the instruction, which was designed for this purpose, will be performed. It is also possible to put the parameters into registers, but some platforms have a limited number of registers, hence this method could cause problems.

C. Simulators

This brings us back to the simulators. As was mentioned before, all the simulators, where the testing is performed, are generated automatically. The generation is based on the instruction set description file, where our special instruction is modelled. In the beginning all the source files are generated by specialized tools. When the generation phase is finished the simulator is built by a Makefile from the generated files. It will be necessary to add into this process the following information:

- information about which instruction (opcode) calls the system function,
- the simulator will have to know the convention for storing parameters,
- the simulator will have to recognize which system function is going to be called,
- the simulator will have to perform the call of the correct system function.

The first three points will be solved within the model of an instruction set. The instruction with the opcode that is not used will be declared. The instruction behavior will be defined in the following way: it will locate the position in the memory where the parameters are stored and according to the value of one of the parameters it will call the corresponding system function. The simulator will have to recognize the system it runs under and call the correct function. For example, in Unix

 ISBN: 978-1-61208-139-7

system, it will be function 'write', and in Windows system, WriteFile. This should be solved by the libc library of the given platform. The following figure demonstrates the call of special instruction.

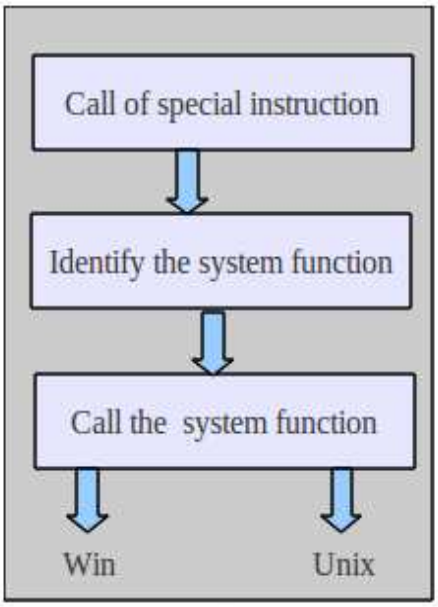


Figure 4. Calling sequence of specialized instruction

The parameters that were placed at the given position on the simulated memory can remain unchanged. They will later be passed to the specific system call.

One important issue is connected with the simulated memory. As we would like to correctly simulate the operations with memory such as malloc, realloc, etc. we need to tell the simulator how much memory it can simulate. This will be done most probably by the special file that will be passed to the linker. This file will contain symbols that will declare how much memory can be used. It is necessary to state how much memory can be allocated. The symbol that denotes the heap end will be used in the sbrk function.

VI. PROCESS OF PORTING

Before the whole process of porting begins we need to download the uClibc. There are two possibilities. It is possible to download only the library or there is a whole toolchain for development of embedded system for a given architecture, the so-called buildroot.

The main advantage of downloading the whole buildroot is that once it is built you get a whole set of development tools including various compilers, linkers, debugers, strip programs, etc. You also get the build of uClibc. These tools are quite useful in the beginning when you remove useless modules from the library, because they can be used for rebuilding the library.

One of the problems we faced is that we need to have the compiler for the architecture we are developing for. In other words, if we want to create a library for testing a C compiler on a given platform we need a compiler for the same platform that is already created. The compiler will be used for building the uClibc. Moreover the compiler must have exactly the same instruction set. In the future we would like to use the generated compiler for building the library. This requires a high quality of backendgen and generated backend.

Because we are going to use the library in the simulator and the simulator can handle only instructions of the specified instruction set, then the library must be translated to the instruction set that is recognized by the simulator. For building the simulator, we can use common gcc for Windows or Unix, because it runs under common system such as Windows or Unix.

This may be the first big problem in the whole process of porting. It is not hard to find a compiler for a given platform. Nowadays, there are specialized compilers for nearly all architectures used in embedded systems. The buildroot for uClibc contains more than a dozen different architectures such as MIPS, arm, mipsel, sparc, etc. There are even different versions of the micro-architectures in case of Microprocessor without Interlocked Pipeline Stages (MIPS), for example.

The problem is that, thanks to the aim of the whole Lissom project, we usually use specialized instruction sets or we use some generic instruction set and add certain specialized instructions. After this customization it is usually impossible to use a generic compiler for building the library.

We could use the compiler that we want to test for building the library but currently it is not stable enough for building large programs. The best solution of this problem is usually building a specialized toolchain including GNU binutils and GNU compiler collection [8]. As it was mentioned, once the generated backend is stable enough it will be used for building the library.

Several issues we faced during the process are closely related to the buildsystem of the library. The library contains a system of makefiles. This system is hierarchical and usually the makefiles from the upper levels are included. So, if for example we would like to compile any test examples that are included in the uClibc we switch to the given directory and call make. This will call all the makefiles from the above directory. This is very effective, because only the makefile in the root directory contains variables defining which compiler, assembler, linker will be used. On the other hand, it is very difficult to modify this system in case we want to build the different parts of the library using different tools.

Currently, we are using for development the set of our tools containing archiver, linker, asembler and compiler. The currently used compiler is called mips-elf-gcc. It is not generated automatically but was created specially for this purpose as our generated compiler is not stable enough for compiling the library. Our compiler has in the current version problems with floating point number, so it usually fails to compile them. Linker and archiver are not generated automatically for each platform but were developed in the Lissom project.

 ISBN: 978-1-61208-139-7

Our tools are not compatible with the tools that were originally used for building the library. Our tools do not support such a wide variety of parameters so some of them had to be erased from the configuration files and some were just changed to suit our needs. There are two main reasons for this. The first is that we simply do not need all the parameters. For example we always build files in elf format, so we do not need parameter to specify this. The second reason is that our files have different (usually text) format, that allows us to debug more effectively.

Currently, we use a set of scripts, which preprocess the flags. In the scripts we erase the flags we do not need and make necessary substitutions.

The buildsystem of the library starts by parsing the configuration file and accord to the content of the file are set different macros and variables. When doing manual changes to the buildsystem we have basically two possibilities:

- change the configuration file or,
- do the changes later in the Makefiles.

The first possibility is cleaner but the Makefiles often check if the option is present in the configuration file and ends with an error if the option is missing. Hence it is more convenient to make the necessary changes in the Makefiles. Thanks to the hierarchical structure it is in most cases sufficient to make the change in just one place.

In the theoretical part, we mentioned the need to link a special file containing information on how much memory can be used. The file will contain symbols defining the beginning and the end of memory space that can be used. It will have the following syntax:

```
#file defining memory boundaries

define start 256

define stop 768
```

Given that the numbers are in kB the simulator can simulate up to 512 kB of memory. Character # denotes comment.

For storing the parameters we have chosen the following approach: the first parameter says which system function is going to be called. In the uClibc it is a list of system functions for Unix systems. The rest of the parameters, that have numbers 2-7, are passed to the function call. The parameters remain unchanged. They are passed to the system function in exactly the same state which were saved in the memory before calling the special instruction. The special instruction itself has no parameters. When the instruction is called, all the parameters have to be stored in the memory at given addresses.

A. Automation of the porting process

For the first time, all the steps were performed manually. In the future we would like to automatize this process as much as possible. Without doubt we could remove the needless parts of the library automatically. The needless parts would be identified by the configuration file and also the special instruction principle could be highly automatic. If we have spare instruction we will choose it and compose it into the simulator. Unfortunately, there are steps that need to be performed manually, for example, we need to provide the runtime file for the simulators and the corresponding sections need to be specified in the ISAC file.

File with the runtime is also one of the files that is written by hand in the assembler. There are also other files written in assembly language and hence are platform dependent. In case of mips platform there were 8 files that contained assembly language. For example syscalls or memcpy functions are implemented in the assembler. In order to minimize number of files written by hand we decided to provide as many files written in portable C as possible. We managed to replace all but two files by C implementations. All that has to be provided is the runtime and syscall mechanism.

VII. CONCLUSION

In this paper, was sketched the idea of porting the library into the simulator. The motivation is quite clear: to be able to use the library functions in the tests that are run on the simulator of the given micro-controller. The special instruction principle was proposed which enables us to forward the call of system function. It also allows us to identify which system function is called. This principle is quite universal and can be used for the majority of platforms. After implementation of this method, we are able to run all the functions that are commonly used, such as I/O functions, memory management and string functions, etc. Moreover we can adjust the library according to our needs. Thanks to the modularity we can enable or disable any module. This may turn out to be an advantage, because the complete library has tens of megabytes and compilation and linking such a library can be time consuming.

ACKWNOWLEDGEMENTS

This research was supported by doctoral grant GA CR 102/09/H045, by the grants of MPO Czech Republic FR-TI1/038, by the grant FIT-S-11-2 and by the research plan no. MSM0021630528.

REFERENCES

[1] Lissom Project, doi :http://www.fit.vutbr.cz/research/groups/lissom, [online, accessed 19. 4. 2011] .

[2] Z. Prikryl, K. Masarik, T. Hruska, and A. Husar, “Generated cycle-accurate profiler for C language,” Proc. 13th EUROMICRO Conference on Digital System Design, DSD'2010, pp. 263—268, in press.

[3] ISAC language, doi:http://www.codasip.com/, [online, accessed 19. 4. 2011].

[4] S. Onder, and R. Gupta, "Automatic generation of microarchitecture simulators," Proc. 1998 International Conference on Computer Languages, May 1998, pp. 80-89, in press.

[5] C.J. Hughes, V.S. Pai, P. Ranganathan, and S.V. Adve, "Rsim: simulating shared-memory multiprocessors with ILP processors," Computer , vol.35, no.2, Feb. 2002, pp. 40-49, in press.

[6] D. August, J. Chang, S. Girbal, D. Gracia-Perez, G. Mouchard, D. Penry, O. Temam, and N. Vachharajani, "UNISIM: An open simulation environment and library for complex architecture design and collaborative development," Computer Architecture Letters , vol.6, no.2, Feb. 2007, pp. 45-48, in press.

[7] Uclibc, doi:http://uclibc.org/, [online, accessed 19. 4. 2011].

[8] GNU Operating System, doi:http://www.gnu.org/software/, [online, accessed 19. 4. 2011].

 ISBN: 978-1-61208-139-7

Updating Inventories with Intelligent Agents

Mary Luz Mouronte López[1,2], Francisco Javier Ramos Gutierrez[2]
Department of Telematic Engineering
[1]Universidad Carlos III de Madrid
[2]Ericsson España
Madrid, Spain
mmouront@it.uc3m.es, mary.luz.mouronte.lopez@ericsson.com

***Abstract*—This paper describes a procedure to automatically solve misalignments between inventories by means of expert agents. We apply this mechanism to maintain the repository of a Network Manager System (NMS), which allows executing different tasks over a transmission network in a Telecommunication Operator: fulfillment, supervision, etc. The consistency in the NMS inventory is maintained by means of connections to corporate systems (repository, planning, and assignment systems) and operations over Network Element Managers (NEMs) regarding to deployment of Network Equipments (NEs) or cards, setting up of circuits, paths and physical links. This repository contains data about managed elements, Synchronous Digital Hierarchy (SDH), Ethernet and Wavelength Division Multiplexing (WDM) components, and not managed elements: Radio, Plesiochronous Digital Hierarchy (PDH), fibers, etc. Sometimes misalignments occur between the corporate inventory and the repository in NMS. These misalignments have to be solved by the technicians. Our method incorporates the necessary intelligence to analyze failures and the mechanisms to solve them in expert agents, therefore preventing technicians from checking and solving main errors. It reduces Operating Expenditures (OPEX).**

Keywords - Transmission Network, NMS, NEM, NE, Expert Agents, CLIPS

I. INTRODUCTION

In this paper we describe a method to solve misalignments between inventories by means of intelligent agents. This procedure is applied to a Transmission **N**etwork **M**anagement **S**ystem (NMS) where the agents incorporate business information used by technicians when they do tasks manually. The solution both improves reliability and reduces human intervention (lower Operating Expenditures).

In a Telecommunication Operator, the NMS executes all business processes related to the transmission network [8] and its services across different vendors: fulfilment, supervision and performance monitoring in the network. It connects to corporate systems in order to automate routine tasks for fulfillment and to update the information about not managed elements. It also connects to vendor-specific **N**etwork **E**lement **M**anagers (NEMs).

This NMS manages **S**ynchronous **D**igital **H**ierarchy (SDH), **W**avelength **D**ivision **M**ultiplexing (WDM) and Ethernet over both SDH and **W**avelength **D**ivision **M**ultiplexing (WDM) networks. Other elements included in the transmission network but not related to management technologies (Plesiochronous Digital Hierarchy network elements, routers, fibres and radio elements) are also registered in the NMS to provide end-to-end vision.

The rest of the paper is organized as follows: Section II details relevant related works, Section III gives an overview about the NMS inventory and its current situation. Proposed technical solution is described in Section IV. Section V summarizes the results of applying this solution while Section VI describes main conclusions and areas for future works.

II. RELATED WORKS

There are not previous works on applications using expert agents to solve inventory problems on the transmission network from a NMS. Nevertheless, similar methods have been developed in others areas such as:

- Plant asset management where various research approaches and systems in this area exist [3].
- Detecting and preventing SQL injection attacks using **G**ene **E**xpression **P**rogramming (GEP) with intelligent agents in Web applications [9].
- Supply chains where a community of autonomous, intelligent, and goal oriented units cooperate and coordinate their decisions to reach a global goal [7].
- Studies to identify key concepts in different expert agent frameworks [10].

III. OVERVIEW

The main features of this NMS are:

- Network model based on standards like ITU-T G.803, ITU-T G.805, ITU-T G.709.
- It provides a vendor-neutral unified network management.
- It connects to the vendor **N**etwork **E**lement **M**anagers (NEMs). The interaction within the plant is carried out using the northbound interfaces offered by the NEMs which in turn deal with the **N**etwork **E**quipments (NEs).
- End to end control of the whole transmission network.
- Network inventory, which establishes relationships between the network vision in the network managers and the existing one in corporate systems. It also offers auditing and discovery mechanisms.

 ISBN: 978-1-61208-139-7

- Complete functional support in the following areas: fulfillment (network deployment and provisioning), supervision and performance monitoring. The fulfillment function allows executing operations over the NEMs regarding NEs, cards, circuits, paths and physical links.
- Simple user interface, based on well-known web technologies, defining operating profiles adapted to each user and enforcing a strict security policy.
- Connection to corporate systems in order to automate most routine tasks for fulfillment and to update information about not managed plant.

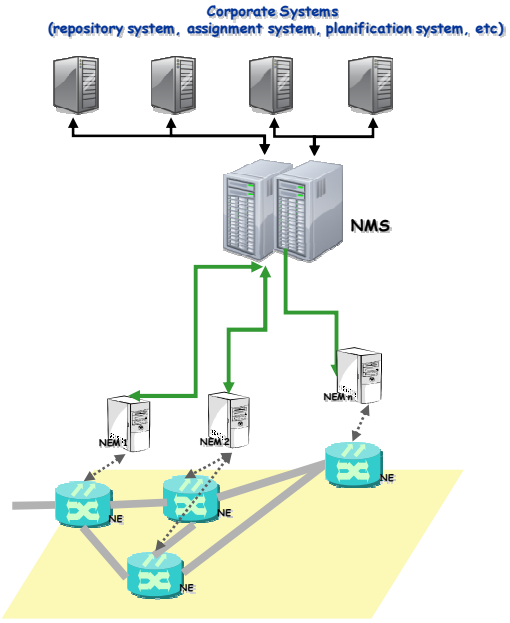


Figure 1. Transmission NMS

A. *Inventory interface between the NMS and corporate system*

The NMS is a system built upon a standards-based network model, backed by an ORACLE DBMS [2], with a business logic layer that allows interacting with the core applications through a CORBA bus.

The NMS receives information about transmission network elements, ports, paths, circuits, link connections, fibers and administrative data from a corporate repository, which contains all deployed plant, managed or not by NMS.

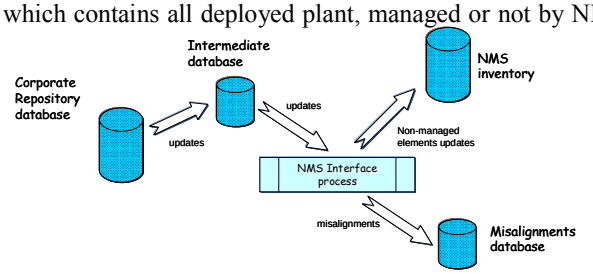


Figure 2. Inventory interface

This repository is a network inventory which uses proprietary operator names to label its elements.

In order to update the NMS inventory, a database-to-database interface has been implemented between the corporate repository and the NMS (fig. 2). This interface works the following way:

- The corporate repository captures each update generated in its database and inserts it into an intermediate database through ORACLE Net.
- A process in NMS reads the updates from this intermediate database and transforms it according to the NMS model.
- A NMS interface process establishes the operation to be performed. If the update affects to a not managed element, the NMS inventory is updated. Otherwise, it determines if received data is consistent with stored data in its inventory. If it is inconsistent, an error is generated and inserted into a misalignments database.
- NMS provides functions to search misalignments and find relationships among them. By means of this set of operations, misalignments are solved manually by technicians. There is a big delay until a misalignment is solved and further misalignments caused by it may appear. Moreover, some operations in the NMS are not executed (remain pending) until the misalignment is solved.

B. *Current situation in the NMS inventory*

After analyzing misalignments in this interface between the NMS and the corporate repository during two months, 1300 errors were found. The percentage of each class error over the total failures is showed bellow (per entity basis).

1. Misalignments in circuits,
   - Link capacity is not registered in the NMS (30%): It does not exist in the path; has different structure or administrative properties.
   - Port is not assigned to the circuit in the NMS (30%): circuit is incomplete or has different administrative properties.
   - Link capacity is not assigned to the circuit in the NMS (22%): circuit is incomplete or has different administrative properties.
   - Port is not registered in the NMS (5%): card or port has different number, tasks have not been done in the NMS or port is in a card depending on other one.
   - Link capacity is not assigned to the circuit in the corporate system (4%): circuit is erroneous in the NMS.
   - Port is not assigned to the circuit in the corporate system (2%): circuit is wrong in the NMS.
   - Circuit does not exist in the corporate repository (0.5%): some operations were not done in the NMS.
   - Others (6.5 %)

 ISBN: 978-1-61208-139-7

2. Misalignments in paths,
   - Path does not exist in corporate repository (30%): some operations over the path were not done in the NMS.
   - Port is not assigned to the path in the NMS (18%): path is incomplete or has different administrative properties.
   - Port is not registered in the NMS (18%): card or port has different number, tasks have not been done in the NMS or port is in a card depending on other one.
   - Link capacity is not assigned to the path in the NMS (15%): path is incomplete or it has different administrative properties.
   - Link capacity is not assigned to the path in the corporate system (10%): path is erroneous in the NMS.
   - Port is not assigned to the path in the corporate system (5%): path is erroneous in the NMS.
   - Link capacity is not registered in the NMS (2%): Link capacity does not exist in the path; has different structure or administrative properties.
   - Other (2 %)
3. Misalignments in link connections,
   - Link connection does not exist in the NMS (42%): it has different structure.
   - Link connection does not exist in the corporate system (34%): it has unlike structure.
   - Link connection is occupied in the NMS and not in the corporate system (7%): occupancy is incoherent.
   - Link connection is occupied in the corporate system and not in the NMS (10%): occupancy is incoherent.
   - Different link connection type in the corporate system and in the NMS (6.5%): it has different structure.
   - Link connection is not in service in the NMS (0.5%): it is included in a pending operation.
4. Misalignments in cards,
   - Card does not exist in the NMS (60%): card number is different or some operations have not done in the NMS.
   - Card in service in corporate system (21%): it is include in a pending operation.
   - Card does not exist in corporate system (18%): some operations have not done in the NMS.
   - Different card number in the NMS and in corporate system (1%): some tasks have not done in NMS.
5. Misalignments in ports,
   - Card does not exist in the NMS (42%): card number is different or some operations have not done in the NMS.
   - Port does not exist in the NMS (31%): card has unlike number, some tasks have not done in the NMS or port is in a card which is depending on other card.
   - Port does not exist in corporate system (19%): tasks have not done in the NMS.
   - Different port position (6.5%): some operations have not done in the NMS.
   - Port in different card (1%): port is assigned to a depending card.
   - Port in service (0.5%): port is included in a pending operation.

Depending on the error, a technician needs to execute some tasks:

- Get additional data from the corporate repository, the NMS inventory and NEMs.
- Analyze obtained data.
- Decide which operations to perform in the NMS inventory.

The actions which can be executed in the NMS inventory are: modify (change its link connections, ports, structure, state) or remove the circuit or path, create or remove the card, change the number of the card, modify the state of the card, create or remove the port, change the number of the port and modify the state of the port.

Technicians have to spend much time analyzing error causes, comparing both inventories and solving misalignments. On a daily basis, the average spent time is about several minutes per failure depending on the error type.

TABLE I. TIME FOR ANALYZING AND SOLVING MISALIGNMENTS BY TECHNICIANS

Paths and Circuits	
Misalignment type	***Average time (minutes)***
Link capacity does not exist or is not assigned in NMS	13.70
Port does not exist or is not assigned in NMS	13.70
Circuit or path does not exist in the corporate repository	7.30
Link capacity is not assigned to the circuit path in the corporate system	11.30
Port is not assigned to the circuit or path in the corporate system	11.30
Cards	
Misalignment type	***Time (minutes)***
Card does not exist in NMS	15
Card in service in corporate system	3
Card does not exist in corporate system	1.10
Link Connections	**Average Time (minutes)**
All types	6.10

 ISBN: 978-1-61208-139-7

Ports	Average Time (minutes)
All types	4.90

IV. SOLUTION

The proposed solution in this paper, allows supervising the errors in the interface between the NMS and a corporate repository.

Our technical solution uses a knowledge base which contains rules to represent the actions executed by a technician when an error occurs. A method based on ORACLE queues is used to capture and notify asynchronous errors in NMS. In the NMS, CORBA operations are implemented to get data from NEMs and to perform tasks in the NMS inventory.

The advantages of the implemented solution are:

- Changes are not required in the interface between the NMS and the corporate repository. Errors are processed in background mode and there are not delays by the processed failures.
- Errors are processed as soon as they happen. Further misalignments are not generated by a previous one.

The architecture of the solution is shown in Fig 3.

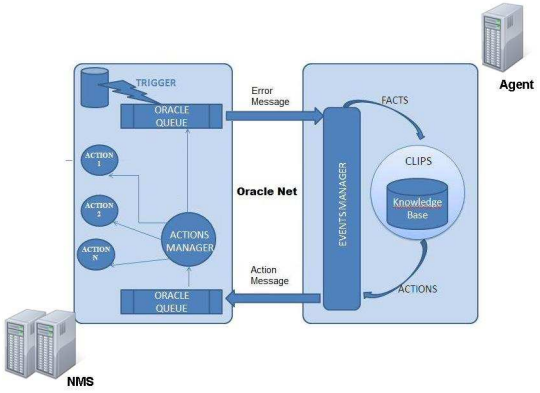


Figure 3. Architecture of the technical solution

C **L**anguage **I**ntegrated **P**roduction **S**ystem (CLIPS) is an open source expert system tool developed by NASA-Johnson Space Centre [1]. It is fast, efficient, free and updated and supported by the original author, Gary Riley.

The components in the architecture are: intelligent agents (*Event Manager* and *Action Manager*) and a *Knowledge Base* which contains different rules.

Intelligent Agent: *Event Manager*

In the NMS, a trigger is fired by an error in the interface and it writes an event in an ORACLE queue. The *Event Manager* (with CLIPS embedded) is connected to database through ORACLE Net. It executes the following tasks:

- Read an event from ORACLE queue. The format of the message in ORACLE queue is:
  message_type|entity_type|Id_entity|entity_name|text|element_type|Id_element|report_name
- Generate a fact and add it to the Fact Base which, is used by the Inference Engine. A fact has the following format:

```
(deftemplate misalignment
  (slot  entity_type (type STRING))
  (slot Id_entity (type STRING)))
  (slot name (type STRING))
  (slot text (type STRING))
  (slot elem_type (type STRING))
  (slot Id_element (type STRING))
  (slot report_name (type STRING))
)
```

- Run the Inference Engine which decides the actions to be performed based on facts contained in the Fact Base and the knowledge base. It contains different types of rules:
  - Correlation rules: to establish relations between facts and to construct new ones.
  - Decision rules about the actions: to establish the specific criteria to shoot actions.
  - Metacontrol: to fix priorities in the rules according to their relevance.
- Send the action to an ORACLE queue. It will be processed by the *Action Manager*.

  The event format is:

  Action_type|entity_type|Id_entity

Intelligent Agent: *Action Manager*

This agent executes the actions generated by the Inference Engine and interacts with the NMS to perform the suitable operations. This intelligent agent read the actions from ORACLE queue and executes the specific tasks in NMS. There are different operations:

- Data acquisition operation: It is executed to get data from the corporate repository, the NMS inventory or NEMs. It can be a search in database or a complex CORBA operation.
- Final operation: It is a task to solve an error or to require a technician. It can be a transaction in a database or a complex CORBA operation.

In case of data acquisition operations, when the task is finished, the *Action Manager* sends a message to the *Event Manager* to inform about the result of the task. The message format is:

 ISBN: 978-1-61208-139-7

Message_type|entity_type|Id_entity|action|value

When the *Event Manager* receives this message, a new fact is generated and added to the Fact Base. The format of this fact is shown below:

```
(deftemplate condictions
   (slot entity_type (type STRING))
   (slot Id_entity (type STRING))
   (slot type (type STRING))
   (slot value (type STRING))
)
```

This process of adding facts to the Fact Base is repeated until a final action is obtained.

A particular case is shown below: misalignments due to a port not assigned to the path in the corporate system

1. *The Event Manager* receives the following event from the ORACLE queue:

   MISALIGNMENT|CIRCUIT|265792011|2/AB.LL ETH 7/2/AB.PA008ETH 1/ETH0/0001/NIGCM /|Port is not assigned to the circuit |020001003000089S000189|MISALIGNEMT IN CIRCUIT

2. *The Event Manager* adds a fact to the Fact Base:

   misalignment(entity_type "CIRCUIT") (Id_entity 265792011) (entity_name "2/AB.LL ETH 7/2/AB.PA008ETH 1/ETH0/ 0001/NIGCM /|") (text " Port is not assigned to the circuit ")(type_element "Port")(Id_element "020001003000089S000189")(report_name " MISALIGNEMT IN CIRCUIT")

3. The Inference Engine decides the required actions to be sent to the *Action Manager*

   ```
   (defrule misalignemt_circuit
   ?h1 <- (misalignment (entity_type
   ?ent_type) (Id_entity ?Id) (entity_name
   ?name) (text " Port is not assigned to the
   circuit " )(type_element
   "Port")(Id_element ?elem)(report_name
   ?report_name))
         (not(process_misalignment
   (entity_type ?entity_type) (Id_entity
   ?Id_ent) (name ?previous_name) (type
   ?type) (pending_elem_number ?number)
   (depending_entity_type
   ?depending_ent_type)
   (depending_entity_Id ?depending_ent_Id)
   ))
   =>
   (retract ?h1)
   (assert ( process_misalignment
   (entity_type ?entity_type) (Id_entity ?Id)
   (name ?name)( adaptation type)
   (pending_elem_number 0)
   (depending_entity_type "")
   (depending_entity_Id "") ))
   (send_action "there_is_request"
   ?entity_type ?Id)
   ```

4. The following message is read from the ORACLE queue by the *Action Manager* :
   there_is_request|CIRCUIT|265792011
5. *The Action Manager* executes the necessary actions in the NMS and sends a new message to the *Event Manager* through the ORACLE queue:
   CONDITION|CIRCUIT|265792011|there_is_request|NO
6. A fact is generated and added to the Fact Base:
   condition (entity_type "CIRCUIT") (Id_entity 265792011) (action_type "there_is_request") (value "NO")
7. The Inference Engine decides the required actions to be sent to the *Action Manager*

   ```
   (defrule misalignment_no_request
   ?h1 <- (process_misalignment (entity_type
   ”CIRCUIT”) (Id_entity ?Id_ent) (name
   ?name) (type ?type)
   (pending_elem_number ?number)
   (depending_entity_type
   ?depending_ent_type)
   (depending_entity_Id ?depending_ent_Id))
   ?h2 <- (conditions (entity_type ”CIRCUIT”
   ) (entity_Id ?id_ent) (type
   “there_is_request”) (values NO))
   =>
   (send_action "modify_circuit" ?ent_type
   ?Id)
   (retract ?h1)
   (retract ?h2)
   ```

8. The Inference Engine generates an action to modify the circuit and puts the following message in the ORACLE queue:

   modify_circuit|CIRCUIT|265792011

9. *The Action Manager* reads the message from the ORACLE queue and modifies the circuit.

 ISBN: 978-1-61208-139-7

V. RESULTS

The solution was implemented in the production environment over 600 misalignments obtaining the following results:

TABLE II. TIME FOR ANALYZING AND SOLVING MISALIGNMENTS BY EXPERT AGENTS

Paths and Circuits	
Misalignment type	***Average time (seconds)***
Link capacity does not exist or is not assigned in NMS	169
Port does not exist or is not assigned in NMS	169
Circuit or path does not exist in the corporate repository	90
Link capacity is not assigned to the path in the corporate system	139
Port is not assigned to the path in the corporate system	139
Cards	
Misalignment type	***Time (seconds)***
Card does not exist in NMS	185
Card in service in corporate system	36
Card does not exist in corporate system	10
Link Connections	**Average Time (seconds)**
All types	42
Ports	**Average Time (seconds)**
All types	63

The solution reduces the time needed for analyzing and resolving problems. Therefore, technicians are required only when the solution depends on pending task in the NMS or produces a loss of quality of service, as shown in Table III.

TABLE III. STATISTICS

Solution	***Percentage***
Misaligments solved by agents	78%
Misalignments where technicians are required in final stages. Technicians decide when the tasks have to be done in network equipments in order to maintain the quality of service.	13%
Misalignments where technicians are required to perform pending tasks in NMS (routing, programmed jobs, etc.)	9%

When we apply this method to resolve all misalignments, it can be necessary to add new rules and actions in the expert agents.

VI. CONCLUSIONS AND FUTURE WORKS

This paper described a good solution to solve misalignments between inventories. In particular, we applied the method to solve the problems between a corporate repository and an inventory of a NMS in a Telecommunication Operator.

This method reduces operating expenditures because it is not necessary to hire technicians to resolve main errors and a solution is provided as soon as a problem is detected. The few unsolved problems can be addressed by those technicians in charge of fulfilment operations in the NMS.

The procedure can be applied to other areas:

- Adaptive graphical user interface [4], [5].
- Diagnosis of problems in different networks [6], [7].

REFERENCES

[1] "CLIPS. A Tool for Building Expert Systems", http://clipsrules.sourceforge.net/index.html, April 2011.

[2] "Oracle9i Net Services Administrator's Guide", http://download.oracle.com/docs/cd/B10500_01/network.920/a96580/toc.htm, April 2011.

[3] "Using Multi-Agent Systems for Intelligent Plant Maintenance Functionality", http://www.tik.ee.ethz.ch/~naedele/WCICA04.pdf, April 2011.

[4] C. Mourlas and P. Germanakos, "Intelligent user interfaces: Adaptation and personalization systems and technologies", Information Science Reference, September 2009.

[5] F. Nasoz and C. L. Lisetti, "Affective user modeling for adaptive intelligent user interfaces", Human Computer Interaction. HCI Intelligent Multimodal Interaction Environments, pp. 421-430, August 2007

[6] L. Bunch et al., "Software agents for process monitoring and notification", Proc. ACM Symp. Appl. Comput., Vol. 4, pp. 94-99, 2004.

[7] J. García, P. Arozarena, S. García, A. Carrera, and R. Toribio. "A Lightweight Approach to Distributed Network Diagnosis under Uncertainty". In "Intelligent Networking, Collaborative Systems and Applications", Springer, pp. 74-80, 2010.

[8] M. L. Mouronte, R. M. Benito, and J. P. Cárdenas, "Complexity in Spanish optical fiber and SDH transport networks", Computer Physics Communications, Vol. 180, No. 4., pp. 523-526, 2009.

[9] S. Kadirvelu and K. Arputharaj. "Intelligent Agent Based Prevention System for Web Applications from SQL Injection Attacks Using Gene Expression Programming", European Journal of Scientific Research, Vol.49, No.2, pp. 286-292, 2011.

[10] W. C. Regli et al. "Development and specification of a reference model for agentbased systems.", IEEE Trans. On Systems, Man, and Cybernetics, Part C, Vol. 39, No.5, pp. 572–596, 2009.

 ISBN: 978-1-61208-139-7

An Information-on-demand E-learning System

Chien-Hsing Chen, Jun-te Huang, Hsiao Ping Lee
Department of Applied Information Sciences
Chung Shan Medical University,
Taichung 402, Taiwan, R.O.C
cy3331@gmail.com

Tzu-Fang Sheu
Department of Computer Science and Communication Engineering
Providence University,
Taichung 433, Taiwan, R.O.C
fang@pu.edu.tw

***Abstract*—Due to the fact that Internet is widely developed, more and more people learn through network. Traditionally, an e -learning system is usually cooperated by Learning Management System (LMS), which is convenient for administrators to manage the course but difficult to get information there through. Until now, e-learning2.0 has become the main trend of current study. It indicates that users are being more interactive with each other. In this paper, we propose a novel information-on-demand e-learning system. The proposed system provides an audio-based information learning service, which lets us learn not only depend on text reading but auditory and visual learning. The proposed system searches and converts the customized information or user-specified articles on the Internet to audio, and no assistive tool is required on user sides. Via modern text-to-speech technology, the conversion can be done automatically. Moreover, our system provides the self-learning areas, in which the users can choose whatever material they like to upload. Our e-learning online platform has a large number of advantages including customized learning, convenience of acquiring information and abundant resources. The proposed system realizes the goal of ubiquitous and friendly access of the Internet for anyone from anywhere. We can learn wherever without time limitation.**

Keywords - e-Learning2.0; information-on-demand; self-learning; text-to-speech

I. INTRODUCTION

E-learning [1, 2] was proposed by CEO forum [3] on education and technology (ET CEO) in June, 1998. Since then, governments, enterprises and schools began to take part in online courses and distance learning [4]. Nowadays, more and more people get the latest information from the Internet anytime, anywhere. Nevertheless, the development of e-learning is still slow, because people do not adapt themselves to the new way of learning and working. Therefore, the quality of e-learning is doubted no matter in academia and enterprises as a result e-learning still plays an assistant role. Even though the government offers large quantities of resources to develop e-learning and apply the system somehow, but they still ignore the most critical point that how to make people want to use e-learning to learn. That is to say, if people are not motivated, people would not have interests to learn actively and automatically.

The term "Web2.0" was brought up by Tim O'Reilly [5] with four concepts, which are:

- All the functions are operated through the browser.
- Users can distribute their own information.
- In-time sharing and interaction between users.
- User-centered.

Currently, Web2.0 [6] is changing the learning and teaching methodology. Interactive web technology helps people to communicate easily. Suppose e-learning can make people connect to the issues they concern and be momentarily updated with latest information, it will provide a preferred and motivational environment for learners. On the other hand, Web2.0 [7] also emphasizes personalization, which everyone can pick subjects they are interested in and arrange the layout as per their preference.

RSS [8, 9] (Really Simple Syndication) adopts these concepts of Web2.0 and become a main technique of information exchange. RSS is used to share news and content of Internet and interchange information for people who follow XML [10, 11] standard format.

In this paper, we build an automatic web-based information-on-demand e-learning system (ODES) to assist the learner to learn the latest information through RSS. Computer science is the main learning resource of our system. First of all, we acquire the news information from RSS and we convert the information into audio type and capture the images through Google API. Then we can fulfill the requirement of e-learning with daily updated information.

With the new e-learning method of our system, we hope to turn the traditional face to face instruction into a worldwide web learning platform, so as to make the user be more active in learning activities, and thus to effectively increase the interaction with the digital learning and realize the goal of learning everywhere and smart living. The rest of the paper is organized as follows. We discuss the present weakness of e-learning and the development of RSS service in Section 2. The implementation of information-on-demand e-learning system is presented in Section 3. The comparison of our ODES and other e-learning system is described in Section 4. Finally, we conclude the paper in Section 5.

 ISBN: 978-1-61208-139-7

II. GROWTH OF E-LEARNING

A. Traditional teaching and e-learning

In the past, the learners must stay in the same space, face to face with their teacher in traditional instruction, and the learners have no freedom, which raise their spiritual pressure. The learners have less chance to arrange the time they want to spend on learning not to mention choosing places to learn for themselves. Nowadays Internet has been regarded as an interactive area for human to exchange their news and information. World Wide Web is also increasingly widely applied to teaching and learning. Learners do not have to go to school, not only stay at home, anyone can learn at any place. E-learning [12] has one of the best advantages is that people learn wherever without time limitation. When we use the Internet as a medium, we can reach any place in the world through web. Owing to our network, we can learn a lot of information by surfing the Internet. This will be the mainstream of education mode in the future. Users can build their own custom-made platforms according to their requirements. In addition, users can break the limitation of where they are, perfectly using their free time to learn. There are some benefits as follows: Reducing learning cost, resources being reused, and no spacious space needed.

The platform of e-learning1.0 [13, 14] usually includes LMS (Learning Management System) and LCMS (Learning Content Management System). LMS is an online teaching platform for learner to manage their learning conditions. LMS provides a catalogue of online courses [15, 16] to control the learning schedule, a registration system to help learners record their learning conditions, and useful tools to assist learners, such as a test system which gives the learners a test on the platform, e-mail groups and forum services for learner to communicate with each other. LCMS is also a platform and used to manage or modify the teaching materials. The resource of elearning1.0 is provided by supplier. Due to the provision of teaching materials are unidirectional, which require abundant of time to pre-make those teaching materials. Then, learners may acquire the content of courses through LMS. E-learning2.0 was proposed by Stephen Downs [17] in 2006. E-learning changed the method of traditional instruction. The information was not unidirectional provided by teacher. Learners can also provide the information to others. This pattern will establish great interaction. E-learning2.0 [18] is composed of many primary elements such as wiki, blog, RSS, social networking, mash up, Ajax [19, 20]. Because the information relies on global digital resources it is updated with a high speed.

B. RSS of e-learning

In traditional web pages, learners only can learn from static web pages. Learners need to find significant resources in disordered data, so that they spend so much time filtering resources. These resources are scattered therefore learners cannot organize all the web content effectively. RSS is an e-learning resource that uses technical framework of XML. This technology allows learners to subscribe the information they want to acquire so as to get the learning resources, and reduce the time of searching in unprocessed data. In tradition, if people want to know whether there is any new information in the websites or blogs, they have to link to the URL and check the information on the web in person. Suppose people want to obtain the information in more than one website, they have no choice but to check all the web pages one by one. Eventually, RSS was created by Netscape that can be used to integrate the significant information from webs and blogs. When information is updated by web administrators, RSS will automatically notify the subscribers of new information. The subscribers can easily learn the content of web pages without linking to the URL. RSS is also widely used on blogs. When authors of blogs write new articles, RSS will notify the subscribers of new articles in blogs.

RSS gradually developed many new services, such as podcast [21]. The podcast is an audio-based online service. The podcast is, normally, a series of periodically released digital media files with some attributes, such as date, title and description. Users pick programs they like and decide when to listen or watch these programs through the podcasting service. In addition, any individual can create a podcast and make people who are interested in to subscribe. The subscriber can automatically receive new podcasts, without going to a specific site and downloading it there from. Podcasting is a broadcasting service on demand. E-learning2.0 [22, 23, 24] brings a different method for users to learn. It focuses on interaction between people for building the best learning experience and result. Podcasting service can be described as a RSS link with media files. All of the content provided to the services must be pre-made by human beings. Therefore, podcast offers a learning method just like users listening to the Internet radio, they can only get information provided by other distributors, but not themselves. If one wants to find some certain specific information, it will be not easy to obtain the information he really needs. He still has to download the media files and check it one by one until the information be founded.

In this paper, our system provides a search tool helping users to find the information they exactly want. It will be much better for users to spend their precious time learning than checking through all materials they do not need.

III. THE INFORMATION-ON-DEMAND E-LEARNING SYSTEM

The information-on-demand e-learning system (ODES) is composed mainly of five modules. The system structure is presented in Figure 1.

These five modules include information capturer, information parser, text-to-speech (TTS) engine [25], photo search engine and e-learning platform.

 ISBN: 978-1-61208-139-7

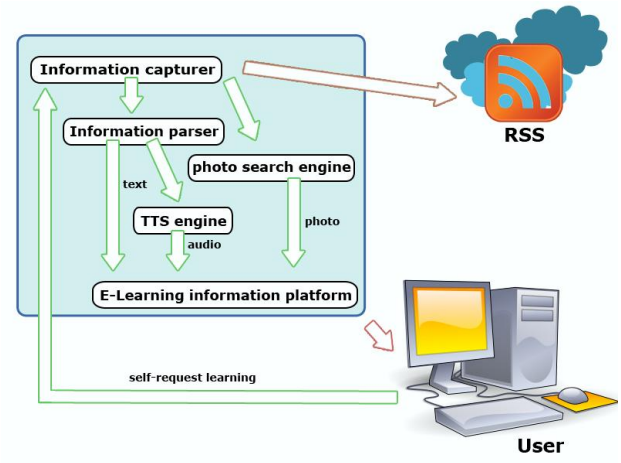


Figure 1. The architecture of the ODES system.

E-learning platform module is an initial area that links the user and information. Here we provide two novel models for users to get their information.

All of the modules are implemented in JAVA-based technology and techniques. The following introduces these 5 modules and two learning models:

A. *Information capturer*

Really Simple Syndication (RSS) is a standardized format of web information, and our system will automatically capture the specific link so as to make our website updated immediately with latest news and information, such as blogs, news headlines, audios, and videos. Our system collect full text with publishing dates of RSS document from other website, RSS can be a transmitting agent via which people communicate with each other. Therefore, our system captures the information of the text and makes it a learning resource. In other words, our system collects information and provides a platform for integrating all the information to form an information pool.

B. *Information parser*

When the remote RSS is updated, the system will obtain the information from RSS, and then update to the system. After getting RSS documents, which will be saved as separate files of XML format, our system then gives those files to information parser for analyzing the XML format, which is packed into a data structure, and in turn abstracts the text information, such as the path, size, category, subject and updated date. There are many ways to parse XML, the traditional analysis method is DOM [26], which is a W3C standard way dealing with XML documents. However, the memory needs to read the entire file and transform it into a structure of tree, thus it will spend more time to analysis, and use excessive resources. Our system is using SAX [27, 28], which will be more efficiently in analyzing, reading and operating XML documents, because the files are analyzed right upon SAX reading the files. SAX is event-based processing model and functions with the generations of events. After processing via event handler, SAX changes the XML files into a series of events. When the events generate, only single corresponding event handler will be called to deal with them. Compared to parsing methods of DOM, SAX is more suitable for the system. The output of the module is in text format, and is delivered to both the TTS engine and e-learning platform.

C. *Photo search engine*

Information parser will deliver the significant sentence to photo search engine, and the photo search engine will find the pictures relating to the text by Google API. The picture will help users to understand the meaning of the foresaid sentence, and more to get information efficiently, as well as to enhance the absorption of learning.

D. *Text-to-speech (TTS) engine*

TTS engine can transform text into audio speech. The TTS engine is an implementation of Microsoft SAPI (Speech API) [29] interface, and is compatible with SAPI 5 specification. After information parser separates the text which will be passing to TTS engine and processed to speech. The TTS engine module converts text data to speech in WAV or MP3 format. There are many speakers to be chosen from TTS engine, including male and female vocals. User can use different settings to change the vocal of the speech. After the text is changed into speech, it will be placed to e-learning platform

E. *E-learning platform*

E-learning platform is the main core to this paper, and is a gateway that people can communicate with each other on it. On the platform we provide two learning methods as follows:

The first one is to directly use the information that stored in our system, the integration of speeches, words, pictures, and to provide a quick access to different subjects, including computer science, information security, digital life, Android, mediaPC, software, wireless, magazines and English learning. When the information is updated, our system will start to update the information to the e-learning platform, and keep all the information latest.

Our system has another feature, that is, English learning area. The English learning area is presented in Figure 2. There are many ways to learn English in the past; you may go to the language school etc., but all these ways will cost you a lot of money. If you use ODES, you can learn English for free, and you do not require any books or magazine to learn. With your computer and network, you will improve English ability, all on your own. Users will hear the voice to

 ISBN: 978-1-61208-139-7

Figure 2. The English learning of the ODES system.

Figure 3. The home page of the ODES system.

read the English topic as well as the content of the article to study language from our learning area.

The second way is self-learning. This area let users upload text files or paste their RSS hyperlinks, and the information parser will analyze the XML and convey the words required to TTS engine. TTS changes the text into audio, Finally, it presents readable information on e-learning platform. It is a different pattern of ways to learn from the past. This kind of way to learn plays a better role as a learning tool, because it is easy to operate and make the users be motivated by their own interests so as to reach the goal of self-learning.

Except for the learning of general people, our system additionally helps the learning of the blind people. Because of physical disabilities, the blind cannot use those systems with visual interface. In order to allow those who visually disable to use our system, the information platform follows the Web Content Accessibility Guidelines in the specification, such as Access Key etc.

IV. RESULTS

The proposed ODES, except the TTS engine module, was implemented on a PC with one Intel Pentium 4 3.0GHz CPU, 768MB RAM and 80GB disk space. The operating system was Microsoft Windows XP home edition. The TTS engine was implemented on a PC with one Intel Core 2 Duo E6550 2.33 GHz dual-cores CPU, 2GB RAM and 160 disk space. The operating system was Microsoft Windows XP professional edition. All of the modules in the ODES were implemented in JAVA language [30]. The home page of the ODES system is presented in Figure 3.

Our system is introduced to teachers from elementary schools to universities. By using our system, they can get the pictures and audios from the Internet as references, which make the courses more variously. Even students and ordinary people can make use of the system to obtain and learn the information they need in different ways. Through our system, we merge the information technology with education to offer multiple learning materials and develop different ways of learning. Moreover the traditional text-based instruction is no longer the main trend. Sounds and pictures make better learning result. User has the autonomy to make the best use of their time, and trigger the motivation of user's self-learning, so that e-learning can bring the maximal effectiveness of learning. To convey the information easily, our system hide the complexity of information. Our web interface of e-learning can clearly present the information; web pages can be operated mainly by a few buttons, including start, stop, next article, previous article, auto-play, and download. Except for using the mouse to manipulate the web, it can still be operated with the keyboard. As mentioned earlier, podcast must be pre-made by human beings. Compared to podcast, information of ODES are automatically acquired from RSS, so the cost is lower than podcast. TABLE I is the differences between podcast and ODES.

TABLE I THE DIFFERENT BETWEEN PODCAST AND ODES.

Difference	podcast	ODES
File output format	MP3	MP3
Input file format	MP3	Text format file
Program Quantity	determined by the number of recorded	determined by the number of RSS
download	Yes	Yes
Portability	High Portable	High Portable
Human Demand	Middle	low
Subscribe	Yes	No
Timeliness	On Demand	On Demand
Equipment Requirements	support Podcast software	network

V. CONCLUSION AND FUTURE WORKS

The system implementation, focusing on completing the following practical benefits:

- Learning anytime at anywhere.
- Assisting foreign language learning.
- Saving time from getting information.
- The easiest way to get the real-time information.
- Providing the customized area.
- Easier and more convenient way for users to learn.

Combining the above benefits, ODES is a very useful tool for us to acquire information or to learn foreign language. For example, if a person wants to teach himself English, he can find his most interested article in ODES, or upload articles himself. The system will automatically search pictures related to his article, enrich the original material. TTS engine can read this article for him, which he can listen to the practical pronunciation and imitate the speaker. He can also choose the voice and adjust the speed of reading according to his own level or habits. Our system offers a basically different idea which is customized and built as per the requirements of users not the one-way information from providers. This point will greatly motivate users to use our system. We also set up some buttons to help operating our system, such as button to play, replay, consistently play and if he wants to keep the audio in his own devices, then he click the save button and save the mp3 files to his notebook, mobile phone, and mp3 player. Therefore, ODES makes the learner learn anytime, anywhere, easier and more convenient.

For the future, the site has the following plan:

- Build a customized voice area, according to the demand of every individual to build a more customized learning area, save personal settings and customized content, facilitate the convenience of management. Combine voice and customized RSS.
- Except for the above benefits of this system, the system will be more emphatic on functionality and humanized interface, allowing users more easily to apply our system at the first time access.

Face to the era of information explosion that our technology is uninterrupted updating, whoever get the latest information will be the one catch the opportunity. Therefore humans keep on changing the way of learning and the method of acquiring news. Our system provide a new pattern of learning news and education, which reduces the time people spend on looking for the information they need and introduce a brand new way for people to change their behaviors from being passive to active. We believe it is the new mark of intellectual life and online learning.

ACKNOWLEDGEMENT

The authors would like to thank the National Science Council of the Republic of China, Taiwan R.O.C, for financially supporting this research under Grants NSC99-2218-E-040-001.

REFERENCES

[1] Wikipedia. E-Learning2.0. http://en.wikipedia.org/wiki/E-Learning_2.0. Retrieved 2011-01-27.

[2] S. Britain and O. Liber. A Framework for the Pedagogical Evaluation of eLearning Environments, JISC, pp. 322–383, 2004.

[3] CEO forum. E-learning history. http://www.ceoforum.com.au/article-detail.cfm?cid=10348. 2008. Retrieved 2010-09-24.

[4] J. S. Liu. A Probe into the RSS-based Online Learning, vol. 6, pp. 62–64, December 2007.

[5] O'Reilly. What is Web 2.0 Design Patterns and Business Models for the Next Generation Software.

http://www.oreillynet.com/pub/a/oreilly/tim/news/2005/09/30/what-is-web-20.html. 2005. Retrieved 2011-01-24.

[6] Q. D. Ding. Research on the New Pattern of Internet Based on Web 2.0. Graduate Degree Thesis of Beijing University of Posts and Telecommunications, pp. 4–5, March 2006.

[7] S. Ma and M. He. E-Learning Based on Web2.0 Technical Characteristics. International Conference on E-Business and E-Government, pp. 1–3, 2010.

[8] D. Winer. Rss 0.93 specification. http://backend.userland.com/rss093, 2001. Re trieved 2010-09-12.

[9] D. Winer. Rss 2.0 specification. http://cyber.law.harvard.edu/rss/rss.html,2003. Retrieved 2010-09-14.

[10] IETF. Xml media types, rfc 3023. http://tools.ietf.org/html/rfc3023#section-3.2, 2001. Retrieved 2011-02-21.

[11] D. Kohn, M. Murata, and C. Lilley. Internet drafts: Xml media types. IETF InternetDraft, 2009.

[12] S. Carliner. Designing e-Learning. American Society for Training and Development, Virginia, 2002.

[13] W. Liu. Ontology in E-Learning System Application Study [D]. Master's degree thesis at Northwestern University in 2006, 2006.2-3.

[14] C. C. Chang and K. C. Hsiao. An e-Learning System for Information Management Education Based on Web Services, International Conference on Computational Aspects of Social Networks, pp. 48–51 2010.

[15] G. Ruan and Y. Lin. Peer Assisted Learning Model and Its Application in Construction of Peer Assisted Online Learning Community, China Educational Technology, vol. 11, pp. 34–37, November 2006.

[16] T. Tian. Online Learning Environment Design Based on Multi-intelligence Theory, Journal of Yangzhou University (Higher Education Study Edition), vol. 11, pp. 66–68, November 2007.

[17] S. Downes. E-Learning 2.0. http://www.downes.ca/post/31741. October 2005. Retrieved 2011-03-24.

[18] S. Wu, C. H. Song, H. D. Chen, and J. M. Zhan. E-Learning Teaching method Research and Ideas, International Conference on Web Information Systems and Mining, pp. 240-243, 2010.

[19] C. Ullman and L. Dykes. Beginning Ajax. Wrox, 2007. W3C. World wide web consortium. http://www.w3.org, 2010. Retrieved 2010-12-27.

[20] E. Pascarello, D. Crane, and D. James. Ajax in Action. Manning Publications, 1 edition, 2005.

[21] Oup.com. Oxford university press — podcast. http://www.oup.com/elt/catalogue/teachersites/oald7/wotm/wotm_archive/podcast?cc=global. Retrieved 2010-12-27.

[22] Y. H. Ma. Research on Individualized Online Learning System Based on Web Mining, Journal of Henan Agricultural Sciences, vol.3, pp. 40–41, February 2007.

 ISBN: 978-1-61208-139-7

[23] Q. Li, W. Rynson, H. Lau, W. C. Elvis, F. Li, V. Lee, B. W. Wah, and A. Helen, Emerging Internet Technologies for E-Learning , IEEE Computer Society, pp.780-1089, September 2009.

[24] S. Hiltz. Online Education: Perspectives on a New Environment New York: Praeger, pp. 133–169, 1990

[25] T. Dutoit. An Introduction to Text-To-Speech Synthesis. Springer, 2001.

[26] P. L. HeGaret. The w3c document object model (dom). http://www.w3.org/2002/07/26-dom-article, 2002. Retrieved 2011-01-25.

[27] M. A. Bodie. The Book of SAX: The Simple API for XML. No Starch Press, 1 edition, 2002.

[28] D. Brownell. SAX2. O'Reilly Media, 2002.

[29] Microsoft Corporation. Microsoft speech api (sapi) 5.3. http://msdn.microsoft.com/en-us/library/ms723627(v=vs.85).aspx, 2010. Retrieved 2010-12-28.

[30] J. Murach and A. Steelman. Murach's Java Servlets and JSP. Mike Murach & AssociatesInc, 2 edition, 2008.

 ISBN: 978-1-61208-139-7

A Net Courseware for "Image Processing"

Yu Jin ZHANG
Department of Electronic Engineering
Tsinghua University
Beijing 100084, China
zhang-yj@tsinghua.edu.cn

***Abstract*—To help the teaching and learning of the undergraduate course named "Image Processing", and to take the advantage of the progress of network for improving the education efficiency, a new and particular corresponding net courseware has been developed. This courseware is developed to support the education, to provide a vivid complement for the textbook, while not to replace the textbook. This paper introduces some design considerations and principles related to pedagogical requirements, the framework and structure for the organization and arrangements of contents, as well as the characteristics of the web-pages and interfaces of the courseware. In addition, some typical web-pages of this net courseware are shown to illustrate the results of different efforts in the design and development. Parts of the courseware have been used by professors and students, and some good feedbacks have been obtained.**

Keywords-computer assisted instruction, net courseware, knowledge component, image processing.

I. INTRODUCTION

Computer Assisted Instruction (CAI), as a powerful technique, plays an important role in modern education. The progress of computer and communication technology as well as multimedia equipments and network infrastructures have greatly enhanced the quality of teaching and learning in recent years [1].

One of the most important tasks to make educational learning effective is the designing of courseware. From the pedagogical point of view, the use of courseware can greatly thrust the interest of learner and stimulate their subjective motivation, which would have significant influence both on the course conducting performance and educational effects.

Different media, such as image/picture, animation, audio, and video have been employed in various coursewares to give students some vivid and dynamic representations to stimulate their deep thinking and to push student's motivation in active participation.

With the wide utilization of networks, net coursewares become popular. They organize the education materials with the help of well organized web-pages [2]. Learners can use network technology to access remotely the contents in courseware. Net courseware becomes an important educational resource. In this paper, a net courseware has been used for 3 years are presented.

"Image Processing" is an important course in image engineering [3]. As a basic course for the discipline, it is conducted for the third year undergraduate students in our university. To help the teaching and learning, the development of a corresponding net courseware has been taken. Three key factors in developing such a courseware are the content selection, the courseware structure, as well as the interface and web-page layout.

The construction of web courses should take into consideration, besides the general pedagogical demands, the particularity of course and learning style, satisfy the condition of operating with the help of Internet, and profit from the advantage of a network environment.

The rest part of this paper is organized as follows. Section 2 introduces the selection criteria in choosing the relative contents. Section 3 gives the framework for organizing the contents and media in the courseware. Section 4 provides a number of typical web-pages in courseware to show the layout and form of interface and contents. Finally, Section 5 summarizes the good points of the courseware.

II. CONTENT SELECTION

The content of courseware is the core of the courseware. Some criteria should be respected. The current trend in the development of discipline under question should be reflected in selecting the sources of information. The courseware should cover the subject matter of the course. In one side, it should correspond to the substance in the textbook. On the other side, it should be complementary to the contents of textbook.

The selection of contents for this courseware has been made in accordance with the goal and function of this course. As this course is given in campus for full time students with a textbook specially written for it, the courseware is designed as an assistant to help the students for extended self learning and reviewing. The following considerations and actions have been made:

- To help the process of course conduction, the material selected and the script designed should be correspond to the materials in the textbook. Taking this factor into consideration, the number of chapters in the courseware has been made to be equal to the number of chapters in the textbook. In addition, for each chapter, the correspondence between the

 ISBN: 978-1-61208-139-7

sections in the courseware and the sections in the textbook is also maintained.

- There is a significant difference between self-learning and the learning by following lecture. In addition, the first learning and the reviewing after learning are also quite different. Therefore, the courseware should provide some variations and supplements. From one side, the courseware can use its advantage in capacity to provide more examples and even more materials than printed textbook. From other side, based on the existing explanation in the textbook, the courseware can just list the compact definitions for a number of concepts to facilitate the sum up and remember.
- The modern theory for education and psychology emphases on the various learning styles to help acquiring knowledge. It is believed that combining different learning styles would greatly improve the efficiency of learning. Considering this factor, the courseware provides a number of demonstrations for illustrating the procedure and for displaying the process results. Such a dynamic manner provides a vivid impression.

III. COURSE STRUCTURE

The structure of courseware is related to the internal organization of contents. From one side, the structure should match the strategy of education, connect different course and learning parts, and make the study easy. From other side, the structure should fit the network environment and the characteristic of computers.

The contents of courseware are to be organized in modules. In particular, it is better to use knowledge component (KC) as a learning/studying unit, and to arrange contents around knowledge points.

The structure of course has provided a guide to learners. A dynamic and layered structure supporting association of related KCs is important. These KCs should be linked to support cross-reference, which is normally enabled by numerous inter-links between related text notes and demonstrations [4]. However, getting lost in hyperspace should be avoided [5].

The following measures have been made:

- A combination of dynamic tree-structure and graph-structure has been adopted in the courseware. The tree-structure can provide a layered outline for the courseware as well as a fast and logical access route for different branches. The graph-structure can provide a cross-grid relationship for the content, and make the browsing and navigation inside courseware easy and intuitive.
- A module implementation has been conducted for the concrete contents. The section has been taken as the logical module of courseware, which has been named “knowledge component”. Each knowledge component corresponds to an independent learning unit, which consists of four groups of contents corresponding to specialized knowledge: the Concepts (for principles and definitions), the Examples (for processing methods and explanations), the Demonstrations (for dynamic process and result displaying), and the Key-points (for summary and references). Each knowledge group has been made quite independent, to perform connected study.
- The web-page has been taken as the physical unit/module corresponding to the section of courseware. The contents in web-page are arranged according to the lecture/study order. On the web-page, not only some concrete contents are directly shown (for compactly representing the concepts and key-points), but also other related contents (examples and demonstrations) are connected by using the hyperlink functionality of web-pages.

The overall framework of the courseware is shown in Figure 1. It can be seen from Figure 1 that the learners have two choices once getting into the courseware. The first is to select the chapter from the Chapter List, and follow the pre-determined learning order. The second is to go to the classified KC lists. From the outline to knowledge component, a tree-structure provides an access mode like using the table of content in a book. While from the list to learning term, a graph-structure supplies a visit mode like using the index of a book.

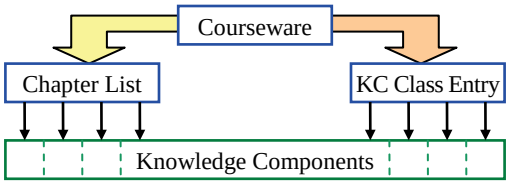


Figure 1. Framework of courseware.

Figure 2 shows the access paths from KC Class Entry to individual KC, via respective lists. All knowledge components are classified into four groups (modules): Concepts, Examples, Demonstrations, and Key-points. The titles of different groups of knowledge components are listed in four corresponding lists. From the concept list, learners can access the definitions of several hundreds of concepts. From the example list, learners can view more than hundred examples with pictures or drawings. From the demonstration list, learners can manipulate a number of dozens of dynamic illustrations. From the key-point list, learners can look over the summary of each section and related reference introductions for further study.

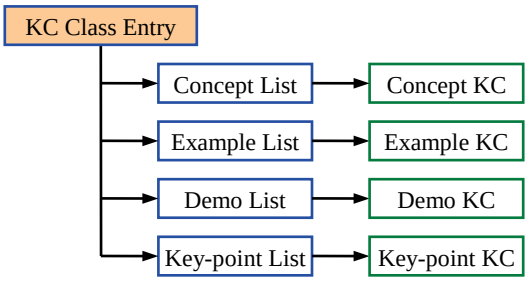


Figure 2. Access from KC class entry.

 ISBN: 978-1-61208-139-7

Figure 3 shows the organization of a chapter in courseware. Getting into a chapter, the summary of this chapter, with four title lists for concepts, examples, demonstrations and key-points in this chapter, is displayed. Beside, a list of sections in this chapter is also provided. A section is a study unit that is formed by a number of KCs arranged in learning order.

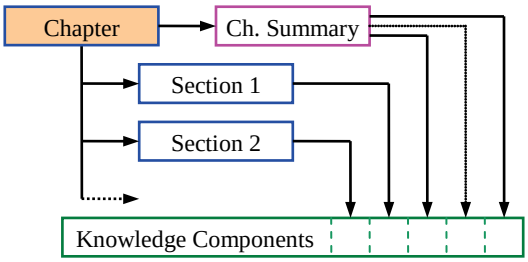


Figure 3. Chapter organizations.

IV. INTERFACE AND WEB-PAGES

The interface of net courseware provides a way for communication between the computer and users. The net courseware is composed of a number of linked web-pages, in which the interface for the learners to use the net courseware is resided. The web-pages convey the information to the learners and receive the control from the learners via interface.

Some careful considerations in the design of web-pages are as follows:

- The form and appearance of courseware should take the advantages of computer capacity, provide intuitive visual effect, encourage the participation of learners in the studying procedure, and make the abstract concept more concrete and more visible to help the understanding.
- A flexible and user-friendly interface is an important criterion to judge the usability of courseware. It should provide fast access to different parts of the courseware. To fulfill this requirement, the hyperlinks between the four lists and the study unit have been established as shown Figures 1, 2 and 3.
- Interactivity plays an important role in courseware. Many interactions among different forms, such as image, video, animation, dialog box, have been included. The styles for each knowledge component and the operating interface have been unified.
- In general, different interfaces in the web courseware should be concise (such as the interface layout), uniform (such as the window locations) and easily operable (such as the keys for interaction). Thus, users could be concentrated more on the course contents themselves than on the learning of how to use the interface.
- Dynamic demonstrations are added in courseware, which would be impossible in textbook. Many tools exist, such as Adobe Acrobat, Macromedia FLASH, Microsoft EXCEL, MATLAB, LabVIEW, and WebCT [6]. By considering various factors such as the security, the data volume, the quality of products, and the effect of demonstration, the demonstrations have been developed with Macromedia FLASH and embedded in web-pages.

Some typical web-pages are given below for examples:

- Web-page for courseware entry (Figure 4)

This page has been divided into three regions: on the left is the table of contents which gives the titles of chapters and sections in order; on the top right are the hot-key for four class lists (Concepts, Examples, Demonstrations and Key-points); and on the low right is the content display region which provides an introduction to the courseware. The top part of Figure 1 is implemented in this page.

Figure 4. Web-page for courseware entry.

- Web-page for KC class entry (Figure 5)

This page will pop out after clicking the concept list on the top of the web-page for courseware entry. Such a page displays the list for concept titles (here the list for concepts is an example, similar pages with other lists can also be obtained), and the definition of every concept is linked behind, as shown in Figure 2.

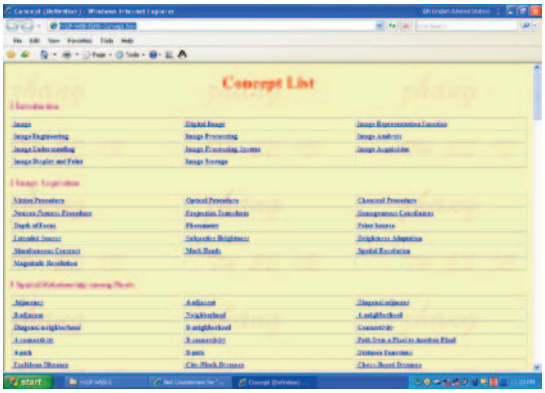


Figure 5. Web-page for list entry.

- Web-page for chapter entry (Figure 6)

This page appears after clicking the chapter title on the left of the web-page for courseware entry. The region on the low right provides a summary of the chapter by giving the title lists of each knowledge component in four groups. From here, all knowledge components in this chapter can be randomly accessed as shown in Figure 3.

 ISBN: 978-1-61208-139-7

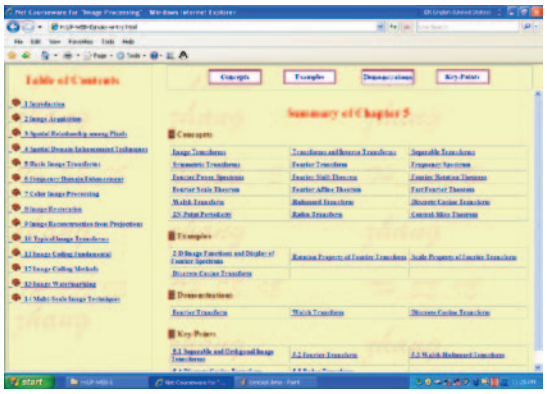


Figure 6. Web-page for chapter entry.

- Web-page for section and content (Figure 7)

This page appears after clicking the section title, which will appear after clicking the book-logo at the left of chapter title (the closed book-logo changes to opened book-logo) on the left of the web-page for courseware entry. From the section list, a selection of a particular section will provide the contents of this study unit on the low right window (as shown in Figure 3), in which different knowledge components are arranged according to the learning order.

Figure 7. Web-page for section and content.

Note in Figure 7, the contents of concepts and key-points are directly displayed, while the contents of examples and demonstration will be shown in new popup windows.

- Web-page for examples (Figure 8)

This page will pop out after clicking the example title in the low right region of web-page for section and content. It can also be obtained if one clicks the same title from the example list on the top right of the web-page for courseware entry, or from the example list in chapter entry (see Figure 6). Examples are illustrated by a number of pictures showing the procedure and the result of particular techniques.

It is often said that “one picture is worth more than ten thousand words”. Using pictures for teaching and learning image courses has great advantages [7], especially for showing pictures on the computer screen with courseware, as the quality of images is much better than that on the printed textbook, and the details caused by noise and the removing effect, for example, could be readily visible.

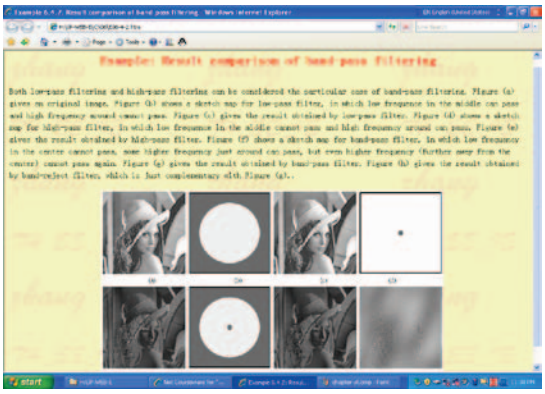


Figure 8. Web-page for examples.

- Web-page for demonstrations (Figure 9)

This page will pop out after clicking the demonstration title in the low right region of web-page for section and content. It can also be obtained if one clicks the same title from the demonstration list on the top right of the web-page for courseware entry. Demonstrations are made by using Flash to achieve interactivity between learners and the courseware. Figure 9 shows a demonstration for color processing. Learners can use the function keys (forward, backward, circle/rewind) in the low right to control the process order and can also click on the web-pages to start the interaction with the demonstration.

Figure 9. Web-page for demonstrations.

Various demonstrations have been developed. Some of them display the dynamic drawing procedure; some of them provide the image operation gradually with sequential pictures, and some of them interactively and selectively perform actions according to learner’s request. Learners could not only view the middle results but also get more impression on the course and flow. Those demonstrations can be used by teachers on the class lecture and by students after the lecture for self-learning.

V. CONCLUDING RMARKS

A new net courseware for “Image Processing” has been designed and developed. The three important factors, the selection of content, the course structure, and the interface and web-pages, have been explained in this paper.

 ISBN: 978-1-61208-139-7

The design process was performed before the start of the course, and the development was made in similar pace with the conduction of the course. Therefore, the professor and students have already the chance to use it once and to provide some useful feedback for improvement.

Some feedbacks from students are: with this courseware, they feel that the course and the subject are more interesting, some abstract concepts become easily understanding, and the emphases are simple to catch in the reviewing.

Some feedbacks from teachers, related to students side, are: less students pose questions on the concepts and process of techniques after class, and more students accomplish their homework with more visual forms [7] (some of them are made by students with their own programming) now.

Some good points of this net courseware include:

- Comprehensive contents: Cover the whole textbook and provide complementary information.
- Combined structure: Both layered and cross-linked structures are employed.
- Uniform outer shell: All materials are embedded into web-page and can be viewed by a net browser
- Interactive demonstration: Push learner's learning interests and motivate learner's visualize thinking

The principle and techniques used in this courseware should be also suitable for use in other coursewares.

ACKNOWLEDGMENT

This work has been supported by the Grant "TH-985-Second Term of Textbook and Teaching Resource (TH-985-STTTR)".

REFERENCES

[1] Y. S. Chee. "Networked Virtual Environments for Collaborative Learning," Proc. of International Conference on Computer in Education'01, 2001, Vol.1: pp. 3-11.

[2] J. Tiffin. "The HyperClass: Education in a Broadband Internet Environment," Proc. of International Conference on Computer in Education'02, 2002, pp. 23.

[3] Y. J. Zhang. Image Engineering: Processing, Analysis, and Understanding. Cengage Learning, 2009.

[4] Y. T. Sung, S. K. Chiou, and K. E. Chang. "Use of Hierarchical Hyper Concept Map in Web-based Courses," Proc. of International Conference on Computer in Education'01, 2001, pp.1133-1137.

[5] M. Hall, and D. Robinson. "Lost in Hyperspace: Linearity Versus Exploration in the Design of Multimedia for Independent Learners," Proc. of International Conference on Computer in Education'98, 1998, Vol.2: pp. 9-13.

[6] N. Shahnam. "Application of Technology in Engineering Education," Computers in Education Journal, 2006, Vol.16, No.3: pp. 100-111.

[7] Y. J. Zhang. "Teaching and Learning Image Courses with Visual Forms," Encyclopedia of Distance Learning, 2nd Edition,, 2009, Vol.4: pp. 2044-2049.

 ISBN: 978-1-61208-139-7

New Methodology for Developing Digital Curricula

Nahla El Zant El Kadhi
Management Information System Department
Ahlia University
Manama, Kingdom of Bahrain
nahla@ahliauniversity.edu.bh

Hanaa Al-Sharrah
Information Technology Department
Ahlia University
Manama, Kingdom of Bahrain
hanaa17@hotmail.com

***Abstract*-- The vast and rapid development in the computer, communication and Internet technologies has significantly affected contemporary educational systems. This paper describes an approach for the deployment of an e-learning infrastructure/environment in addition to a new methodology of digital curricula development for an efficient E-learning implementation and adoption. Hence, the purpose of this research is to study and analyze the role of implementing the E-learning framework and online digital curricula in K12 schools. Success factors of implementing the E-learning framework and project management perspectives of online digital curricula development are also presented and discussed.**

Keywords-E-learning; e-content; learning design; blended learning; e-learning model; web-based educational system.

I. INTRODUCTION

Recent years have witnessed an enormous growth in multimedia applications and telecommunication technology like audio/video conferencing, and live video streaming, to name a few. Education technologies have been developed in the recent years where new technology-based learning methods and channels have been emerged. This paper reports the main components of a successful E-learning and introduces a new framework and model to implement E-learning in K12 schools. It becomes widely known that the national initiatives in the region to implement different kinds of E-learning technologies in the public schools struggle in most the cases. This is due to the fact that such initiatives usually did not look at the "big picture" of E-learning projects and did not implement an integrated framework. Hence, the new framework is proposed to ensure the integration of all the components and success factors of an efficient E-learning system. This paper introduces an enhanced methodology of developing the online digital curricula (e-content). This process mainly handles the digitizing of the conventional curricula into an interactive digital one. Creating an online content is a challenge that faces E-learning adopters. One of the most important elements in creating online content is the level of interactivity with the user to keep him attentive all the time. This paper introduces an instructional design model based on Bloom's levels and the first principal. The evaluation process is based on Kirkpatrick's model.

II. LEARNING MODEL

This section presents the learning model that represents the foundation to understand the proposed framework and blended learning for K12 schools. The model is based on the interaction between the instructors, the learners and the content. According to Spiro [6], cognitive flexibility is the "ability to spontaneously restructure one's knowledge in many ways". The proposed framework aims at enhancing the learning process by providing a better learning environment that "blends" the educational technology with the conventional face-to-face environment. The interaction between the three main components in the presented learning model is enhanced by using the technology. Hence, this model is based on the blended learning type. This "Blended learning programs may include several forms of learning tools, such as real-time virtual/ collaboration software, self-paced Web-based courses, Electronic Performance Support Systems (EPSS) embedded within the job-task environment and knowledge management systems" [4].

III. DELIVERY ENVIRONMENT

Al-Sharhan introduced a delivery model for the new E-learning environment as explained in [7]. The elements of the environment are the learning management system, multimedia equipped classrooms (smart classrooms), and network or the Internet. The instructor guides the learning process by utilizing the online content where students access the content via the Internet. The LMS tracks the learning activities and provides the instructor with report about the learning process. Fig 1 depicts this model which was mainly introduced by Al-Sharhan in [7] and we have added to it the external environment.

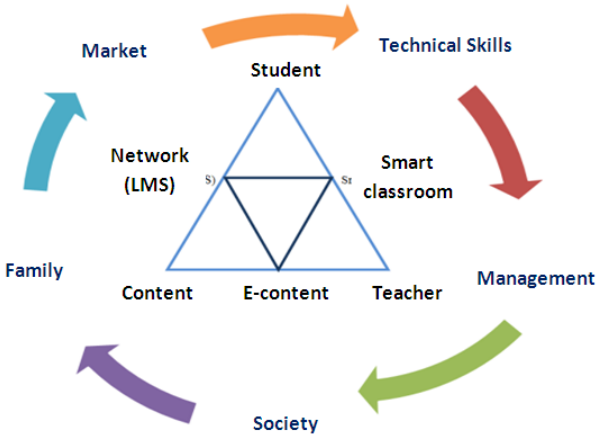


Figure 1. A new E-learning Model.

Fig 2 depicts the E-learning projects incorporated in full-fledge E-learning implementation in K12. All these projects are highly interconnected where the failure in one chain may cause serious problems in the whole implementation.

 ISBN: 978-1-61208-139-7

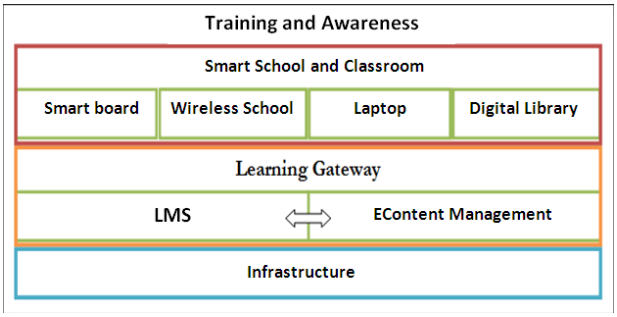


Figure 2. E-learning project.

IV. E-LEARNING INFRASTRUCTURE

The Infrastructure projects aims to provide a high-performance data center in the head office and the required computing devices in the schools. It also provides all the network facilities in the schools and Head office to work in both centralized and decentralized manners. The data center should provide a collection of virtual and physical devices to support the higher layer of applications such as the learning gateway and portal, Learning Management system, collaboration tools, disaster recovery, security and other applications. The design of the data center should be scalable since the E-learning projects usually implemented in phases. With the time, complex and high cost devices are added to incorporate more users and applications. A successful and scalable design of data center network must have, at minimum, Scalability, Simplification, Sharing, and Security.

V. DATA CENTER DESIGN

Today's data centers may contain hundreds of computers with significant aggregate bandwidth requirements. The network architecture typically consists of a tree of routing and switching elements with more specialized and expensive equipment moving up the network hierarchy. During the design of the data center, the designers should pay attention to the increasingly growing applications that possesses special requirements. In today's data center design, the typical architectures consist of either two- or three-tiers networking routing. A three-tiered design usually has a core tier in the root of the tree, an aggregation tier in the middle and an edge tier at the leaves of the tree. A two-tiered design has only the core and the edge tiers. However, in modern design for new applications N-tier design should be considered to support the applications with multi-tier. While the N-tier logic development is suitable for large, medium, and small-scale applications, and Web applications, it can also enhance system application security, performance and scalability for future expansion. The Advantages of utilizing N-Tier architecture can be summarized as follows:

- Security of data and application can be easily maintained.
- Business rules will be separated into a component that is easy to maintain, use, and reuse.
- The architecture supports high scalability and expansion where every tier is independently maintained and expanded without affected the components of other layers and tiers.
- Data storage is centralized into an independent layer for more efficiency coding and implementation.
- The architecture provides flexibility where components can be distributed to different physical machines at any time.

VI. SMART CLASSROOM

The Smart Classroom project aims at applying smart technologies in a physical and conventional classroom. The project bridges the gap between modern technology-based and traditional classroom activities in terms of the teacher's and student's experiences. More specifically, the components of the smart classroom enable the teacher to utilize modern technology to enhance the teaching experience. All the teaching activities can be recorded and hosted on the learning gateway for future consideration. In addition, the smart classrooms components will provide the teacher with efficient tool to manage the class and provide the students with an exceptional teaching and learning experience.

Smart Classroom Components: The Smart Classroom system contains several component technologies that make the interaction between the teacher and students efficient and transfers the role of the teacher to be a facilitator of the learning process. In the proposed E-learning framework the main components are, smart wireless interactive board and data projector, teacher station, laptop for each student, class management system, and wireless network to connect the laptops to the learning gateway. Teachers in the Smart Classroom can utilize different smart components inside the class or freely using conventional teaching methods to instruct students in a blended methodology.

VII. WHAT IS E-CONTENT

E-content can be defined as the process of digitizing the conventional learning subjects and transferring them into interactive multimedia based subjects. Creating an online content is one of the serious challenges that face E-learning adopters. the most important elements in creating online content is the level of interactivity with the user to keep him attentive at all times. When building and developing the e-content one should keep eyes on four important issues; namely:

- The Instructional Design process.
- The standards of building the e-content.
- The sharable Learning objects repository.
- The technology used in the development.

The e-Content is designed using sequencing of learning objects (LOs) or Sharable Content Object (SCO). SCOs are small chunks of information, indexed Meta-Data, self-contained and explanatory, reusable, have aggregation capability, and communication capability with LMS. In this paper, we define the SCO to have the following parts: the Learning Objective, concept explanation, exercises, and evaluation. SCOs are transferred into a digital format according to specific standards. SCOs will be stored in shared Learning Object Repositories (LOR). The course will be added to the Learning Management Systems (LMS). This concept is explained in Fig 3, while Fig 4 explains the relation between the SCOs and the online course.

 ISBN: 978-1-61208-139-7

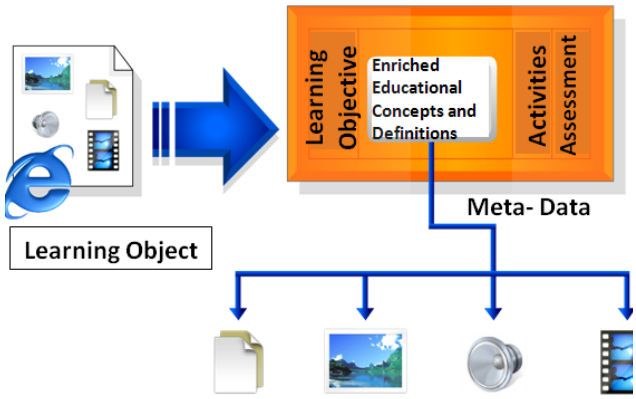


Figure 3. Learning Object Concept.

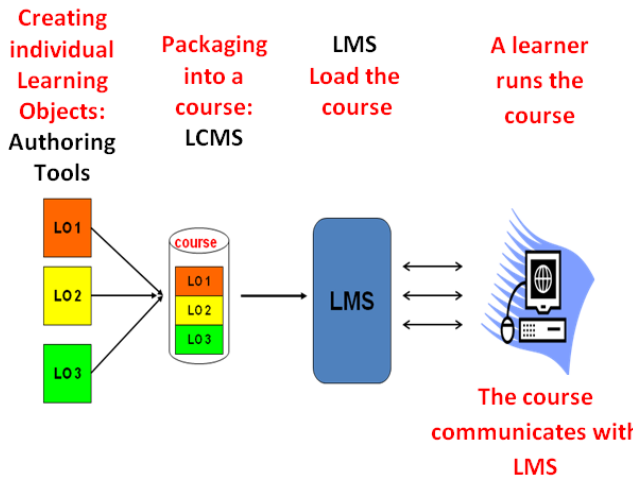


Figure 4. SCOs and e-content.

VIII. CULTURAL LEARNING OBJECT

Cultural factors are very important for a successful E-learning implementation. Culture, however, is a broad and concept with no unified definition. Fernandes believes that cultural issues and considerations are something in which people take pride, and that it must be considered and respected in the user interface [3]. Such considerations are importance because cultural design will touch several areas of a users' culture consciously and unconsciously. Hence, cultural considerations increase the complexity of designing learning objects and e-content interfaces because more variables are added. Several cultural factors must be considered during designing and developing the learning objects and the e-content [1].

IX. ONLINE COURSE DEVELOPMENT

The phases of building online courses are as following:

- The content is divided into small chunks as raw LOs it should be before; this is performed by SME (Subject Matter Experts) in collaboration with ID (Instructional Designer) according to an agreed scheme of course structure.
- The LO Storyboarding takes place by the instructional designer.
- The storyboard is sent to the production unit –multiple production lines exist.
- LO is developed according to the storyboard which was authored by ID and under supervision of SME.
- LO is produced according to E-learning standard (IEEE-LOM) in the form of a SCO (Sharable Content Object)
- LO is stored on a learning object repository (LOR).
- LOs are packaged according to an agreed scheme of course structure which forms the standard SCORM 2004 courseware. Courseware is uploaded and stored on Learning Management System (LMS)

Fig 5 describes the life cycle of online courses development.

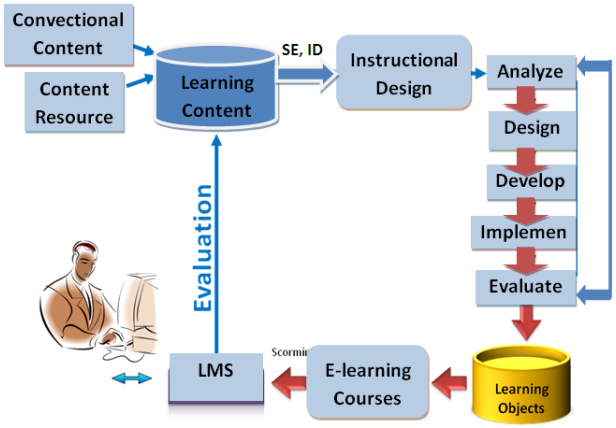


Figure 5. The life cycle of course development.

Courseware Development Methodology through Learning Objects approach goes through two major stages, an initial stage followed by execution stage, each has its own phases.

The initial stage is considered a preliminary stage required for courseware development. A large scale analysis takes place to identify: course materials, learning goals and objectives. Audience analysis is carried out to identify their learning styles and capabilities. The instructional deign strategy is determined according to the results of analysis, and then the course outline (structure) is introduced. Educational material is divided into small chunks of information and data "raw learning objects". Then the interface and templates design and development for its functional requirements is carried out. Eventually, a complete production and development for a prototype of an entire learning object is implemented. Such prototype is introduced to the concerned stakeholders for review and approval, and upon all that the execution phase is to be launched. The initial stage comprises of the following subsequent sequential phases: the analysis phase, design phase and the prototype development phase as shown in Fig 6.

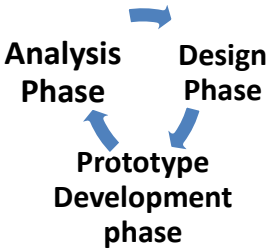


Figure 6. Initial Design.

The execution stage is considered the real start of a large scale production of learning objects. It starts right after MOE stockholders approves the prototype that gives a general and future vision of what to be expected from the execution stage.

 ISBN: 978-1-61208-139-7

The prototype leads to minimization of risks to their lowest rates. The execution stage composes of subsequent sequential phases; those are storyboarding phase, development phase and packaging phase shown in Fig 7.

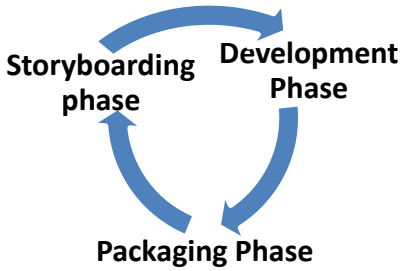


Figure 7. Execution phase.

X. EXISTING METHODOLOGY TO BUILD WEB-BASED EDUCATIONAL SYSTEM

Success factors in an E-learning system are related to students including the disable, instructors, the developed course content and the delivery medium and environment. A new framework for the developing and delivery of the online content or courses is presented and mapped to the activity theory. This mapping explains and organizes the interactivity relations between instructors, students and online learning objects as they are bounded by the technology, Learning management system (LMS) and educational and external environments. ADDIE Model is one of the methodologies used to build educational web that includes 5 phases [10]:

- Analyze - analyze learner characteristics.
- Design - develop learning objectives, choose an instructional approach.
- Develop - create instructional or training materials.
- Implement - deliver or distribute the instructional materials.
- Evaluate - make sure the materials achieved the desired goals.

XI. E-LEARNING EVALUATION

Learning and online learning evaluation mainly deals with the education quality and its management. This process is part of a broader education evaluation process because of its practical issues and aspects in addition to its complicated theoretical approach. The evaluation content includes evaluation of learning environment, evaluation of learning style, evaluation of comprehensive ability and so on. However, the evaluation of the learning objects and learning process in general is not an easy process due to the pedagogical factors involved in it. In addition to the lengthy process, learning evaluation is inherently complex. This is due to the fact the several complexity factors are associated with the sustainable and dynamic interactivity of the various dimensions of the learning process such as learning goals, instructors, learners, and instructional technologies. Hence, a flexible framework of evaluation is required in order to realistically evaluate the E-learning outcomes. The evaluation model or framework must incorporate all the issues related to the learning process. The performance evaluation of E-learning is generally divided into the formative evaluation and the summative evaluation. Formative evaluation is performed by experts during each phase of the implementation progress of E-learning. Based on the evaluation result, an immediate action is taken to improve the current and future implementation, to improve the final results. Summative evaluation, is the evaluation performing after that all the E-learning deployment is completed, to understand the final implementation results and compare the difference of learning effectiveness of students and subjects before and after the implementation. From a pedagogical point of view, E-learning performance evaluation is considered to be purely formative evaluation. This means that the focus of evaluation shifts from the learning results to the learners' learning process. This basic idea of E-learning performance evaluation has been widely accepted by the education community [2]. It is very important that the pedagogical approaches are evaluated while creating E-learning content. Simple pedagogical approach will lack flexibility and approach. On the other hand, a complex approach will take time to develop and consume setup costs. Therefore, a perfect pedagogy should adopt the middle path by creating effective educational material while running parallel to engaging learning experience. In this work, the evaluation of the E-learning objects or the E-learning process is based on the Kirkpatrick's model. It consists of four levels, namely the reaction evaluation, learning evaluation, behavior evaluation and results evaluation. In Kirkpatrick's model, the evaluation process of learning or training program should always begin with first level, and then, should move sequentially through levels two, three, and four. The information from each level serves as the input or the base for the next level evolutional.

The first level which is the reaction evaluation measures how students in classroom react to the learning object and the online lessons and the perception of the students. The evolution of this level depends on answering questions regarding online lesson and learning objects acceptance. Example of these questions may include: Did they like a specific learning object? This type of evaluation can be conducted using an online survey, which is part of the LMS and learning gateway. In addition, the participants' reactions have important consequences for learning. The second level named learning evaluation aims at assessing the learning capabilities beyond learner satisfaction and targets assessing students' developments in terms of skills, knowledge, or attitude. Naturally, the measurement outcome at this level is more difficult than the first level. Hence, methods of the evaluation differs from the first level and range from testing techniques to group assessment and self-assessment. A useful technique here is to have a pre-test and post-test in order to determine the amount of learning that has accumulated. The third level is the behavior evaluation that measures the impact of the learning occurred by the online lesson on learners' behavior. It measure the amount of change or the transfer of the learners behavior due to the learning material and the knowledge accumulated in the online lesson or in a certain group of learning objects. For many educators the third Kirkpatrick's level evaluation represents the truest assessment of learning effectiveness. However, measuring at

 ISBN: 978-1-61208-139-7

this level is not an easy task due to the fact that is almost impossible to predict when the change in behavior will happen. Hence, this level of evaluation requires important decisions in terms of when to evaluate, how often to evaluate, and how to evaluate. The fourth level of Kirkpatrick's model measures the success of the learning process in terms of improving the performance of the educational management.

XII. E-CONTENT DEVELOPMENT: E-CONTENT PROJECT MANAGEMENT

The E-learning project starts with an *initiation* phase and then *planning* for the project activities and tasks. Once the planning phase is completed the *execution* phase starts along with *monitoring and control* process to ensure aligning the execution with the plans. The project is finished with a *closure* phase. The project life cycle is depicted in Fig 8.

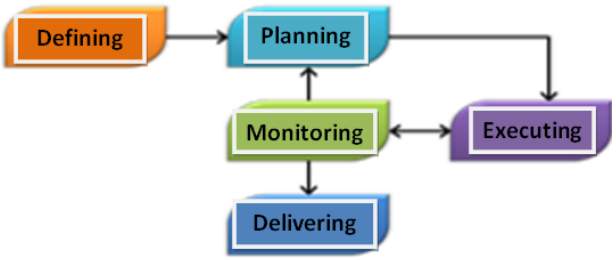


Figure 8. Project Life Cycle.

Defining / Initiation phase - The first phase in the e-content development is the initiation or the defining phase, this phase includes several activities and processes such as defining the scope of the project, scope planning and defining the work breakdown structure (WBS) (see Fig 9).

Defining the scope of the e-content project - The scope of the e-content project mainly concentrates on digitizing the curricula of the subject of year 10-12 in the ministry of education in Kuwait .The objective is to design the "online" course according to the best practices and international standards, namely, SCORM standard . Since this project is just member of a portfolio, the target is to host all the "online" courses on a "content" server and create online sessions based on clustering methodology that will distribute the different sessions on different cluster for good load balance. In order to get of the best of the E-learning the teacher should be able to track and follow his student's activities. Hence, the entire subject must be managed by a Learning Management System (LMS) that will create all the learning sessions.

Defining the Work Breakdown Structure - The Work Break Down Structure (WBS) is defined as a tool used to state and group a project's discrete work elements in a way that helps organize and define the total work scope of the project [9]. The main structure of our WBS for the e-content project specify that each learning object must reflect an integrated object that reflects a scientific concept. It should incorporates, Learning Objectives, Concept Explanation, Training exercises, and Evaluations. The Learning object also should be linked to more knowledge areas and external resources such as websites, external files and other teaching material to enrich the learning experience.

Planning Phase - the second phase in the project life cycle involves a set of processes and plans to help guide the project team through the execution, monitoring and closure phases of the project. The plans created during this phase help manage time, cost, quality, communication, risk and HR management. It also helps the project team to monitor, control, and manage the changes during the execution phase. Fig 10 illustrates the planning process.

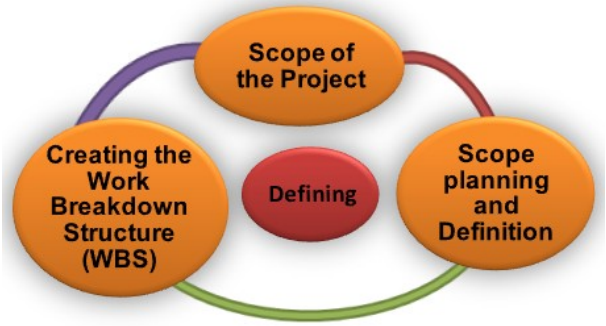


Figure 9. Initiation/Defining Phase.

HR Management - HR management process includes a variety of activities like internal search for employees who are willing to join the E-learning, or recruit experienced personnel, from outside. Selecting the internal staff requires evaluation of their knowledge, experience and performance levels. Training is to be involved if it will help the employee to handle situations. The required manpower for the content development are, Project Manager, Subject Experts, Instructional Designers, Graphic designers, Animators, Quality assurance engineers, Multimedia Developers, and Audio/Video specialists.

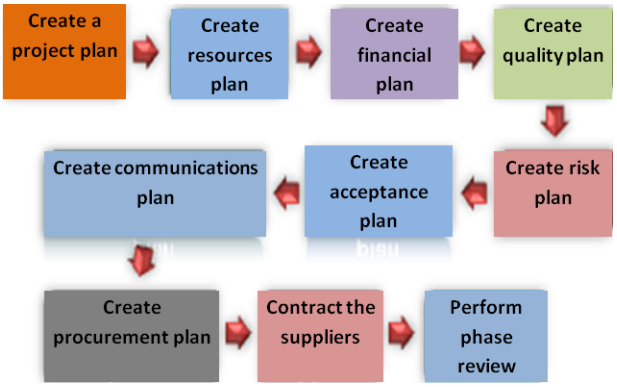


Figure 10. E-content Project Planning.

Risk planning - The e-content development project has several risks factors that need to be considered during the execution phase. The risk factors can be classified according to the following:

- Human related factors - availability of subject experts, the availability of the Instructional designers, and Developers.
- Technology based factors - tools to be used, Technological risks, and Incompatibility.
- Stakeholder Lengthy process - The approval process.
- Intellectual rights.

Quality planning - It is necessary to analyze various factors in order to deliver efficient teaching methodologies such as:

 ISBN: 978-1-61208-139-7

- Ease of use to implement the course and host it on the Learning Management system in order to be tracked and the tracking points must be clear for the teacher to understand the performance reports.
- Navigation: Students must not suffer of any difficulty to navigate in though the online course
- The amount of text in each electronic page.
- The amount of animation in each electronic page.
- Consistency, accessibility, interactivity & instructional issues, and content accuracy.

Managing the project - Change management process is a crucial one in each project since it ensures that the changes to the project scope, deliverables, timescales or resources are formally declared and handled prior even to the development. Applying and implementing change management is a difficult task; especially in large projects. To ensure that changes are monitored through to completion, a change management record and tracking procedure is maintained. This allows the project manager to identify any outstanding changes and to measure the actual impact of each change once implemented. The change record provides the project managers and teams with information about the changes conducted in the project at different levels. It can provides record changes within the project, monitor the change status and its project impact, record the status of all change approvals, identify and report on any change management issues, and control the amount of change required to meet your objectives.

Execution Phase - In this stage all courseware will be developed according to courseware development life cycle. Different processes are incorporated in Fig 11.

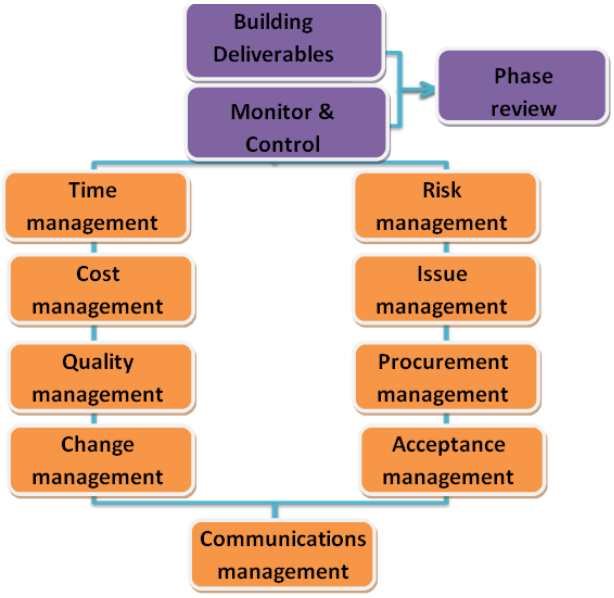


Figure 11. E-content execution phase.

Closure phase - Closing the project means delivering all the required deliverable according the agreed specifications. The project manager needs to make sure that the project closure criteria have been fully satisfied and that there are no outstanding items remaining. Also, he/she should identify a release plan for the project deliverables, documentation, supplier contracts and resources. Finally, the project manager must initiate a communication plan to inform all project stakeholders that the project has now been closed.

XIII. CONCLUSION

This paper presents a new implementation framework of the blended learning in K12 schools and a methodology for developing digital curricula or online courses. The proposed E-learning framework incorporates several components that are highly dependent in order to ensure an efficient implementation of the E-learning project. The evaluation and performance measure is important to guide the E-learning adopters on their directions and future impacts. The model incorporates four levels of evaluation, i.e., reaction evaluation, learning evaluation, behavior evaluation and results evaluation. However, each of the four levels presented by the model requires an empirical study to measure and evaluate the E-learning system. Another direction here is to work on an enhanced Kirkpatrick model based on fuzzy logic concepts. The main driver behind this direction is the fact that it is widely known that the first level of Kirkpatrick's model is most important level since it controls whether or not proceeding to the upper levels in the model. However one needs to design and conduct several evaluation forms, distribute them to the learners and do the analysis. This process is a lengthy one and time consuming. An enhancement is to build a fuzzy engine for the first level to be used by the teacher to evaluate the students' performance. Also the interaction between the teachers, students, smart class room technology is a very important area to be analyzed and thoroughly studied. Another area is the study of the motivations to efficiently utilize smart classroom and how to create a rich teaching environment in it.

REFERENCES

1. Al-Hunaiyyan A., and Al-Sharhan S. "Blended learning Design: Discussion of Cultural Issues." International Journal of Cyber Society and Education, 2008, pp. 17-33.
2. Dali H. "Design and Implementation of E-learning Performance Evaluation System." International Conference on Computer Science and Software Engineering, 2008, pp. 376-380.
3. Fernandes, T. Global Interface Design. Chestnut Hill, MA: AP professional, 1995, pp. 242-254.
4. Khan, B. "Building Effective Blended Learning Programs." Journal of Educational Technology, 2003, pp. 51-54.
5. Al-Fares M., Loukissas A., and Vahdat A. "A Scalable, Commodity Data Center Network Architecture." SIGCOMM'08, ACM, 2008 pp. 63-74.
6. Spiro R., Feltovich P., Jacobson M., and Coulson R.. "Cognitive Flexibility Constructivism, and Hypertext: Random Assess Instruction for Advanced Knowledge Acquisition in Ill-Structured Domains."1995, pp. 24–33.
7. AL-Sharhan S., Al-hunaiyyan A., and Gueaieb W. "Success Factors for an Efficient Blended learning." Proceedings of the 10th IASTED International Conference on Internet And Multimedia Systems And Applications. 2006, pp. 77–82.
8. Keller, J.M. "Development and use of the ARCS model of motivational design." Enscheda, the Netherland, 1983.
9. George H. Effective Work Breakdown Structures: the project Management Essential Library Series, ISBN 156726-135-3.
10. Molenda M., Pershing A., and Reigeluth M. "Designing Instructional Systems." The ASTD Training and Development Handbook 4th ed. , McGraw-Hill 1996, pp. 266-293.

 ISBN: 978-1-61208-139-7

Ontology-based Architecture for Reusing and Learning Through Context-aware Annotations Memory

Nadia Aloui/University of Sfax
MIRACL Laboratory
Sfax, Tunisia
Aloui_nadia@gmail.com

Faiez Gargouri/ University of Sfax
MIRACL Laboratory
Sfax, Tunisia
faiez.gargouri@isimsf.rnu.tn

***Abstract*— This paper describes a conceptual model and an ontology-based framework for reusing and learning through context-aware annotations memory. This memory manages annotations and adapts actors (learner, tutor, teacher, and co-author) behaviours with the various contexts of their activities. It offers a great reutilisability to share and have a better quality of learning. The annotation model that we propose is composed of three facets: cognitive, semantic and contextual. The architecture of our annotations memory contains many modules based on Web Services. This facilitates its integration with the other tools used by the actors of a computer environment for human learning like, for example, the e-Learning platforms and annotations tools.**

Keywords *- CEHL; context-aware; annotation; adaptation; learning.*

I. INTRODUCTION

Teachers, tutors or co-authors, during their activities of teaching, as well as learner, during its training, in a Computer Environment for Human Learning (CEHL), manage a significant number of learning objects to support their activities. These objects are most of the time in numerical format. In order to memorize the learning objects elements,, these actors create, on the objects, different type of annotations, in order to re-use them like a working memory. Each actor thus, constitutes an external memory for his learning and teaching activities.

The external memory created, is composed of all learning objects and their annotations. It allows the teacher, for example, to memorize an idea to find it, thereafter, in a fast way, or to create an annotation in a context and to re-use it in another context. This external memory is useful and exploitable for all the actors of a computer environment for human learning; it must be well structured according to the semantics of the used annotations and to adapt them to the current context of the actor's activity. Thus, it will make it possible, for each actor of a CEHL to exploit in a fluid way its annotations.

In this work, we propose a new general architecture of the adaptive annotations memory with a detailed description of the functionalities offered by each module of the architecture of ontology-based framework for reusing and learning through context-aware annotations memory (OARLCAM). This architecture is based on a context meta-model and an annotation meta-model. The first one represents the learning context of the various actors of the annotations memory. The annotation meta-model describes the semantics of the annotation to be able to re-use them, share them and learning from knowledge included in these annotations according to a pedagogic objective.

This paper is organized as follows. The first section presents our research field. The second section exposes our research problematic by explaining the need for an adaptive memory of annotations for the different actors of a CEHL and by pointing out the basic concepts used. In the third section, we present the state of the art of the related approaches to our research by showing their advantages and their limits. In the fourth section, we propose and describe the architecture OARLCAM conceived to solve the whole of the released limits.

II. BASIC CONCEPTS AND PROBLEMATIC

The various actors of a CEHL find difficulties to create and consult the annotations associated to learning objects, during their learning activities. Generally, they use manual annotations which are very difficult to manage, to exploit, capitalize and share. These actors annotate in a different ways the learning objects according to their training activities and their learning goals. For this purpose, the existence of an annotations memory is necessary for these various actors. Indeed, it allows the automatic management of all these annotations making possible their capitalization and sharing. It is a learning based on annotations affixed by the various actors according to a particular learning context and a special learning objective. The numerical annotations constitute an added value for the learning objects. They add knowledge to the original contents of these objects.

By re-using the annotations stored in the memory annotations, a teacher for example, can re-use these annotations and those of the co-authors with the knowledge witch contains to improve his process of teaching and to be able to teach in a best way a given learning object.

He can also profit from all the critical, explanatory and prescriptive annotations posted by the tutors to improve the contents of the learning object and thus the quality of the training. He can even recover the learner's annotations and take them into account to add explanations to the learning object. A learning object can be annotated several times, in different contexts, in different places and moments and by heterogeneous tools. Then, it quickly becomes illegible at the time of its consultation. On another side, the actor of a

 ISBN: 978-1-61208-139-7

CEHL creates an annotation, in a given context, to re-use it in another context.

As a result, of these two remarks, we think that an annotation meta-model is essential to unify the various annotations produced by the actors. Moreover, it must be able to adapt itself to the context which surrounds the activity of annotation. Before presenting our architecture, we briefly expose the basic concepts, around our research.

A. Annotation

The state of the art relieves several definitions of the Annotation concept. The most adapted to our work was given by Mille [13], who considers an annotation as being "a trace of the activity of the reader, perceptible on a document as a mark, placed with a specific aim, and in a specific place of which it cannot be dissociated". In fact, the annotation object belongs to the document. An annotation has an objective that is well defined, because the annotator does not annotate for nothing, but for an essential goal.

According to this definition an annotation is regarded at the same time as an object and an activity [1]. An activity of creation of the annotation object aims to realise a user objective in the learning object.

B. Context and Context-awareness

According to the dictionary free on-line dictionary,Howe [8], the context is all that surrounds and gives a meaning to another thing. In this case, a definition of the context cannot be given in an isolated way without taking into account the element concerned with the context. This definition shows that the context must be external to the element concerned with this context.

A formal definition for the context is given by Azouaou and Desmoulins [1] as follows:

The context of an element X is the whole of properties P of any element of Y such as:
Y is around X.
Y gives meaning to X.
P is relevant for X.

In the literature, we find several definitions of context-awareness (conscience of the context). Schilit and Theimer [15] define the conscience of the context as being all the applications adapted to the context. Another definition given by Dey [5] specified that a system is considered as context-awareness, if it uses the context to give relevant information or a service to a user, knowing that the relevance depends on the activity of the user.

In addition, Context-awareness emerged in the fields of mobile and pervasive computing like a technique to design applications with a conscience of the environment, to ensure high level autonomy and flexibility. The context-awareness or the conscience of the context is known under other synonyms like *adaptive* or *reactivate* [4].

C. Personalization and of Adaptation of the CEHL

In order to be able to start the reflexion on the adaptation and personalization, we start by defining these two concepts.

The personalization of information consists in providing to one learner a relevant information equivalent to its preferences and its needs [3].

The adaptation is a transformation of an organization (or a genetic material…) in order to give it more adequacy with a new environment (a natural environment, a new political situation, a technology…) stimulating this adaptation. More particularly, in the context of adaptation of one system to the user, this generic term includes two specificities: adaptability and adaptivity.

In this way, Moisuc [14] explains, "*the adaptability is the capacity of the system to adapt to the personalizations explicitly requested by learner for example, while adaptivity indicates its capacity to meet the needs for learner without his explicit intervention"*.

From this definition, we understand that the adaptation, whatever static (adaptability) or dynamic (adaptivity), requires a recognition of the learning context in order to be able to adapt the training according to the context and the learning objective.

The following section presents researches related to our problematic.

III. APPROACHES RELATED TO OUR RESEARCH TASKS

During last years, several researches were carried out to delimit the needs of CEHL actors and to recense the main elements helping to develop e-learning systems better adapted to their trainings and their needs.

Although they are diversified, those researches do not take into account the capitalization of the learning experiments which can be exploited later by other actors. In fact, we think that the major stake of e-learning is to integrate an approach to re-use the learning annotations within a CEHL

A. Adaptation Approaches and personalized e-Learning systems

In the last decades, the scientific researches were oriented to the adaptation and the personalization of the HMI, in particular in the CEHL field. Indeed, several researches support the personalization in this field to guarantee a better satisfaction of learning.

One approach consists to allow teachers to make scenarios for all the learner's uses possible of the system. A teaching scenario describes goals and learning situations while defining how the learning objects will be implemented in a precise context of training [10]. However, the teaching scenarios help the teachers to integrate the CEHL into their work practices, but do not allow creating sequences of activities adapted to each learner's competences.

A second approach of Duclosson, Daubias, and Riot [6] and Leroux [12] devotes a part of the e-learning system to be personalized by the teacher. Thus, teachers can parameterize the generation of the activities or select the activities which are appropriate for their learner. This personalization is done manually by the teacher, without bond with possible learner profiles, and can't be considered as based on a generic or unified model. In fact, each teacher can use several e-

 ISBN: 978-1-61208-139-7

learning systems and must control several environments of personalization to succeed in defining in each case his teaching choices. Moreover, teacher himself needs to profit from its experiments and those of the other teachers, in addition to the tutor ones. We notice that this second approach is interested just in the training of learners and is not based on models.

A third approach consists in personalizing the e-learning systems automatically so that their contents are adapted to knowledge of each learner. This personalization can progressively be made throw the learners' answers (as a result to their behaviours) and uses the stereotypes associated to the learners [7] or using the learner model according to the e-learning systems. This automatic personalization is adapted to the system's knowledge about learner but is not always adapted to the teacher's learning goals.

Each one of these approaches answers part of the problem, but does not provide a solution to the whole: the adaptation of CEHL to the activities of the teachers and the adaptation of their contents to each one of their learners [11].

We present, in the following, some examples of projects working on the adaptation and the personalization in the e-learning systems.

B. *Approaches of annotations memories in e-learning*

The recent work on the annotations memories, Ouadah, Azouaou and Desmoulins [9] propose an annotation tool context-aware for an external annotations memory for the teacher. "*We are based on two architectures to propose the general architecture of our adaptive annotations tool. This tool must be able to identify the current context of teacher's activity in order to adapt to his behaviour and the changes of its situations of activity*".

This approach supports the teacher at his annotation activity to re-use it in another context. However, it is not an annotations memory for a general learning. Indeed, it can be integrated into an annotation tool and not into a computer environment for human learning. In addition, it is dedicated to only one actor who is the teacher.

In this way, we deduce the lack, in the literature, of an adaptable approach for the training of all CEHL actors based on annotations according to a given context.

IV. GENERIC ARCHITECTURE BASED ON WEB SERVICES

In our work, we propose an approach to mitigate the limits illustrated above by introducing a personalized learning architecture based on annotations (experience feedback). This ontology-based architecture OARLCAM makes possible to capitalize and re-use a collective annotations memory for training. Its main goal is to provide all actors with a best training relative to their learning objectives and the current context extracted from the annotations memory, the learning objects warehouse or both. Such a system can be used as an assistance for the original authors, co-authors, tutors and learners.

Our contributions can be presented as follows: i) modelling the semantic of various annotations used in CEHL by a top level ontology, ii) modelling the various contexts of training by a context top level ontology and iii) proposing the approach OARLCAM to automatically exploit this knowledge to generate the learning objects with an added value of annotations adapted and personalized.

The following figure presents our modular architecture.

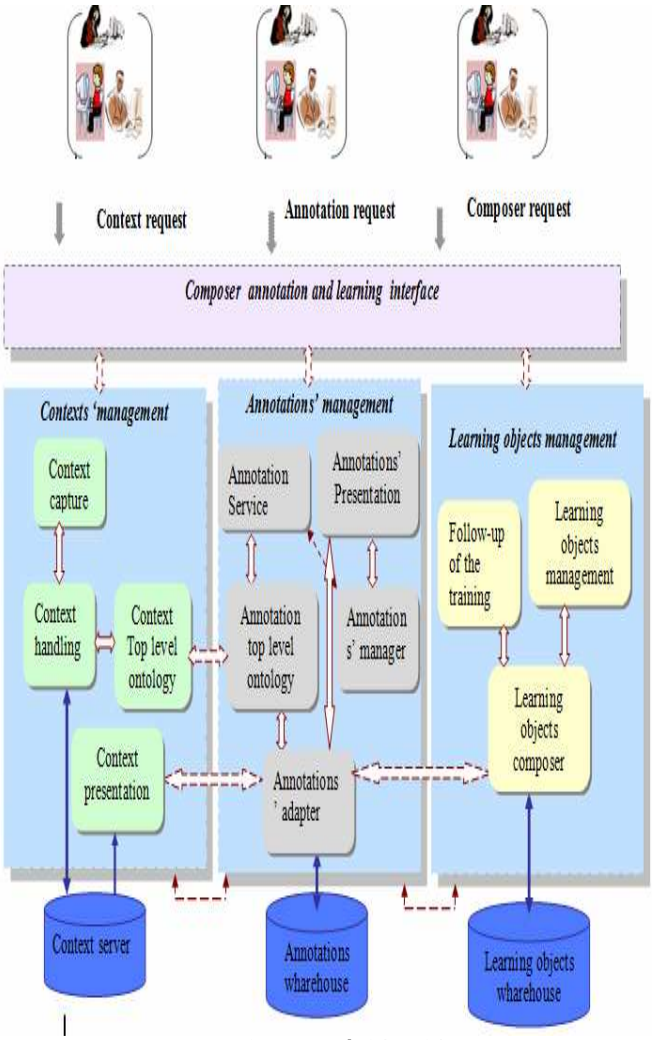


Figure 1. Our approach OARLCAM

This modular architecture is based on a context-aware approach coupled to some ontological engineering techniques in order to build a learning annotations memory, unified by an annotation top level ontology and a context top level ontology for an appropriate learning for all actors. It allows associating the adequate annotations with the appropriate learning object according to learning objective in a given context in order to improve the educational content and to activate the training process.

This architecture OARLCAM assures the reuse, the adaptability and the interoperability between our framework and the various tools used by the various actors, whom can use it as an external memory.

Our architecture includes three subsystems : i) the subsystem of contextualisation, containing the modules of context capture, context handling, context server, context

 ISBN: 978-1-61208-139-7

presentation and a context top level ontology of training ii) and the learning subsystem, containing the modules of learning objects management, learning objects composer and follow-up of the training; and iii) the annotation subsystem, containing the annotation module, the annotations' manager, the annotations' adapter, the annotation top level ontology, the annotations' presentation module and the annotations warehouse, for later re-uses. The following sections describe the modules of the architecture.

A. Subsystem of contextualisation

1) The context capture module: The capture of the context is carried out using a whole of services which interact with sources of context (operating system, learning objects manager, organizer, etc…). This interaction can be made in a direct way if context information is accessible, or in an indirect way thanks to an export operation of context from the context source and an import of these data on the level of this module.

2) The context handling module: Context informations, provided by the context capture module, are treated according to our context top level ontology in order to be stored in the context server. This treatment consists in making a mapping between data types of the context source and our context model.

3) The context server module: The context informations are stored in XML format in order to facilitate their sharing and their use by the adaptive application and to keep the contexts history.

4) *The context presentation module*: context information is presented using the Web services standards (the details of this presentation are not in this paper). Each Web service can be consumed by other applications to be adapted to their context. This module is a service which gives information about the current context to the annotations' adapter module, knowing that those informations are extracted from the context server.

5) *The context top level ontology module.* It is a generic and exhaustive context ontology which provides the proprieties of context related to learning provided by our annotations memory.

The context top level ontology is conceived to solve the limits and the insufficiencies of the existing context models. For a given learning objective and a given context, we must extract the adequate annotations from the annotations memory. We use then a mapping method for determining the similarities between a learning context (context top level ontology), annotation semantics (annotation top level ontology). The annotation top level ontology contains three facets: cognitive, semantic and contextual. On another side, the context top level ontology contains six facets. Four of these facets (user, activity, environnement, collaboration) where defined by [9]. We add two more facets composition and objective to have an exhaustif context model which take to account the context of reuse of learning and learning objective.

6) Algorithm of similarities between different contexts: This algorithm aims to determinate similarities between different contexts of reuse of learning annotations; we take into account both semantic and structured similiraties between the concepts of ontology concepts. For example, the concepts of context ontologies of learning : context 1 ontology and context 2 ontology of learning. The algorithm below describes the process to find similarities between contexts :

Algorithm: Similarities
INPUT :
1) *CO1* and *CO2*: Context Ontology 1 and Context Ontology 2
2) V_{SS}: Semantic vector of similarities
3) V_{SSt} **:** structural vector of similarities
4) Sim_{St} : Weight of structural similarities
5) Sim_S : Weight of semantic similarities
OUTPUT : V_{SG} : global (Semantic and structural) vector of similarities
Begin
/* go to each concept of context ontology 1 */
For each *(CCO1 ε CO1)* **do**
/* go to each concept of context ontology 2 */
For each *(CCO2 ε CO2)* **do**
If *CCO1*.type==*CCO2*..type **then**
/*Extract semantic similarities between *CCO1* and CCO2 of V_{SS}*/
Sim_S=EXTRACTSIM (V_{SS}, *CCO1, CCO2*)
/*Extract structural simiraties de *CCO1*et *CCO2* of V_{SSt} */
Sim_{St}=EXTRACTSIM (V_{SSt}, *CCO1, CCO2*)
/*calculate global similarity*/
$Sim_G = Sim_S + Sim_{St}$
/* Add CCO1, CCO2 and Sim_G in V_{SG}*/
Add ((*CCO1, CCO2*, Sim_G), *VSG*)
Return (V_{SG})
END

The proposed algorithm of similarities has as input the two ontologies of context 1 and context 2, the two vectors of semantic and structural the similarities (VSS, and VSST), as well as the weights related to the semantic and structural similarities (SimS and SimSt). It produces in result a vector of global similarity, VSG. The function EXTRACTSIM extracts the value of the similarity corresponding to the two concepts (CCO1 and CCO2) from the vector of similarity (VSS or VSST). For each couple of concepts, CCO1 and CCO2, having the same category of two context ontologies , CO1 and CO2, the global similarity is calculated as follows:

SimG (CCO1; CCO2) = SimS (CCO1; CCO2) + SimSt (CCO1; CCO22)

B. Subsystem of the learning objects management

1) The learning objects management module: This module is used to create, to add, to remove, and to modify learning objects.

2) The learning objects composer module: This module serves to compose the learning objects with existing annotations for a given learning objective and a given context.

 ISBN: 978-1-61208-139-7

3) The follow-up of the training module: This module serves to save the learning activities history of our architecture's actors

C. Subsystem of annotation

1) The annotation module: This module allows actors to add an annotation according to our annotation top level ontology.

2) The annotations' manager module: This module manages the annotations affixed according to our annotation tool, for example, to add, to modify or to remove an annotation.

3) *The adapter annotations' module:* This module adapts the annotations stored in the annotations warehouse according to the context of training (objective) provided by the service of context presentation. This service has also as a role to provide the result of the adaptation for the services of learning objects composer to combine them with these objects (annotations source) or to provide them directly to the annotations' presentation module (the result of the request). For example, one learner during his revision wants to extract from a learning object all the explanatory annotations related to this object (in an explanatory context). He also wants that this annotation will be posted only during its next envisaged revision in a given day and a given hour. In this case, the service of annotation checks the properties of the context provided by the context services and posts the annotation only if the context is verified.

4) *The annotation top level ontology module:* It is a generic and exhaustive annotations ontology which provides the semantics of different learning annotations. We develop this ontology to mitigate the various insufficiencies in the state of the art.

5) *The annotations' presentation module:* This module presents, for the architecture various actors, the annotations adapted to the learning context. These annotations are treated to be adequate to the actors' requests and their learning objectives.

D. Communication between the various modules with XML

The use of the Web services to publish the context data facilitates the interoperability of our framework with any application sensitive to the context. On another side, XML is currently the standard language used for the data exchange on the Web. For this reasons we adopt it for the exchange of the data between all the modules composing our architecture. In the same way, the communication between the various services is done using the SOAP protocol which is based on XML.

E. Illustrative scenario

This section gives an illustrative example of our architecture's use. One learner at the time of its revision for an examination of the first session, for example, wants to recover all the explanatory and analytical annotations relating to a given learning object. He then makes training from our system by specifying its learning objective and its learning activity (examination). This learning is located in a C1 context (learning objective=revision, pedagogic activity=examination, date=d1, hour=h1, place=p1, learning domain= data bases) and this learning can be re-used by the same learner or another actor in a C2 context. For example, a preparation context of a part of learning object by one teacher (learning objective= course conception, pedagogic activity=course preparation, date=d2, hour=h2, place=p2, learning domain=data bases).

Thus, the training provided to this learner through our framework must be adapted according to his context and his learning objective.

Then we note that, the same learner can re-use the same context C1 another time for its next planned revision. For example, if the properties values of the learning contextual facet are (date=15/11/2010, learning objective =revision, pedagogic activity =examination), learner specifies also the context of re-use of these annotations, for example, for the final examination, by choosing directly the next context C2 (date=15/01/2011, learning objective =revision, pedagogic activity =examination). Automatically, the 15/01/2010, during the nearest examination, the annotation is posted for learner, to remind him his second revision.

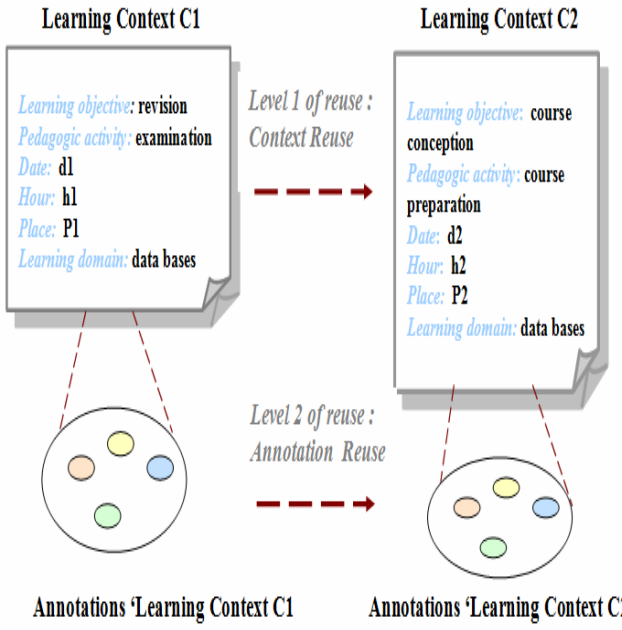


Figure 2. Reuse of context

In fact, our framework is able to provide a re-use on two levels. The first re-use (and sharing) is of annotations as well as the knowledge which is included there. The second re-use concerns the context. It means that a given context is reused in another context covering the same contextual properties and stored in the context server for a several re-uses later by the same actor or others.

 ISBN: 978-1-61208-139-7

Moreover, the annotations can be presented with or without their learning objects. In the first case they can be used to obtain an added value for the learning objects (learning objects composed with annotations). In the second case, they constitute an important knowledge to be capitalized.

V. CONCLUSION AND FUTURE WORKS

This paper represents a preliminary study to realize an adaptive annotation memory for context-aware training for the various actors in a CEHL. This memory can satisfy the need for training according to a given objective and a given context of all the actors in terms of utility, re-use, sharing and adaptability.

The running mechanism of our framework articulates around a whole of modules. Each module allows a functional need well defined and is composed of Web services. This framework aims to facilitate its integration, its interoperability with other e-learning systems. Moreover, it is based on three independent but communicating subsystems. The first is a subsystem for the training context capture, its treatment, its storage and its sharing by context sensitive applications. The second subsystem is devoted to the management and the adaptation a given annotation, from annotations memory, to the captured context. The third subsystem is planned for the management and the composition of the learning objects with the annotations extracted by the second subsystem.

Several perspectives are possible for this work. In particular, we aim to use the techniques of data warehousing (wrapper, monitor, etc.) to extract annotations from the annotations data bases and warehouses relative to other e-learning tools. Thesis extracted annotations will be total forwarded to our annotations memory according to our annotation top level ontology.

Also, we think that adding an ontology for the automatic deduction of the context annotation reuse, deduced for each actor, would be an important enrichment to our architecture and its functionalities.

REFERENCES

[1] F. Azouaou and C. Desmoulins: Taking Teaching Context into Account for Semantic Annotation Patterns. EC-TEL 2006, pp. 543-548, last access date 12.02.2011.

[2] D. Mille, Modèles et outils logiciels pour l'annotation sémantique de documents pédagogiques. Thèse., in Département informatique. 2005, Université Joseph-Fourier : Grenoble. last access date 09.02.2011.

[3] A. Anli, C. Kolski, and M. Abed, Principes et architecture pour la personnalisation d'information en interaction homme-machine, Application à l'information transport, IHM, Toulouse, pp. 123-130, 2005, last access date 10.01.2011.

[4] T. Chaari and F. Laforest LIRIS, L'adaptation dans les systèmes d'information sensibles au contexte d'utilisation: approche et modèles. Conference Genie Electrique en Informatique, Sousse, Tunisie ; 2005, pp. 56-61 last access date 02.02.2011.,

[5] A. Dey, Providing Architectural Support for Building Context-Aware Applications, in College of Computing. 2000, Georgia Institute of Technology, pp 23-35, last access date 05.01.2011.

[6] N. Duclosson, S. Jean-Daubias, and S. Riot, AMBRE-enseignant : un module partenaire de l'enseignant pour créer des problèmes, EIAH'2005, Montpellier, 25-27 mai 2005, pp. 353-358, last access date 04.02.2011.

[7] S. Girard and H. Johnson, DividingQuest : opening the learner model to teachers, EIAH'2007, Lausanne, 27-29 juin 2007, pp. 329-334, last access date 05.02.2011.

[8] D. Howe, Free online dictionary of computing. London, UK, Imperial College Departement of Computing, last access date 12.02.2011. .

[9] A. Oouadah, F. Azouaou, and C. Desmoulins, Modèles et Architecture d'une Mémoire d'Annotation context-aware pour l'Enseignant. 2ème conférence internationale sur l'informatique et ses applications, CIIA09, 03-04 Mai 2009, Saida, Algérie, last access date 10.02.2011.

[10] A. Lejeune and J.-P. Pernin, A taxonomy for scenario-based engineering, CELDA 2004, Portugal, pp. 249-256, last access date 05.02.2011.

[11] M. Lefevre, S. Jean-Daubias, and N. Guin. Personnaliser des séquences de travail à partir de profils d'apprenants, EIAH 2009, Le Mans, 23-26 juin 2009, pp. 1136-1141, last access date 04.02.2011.

[12] P. Leroux, Machines partenaires des apprenants et des enseignants – Étude dans le cadre d'environnements supports de projets pédagogiques. Habilitation à Diriger des Recherches, Université du Maine, 2002, last access date 06.02.2011.

[13] D. Mille, Modèles et outils logiciels pour l'annotation sémantique de documents pédagogiques. Thèse., in Département informatique. 2005, Université Joseph-Fourier : Grenoble, last access date 24.12.2010.

[14] B. Moisuc, Adaptation dans les systèmes d'information spatio-temporelle interactifs, INFORSID, Hammamet, Tunisie, 2006, last access date 25.12.2010.

[15] B.N. Schilit and M.N Theimer. Disseminating active map information to mobile hots, IEEE Network, 8(5), pp. 22-32, 1994, last access date 28.12.2010.

 ISBN: 978-1-61208-139-7

Transforming Source Code Examples into Programming Tutorials

Roger Rudolph Gajraj
Department of Computing & Information Technology
The University of the West Indies
St. Augustine, Trinidad
roger.gajraj@uog.edu.gy

Malcolm Williams
Department of Computer Science
The University of Guyana
Turkeyen, Guyana
malcolm.williams@uog.edu.gy

Margaret Bernard
Department of Computing & Information Technology
The University of the West Indies
St. Augustine, Trinidad
margaret.bernard@sta.uwi.edu

Lenandlar Singh
Department of Computer Science
The University of Guyana
Turkeyen, Guyana
lenandlar.singh@uog.edu.gy

***Abstract*—One popular approach to teaching computer programming is to use example programs to demonstrate programming concepts. We propose to increase the pedagogical value of example program source code by transforming them into self-explaining tutorials within a learning integrated development environment. In this paper, we present a stepwise instructed implementation of annotated example code. Source code with instructor comments is parsed and processed to create an intelligent learner environment. Students are guided step by step to develop the program solution. Explanations are auto-generated for each line of code; these come from an author's comments as well as extended explanations dynamically generated for certain coding constructs. Explanations are presented to the learner in multiple modes using the full range of multimedia displays. Source code examples can be used as self-contained tutorials.**

Keywords-e-learning; educational technologies; programming pedagogy; source code examples; integrated development environment

I. INTRODUCTION

Teaching by example is a well established method for teaching computer programming. Experienced teachers of programming prepare suitable examples of program code which illustrate the particular programming construct that is being taught; students learn from these examples by seeing how the construct is used in problem solving as well as familiarizing themselves with syntax and semantics of the language. These example programs may be made available to the student in digital form and students may be encouraged to run the code in some IDE. Many programming textbooks use this approach of teaching by example.

In this paper, we present a computer aided example-based approach to teaching and learning programming. Students are guided step-by-step to construct the examples that the instructor has already prepared. We also present a learner's integrated development environment (L-IDE), called CSmart, which facilitates this guided instruction. The instructor creates annotated example programs which can illustrate the use of a programming concept. When the student selects that example problem, he is not shown the solution code immediately; rather he is guided through step-by step instructions so that he can develop the program solution which the instructor/expert has prepared. This is done dynamically, in real time, where the example source code with instructor's annotation is parsed and used to generate explanations of the programming construct in focus. The explanations generated are a combination of text and visual representations. The intelligence is built into CSmart so that instruction is not static but relates to the point in the program that the student is working on. CSmart stores values of variables in the program and is able to present to the learner exactly what is happening with the code that he is typing in.

This example-based pedagogical approach for teaching and learning programming actually combines the traditional approach of learning by example with learning by doing and learning using visualization. It can be used for novice programmers where the instructor examples are simple one-task programs, with possibly just one construct (a FOR loop, for example). The approach can also be used to develop more experienced programmers where the instructor examples focus not so much on syntax but on expert techniques for problem solving. Most instructors have a bank of example programs that they reuse with different student groups. The CSmart environment allows the instructor to create a repository of annotated examples, categorized and sequenced in any manner that the instructor prefers.

In the following sections, we present our pedagogical approach after highlighting related work and discussing their approaches. The CSmart environment is then detailed to show how it embodies our novel pedagogical concept by processing source code examples as tutorials themselves.

 ISBN: 978-1-61208-139-7

II. BACKGROUND

A. Related Literature

In a study of difficulties faced by novice programmers, both students and teachers valued example programs as the most useful type of learning material [1]. Most sources of static examples are accompanied by explanations in varying degrees of detail and usefulness. Programmers conventionally seek example code in resource repositories such as tutorial web-pages, forums and books. These examples can then be implemented in a separate integrated development environment (IDE) and then compiled to run the resulting executable program. This is done to gain deeper understanding of what example code really does at runtime execution.

Using source code to aid in the learning of programming has been explored by environments such as 'WebEx' [2], 'Jeliot' [3] and 'BlueJ' [4]. 'WebEx' presents example source code with annotations explaining relevant lines of code. A read-only exploration of an example is done by altering the visibility of annotations. 'Jeliot' focuses on auto-generating visualizations of source code's execution at run-time. 'BlueJ' also parses source code like 'Jeliot' but to produce visualizations promoting object oriented concepts only. In addition, 'BlueJ' is a simple IDE that provides graphical and textual editing of source code which can be further compiled.

Of the aforementioned environments, only 'WebEx' focuses on directly leveraging pedagogical value from example source code by presenting textual explanations of individual lines of code. However, 'WebEx' hardly differs from reading static commented source code because it only allows interactivity with reading annotations which is similar to reading extended comments. Also, there is no IDE support for a user to gain insight from the experience of an example program's behavior at run-time.

None of the environments mentioned provide guidance for the actual activity of programming implementation. In addition to an instructor explaining example code in traditional lab environments, students are often guided to implement the example and encouraged to change and apply the example in order to solve a similar problem. Guided implementation of example code could assist in reducing cognitive overheads while learning to program.

B. Pedagogical Approach

We propose to increase the pedagogical value of source code by transforming them into self-explaining tutorials within an integrated development environment. This pedagogical approach to teaching and learning programming actually combines the traditional approach of learning by example with learning by doing and learning using visualization. This 3-pronged approach combines strategies which individually are well established in the literature as being effective for teaching and learning programming [1, 5].

Each line in a source code example can be presented as an instruction for a learner to actually type into the environment. We posit that this instructed implementation activity promotes learning by doing. In addition, the guidance to implement working code may reduce cognitive overheads that may hinder the learning process.

Teachers and books often try to explain source code examples line-by-line. Comments belonging to selected lines of source code can store an author's explanation of its intended purpose. As a result, we propose to use comments to annotate lines of source code. These annotations can be positioned strategically within an IDE at the same time the instructed implementation activity is carried out by the learner.

Visualization is also used as an effective pedagogical technique [5]. Visualization of what certain code constructs are doing can be provided as graphical annotations alongside the textual annotations. The approach taken is different from algorithm visualization, which focuses on visualizing an entire algorithm. We rather focus on visually explaining a line of code by itself. Visuals are presented as metaphorical representations of code semantics.

We call this entire pedagogical approach "stepwise instructed implementation of annotated example code".

III. LEARNER INTEGRATED DEVELOPMENT ENVIRONMENT

Traditional IDEs provide little support for novice programmers who attempt to learn programming from example code. Beginners may have to type example code into an IDE either from book sources or "copy and paste" from electronic media sources. We propose a "learner's integrated development environment" (L-IDE) concept which supports the use of example code to aid in the learning process.

We have created a L-IDE model which parses actual source code files into a presentable format more amenable to learning. A beginner is guided through implementation of an example program. Code comments are extracted and provided as annotations for respective lines of code. Visualizations are generated dynamically from code which attempt to give a graphic representation to aid understanding of abstract programming concepts. After guided implementation of example code, a beginner programmer has the option to edit the example and/or compile and execute the code with possible error feedback.

The following sub-sections describe the features of our L-IDE called 'CSmart'. Currently, CSmart is limited to parsing source code of the C programming language.

A. Information Extraction

Tutorials are created dynamically where example source code files are parsed and presented in a richer format that aids learning. Figure 1 shows how an actual line of code and relating comment is parsed from a source code file and displayed to the learner as an informal instruction within CSmart's editor.

B. Stepwise Guided Implementation

In applying the step-wise guided technique, the learner is instructed to type out each line of code in an example. This stepwise approach fosters learning through learning by doing. In the 'CSmart' L-IDE, the example code that is to be typed out, as part of encoding the program, is placed under

the cursor – within a tooltip – in full view of the programmer, while they type the code on the line above the tooltip guide (see Figure 1 and Figure 2).

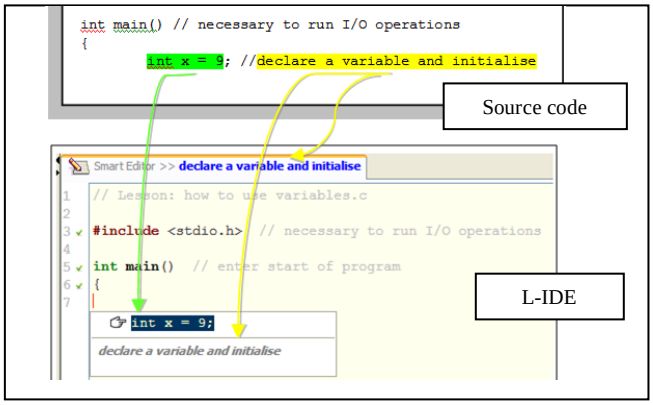


Figure 1. Code and comment extraction from source code

Similarly, the comments that explain a line of code are placed just beneath the line of code. This allows very little overhead of copying code because it is very close to the input area, specifically, where the cursor is located within the editor area. We posit that the strategic presentation all of these pieces of information (i.e. presenting a lot of information in a small space without overloading the user) increases the information content of the user interface, and facilitates learning.

C. *Explanation of Code*

Example source code demonstrates select programming concepts which can be explored by the learner in order to foster learning and understanding. A line by line explanatory approach was used to explain example code. This allows learners to focus on the semantics of the programming language. To best explain examples visually and textually, Clark & Mayer provide the following tested guidelines in [6] which were considered and incorporated into the design:

- "multimedia principle" which involves appropriate use of text, audio and graphics
- "contiguity principle" ensuring words align to graphics
- "segmenting principle" where small manageable chunks of data are presented at a time to the user

These principles were applied to explain lines of code visually, textually and audibly. These techniques are further illustrated in the proceeding sub-sections.

1) *Annotation from Comments*

Comments from source code files are used to provide auto-generated annotation of the corresponding lines of code that the learner types out in the L-IDE text editor. Where no comments exist in the example source code, no auto-generated explanation for those lines of code is generated. Authors of tutorials (example code) are encouraged to create better tutorials by including comments in each line of code. It is the human tutor's responsibility to ensure appropriateness of explanation of lines of code.

In addition, there is an automatic display of extended explanations of certain coding constructs. Certain coding constructs may be keywords or popularly used functions of the C programming language. These extended explanations intelligently appear next to the editor area whenever certain coding constructs are used in a line of code in order to augment the learning process.

For example, a line of code and comment within a source code file may be:

int x = 9; //declare a variable and initialise

From the previous line of code above, Figure 2 shows how the comment – as an explanation – is displayed just under the instructed line of code to be typed out.

2) *Information pane for Keywords and Functions*

The CSmart environment displays additional information when keywords and functions from the C standard library are found while parsing lines of code. This intends to give a learner a better understanding of how to use keywords and functions.

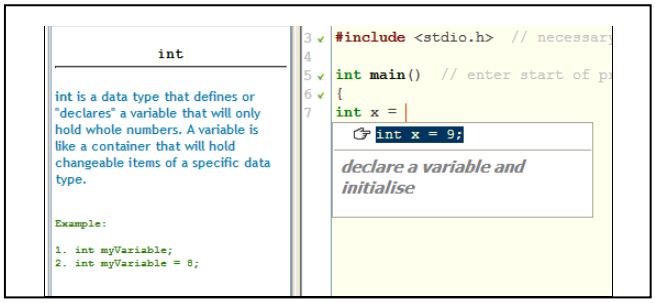


Figure 2. Textual explanation of instructed line of code to be typed out

For example, the left pane in Figure 2 beside the editor dynamically displays additional information on the 'int' - integer - data type that is found in the line of code.

3) *Visual explanation of source code*

Visual explanations of code constructs are automatically generated just under comments that explain an instructed line of code that is being typed out. Visuals attempt to explain coding constructs with very little prerequisite of programming concepts. Beginners may understand programming concepts easier by being able to relate to visuals.

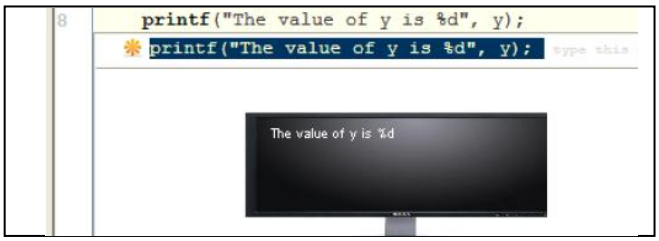


Figure 3. Visualization for output function 'printf'

For example, the popular 'printf' statement for output to screen is portrayed (see Figure 3) by showing the output on

an image of a computer monitor. The intention is to get the learner to understand what printf does – it outputs text to screen. In another example (see Figure 4), variables are portrayed as containers because a container can hold an item like a variable holding data.

Calculation and assignment based visuals are dynamically generated from a trace-table data structure that keeps track of the data in variables. The trace table is used to show what data is moved and/or replaced within variables during calculations and assignments. An extension of the java programming language called 'Groovy' [7] was used to evaluate loop and conditional expressions and to provide the inputs into the visualization module. Figure 4 shows a calculation where the data in variable 'x' is added to the number seven (7) to give the sum of sixteen (16) that is further animated by moving the number sixteen (16) into the 'y' container representing the 'y' variable. This animated path is represented by the broken line in Figure 4.

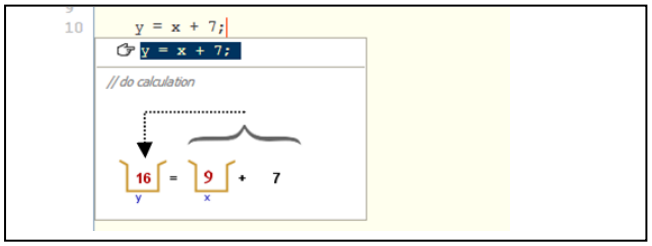


Figure 4. Visualization of calculation and assignment of data to variable

D. Compilation and Execution

The 'CSmart' L-IDE provides the opportunity to compile and execute implemented source code during a tutorial. Learners can see how a program behaves from its implemented example code.

Additionally, users may experiment with example code in terms of changing literals and variables in the original example. Compilation of modified code with error feedback is possible with additional support from an incorporated C compiler.

IV. Discussion

The learner integrated development environment's ability to create tutorials directly from source code infers a high ease of adoption by instructors as a teaching tool. There is neither a new method nor syntax to learn in order to create a tutorial. Tutorials are source code examples themselves and the quality would vary based on how instructors annotate lines of code with comments. Repositories could store high quality source code examples for re-use.

Instructors can focus on helping students individually instead of losing time to write and explain examples verbally. Students are able to go through tutorials at their own pace via a computer based pedagogical resource which could promote better learning.

As classroom sizes increase, students gain less attention from instructors. However, all students will be exposed to the same quality of pedagogy via a L-IDE such as 'CSmart'.

V. Conclusion and Future Work

This paper presents a work in progress, where a development environment is being used to deliver programming tutorials from source code examples. However, the pedagogical presentation of visual and textual explanations together with interactive implementation is constrained by the content within a source code example. Therefore, the quality and context of a tutorial is controlled by an example source code's content; code is instructed to be typed in an editor while it's relating comments are presented as annotations.

Experimental evaluation of the combined pedagogical approach, within the L-IDE, is the next step for future research. Initial evaluation of this pedagogical approach, within the 'CSmart' environment, was encouraging as reported in [8]. Participants in the evaluation rated their understanding of tutorials at a mode of eighty percent (80%). All participants reported that the L-IDE was helpful and they felt that they learnt some programming. These results were further supported by the following unsolicited comments from the participants:

- "made concept simple"
- "it helped to make things a little more understandable"
- "at least I learnt something by doing the tutorials"
- "boosted my knowledge"

Visualization could be enhanced by fading from a metaphorical representation of source code execution to a machine level representation in order for learners to gain a deeper understanding of how a program works at run-time.

If future evaluations show a positive effect of the pedagogical approach in a real learning process, students can use a L-IDE environment to learn programming on their own, just as if an instructor is there to guide them in person.

References

[1] E. Lahtinen, K. Ala-Mutka, and H. Järvinen, "A study of the difficulties of novice programmers," Proc. 10th annual SIGCSE conference on Innovation and technology in computer science education (ITiCSE '05), ACM, 2005, pp. 14-18, doi:10.1145/1067445.1067453

[2] P. Brusilovsky, I. Hsiao, and M. Yudelson, "Annotated program examples as first class objects in an educational digital library," Proc. 8th ACM/IEEE-CS joint conference on Digital libraries (JCDL '08), ACM, 2008, pp. 337-340, doi:10.1145/1378889.1378946

[3] A. Moreno, N. Myller, M. Ben-Ari, and E. Sutinen, "Program animation in jeliot 3," Proc. 9th annual SIGCSE conference on Innovation and technology in computer science education (ITiCSE '04). ACM, 2004, pp. 265-265. doi:10.1145/1007996.1008099

[4] M. Kolling, B. Quig, A. Patterson, and J. Rosenberg, "The BlueJ System and its Pedagogy," Journal of Computer Science Education, Special Issue on Learning and Teaching Object Technology, Vol 13, No 4, Dec 2003.

[5] J. Bergin, K. Brodie, and M. Patiño-Martínez, "An overview of visualization: its use and design: report of the working group in visualization," Proc. 1st conference on Integrating technology into computer science education (ITiCSE '96), ACM, 1996, pp. 192-200, doi:10.1145/237466.237647

 ISBN: 978-1-61208-139-7

[6] R. Clark and R. Mayer, "E-learning and the science of instruction," 2nd ed., San Francisco: Pfeiffer, 2008.

[7] D. Koenig, A. Glover, P. King, G. Laforge, and J. Skeet, "Groovy in Action," Manning Publications Co., Greenwich, CT, USA, 2007.

[8] R. Gajraj, "A Computer Based Programming Pedagogy: stepwise instructed implementation of explained example code," Thesis, unpublished.

 ISBN: 978-1-61208-139-7

Logitboost-SO Learning Algorithm for Human Iris Recognition

Wen-Shiung Chen
Dept. of Electrical Engineering,
National Chi Nan University,
Nantou, Taiwan
e-mail: wschen@ncnu.edu.tww

Lili Hsieh
Information Management Dept.
Hsiuping Institute of Technology
Taichung, Taiwan
e-mail: lily@mail.hit.edu.tw

Wei-Chih Tang
Dept. of Electrical Engineering,
National Chi Nan University,
Nantou, Taiwan
e-mail: 97323559@ncnu.edu.tw

***Abstract*—Boosting has been used extensively in the field of machine learning. This work intends to apply boosting method to iris biometrics. The recognition performance of boosting-based classification can be improved greatly by means of different re-weighting rules and different voting regulations based on the assigned weights. This paper proposes a novel Logitboost-SO algorithm which integrates similarity-oriented concepts into additive logistic model. We modify the existing manner of combining classifiers with Logitboost by utilizing multi-weight update rule to refine boosting algorithm. The experimental results show that Logitboost-SO applied to iris recognition is better than existing boosting algorithms.**

Keywords–Biometrics; Iris Recognition; Boosting; Adaboost; Logitboost..

I. INTRODUCTION

Iris is a thin circular organ which lies between the cornea and the sclera in human eyes. This trait has always been used for high security applications since jailers identified criminals by their irises in 18th century Paris prison. Among all the biometric traits, human iris has rich texture information and excellent uniqueness, which has been proved by Daugman in [1]. He presented the results of 200 billion iris pair comparisons to give the conclusion that the iris is fit for national scale deployment of recognition. After that, there are also many other researchers who propose different novel iris recognition systems [2]-[6]. The framework of recognition includes three units: pre-processing, feature extraction and learning/classification. This work aims at the learning and/or classification step. In any biometric recognition system, some known data are fed into the machine learning sub-system for training a classifier, and then the optimal classifier learn well and created after training. However, Logitboost has not been paid much attention on iris. As a superior boosting, Logitboost has only been regarded as a learning tool and its potentials are ignored. Even so, there are many fields including tumor [7], protein [8], and text [9] classification in which the researchers proposed the variants of Logitboost to discriminate the classes. In these papers, the base classifier of Logitboost is a breaking through point and provides Logitboost lots of improvement space. Here we choose a boosting algorithm as machine learning.

Boosting is a learning method which finds a way to combine "weak" classifiers and build up a "strong" classifier. Since Freund and Schapire [10] invented Adaboost in 1996, there are many works about how to improve Adaboost and some variants of Adaboost are proposed. Logiboost [11] applied backfitting to a logistic regression model and gives limited weights to mislabeled samples. SOBoost [12] operated similarity oriented rules. In this paper, we propose a novel boosting algorithm, called Logitboost-SO, which exploits multiple re-weighting rules to integrate the similarity oriented concepts into logistic regression model.

This paper is organized as follows. In Section II, we give the motivation to this work and then propose a new boosting learning algorithm. Section III states the experiment on iris recognition and discusses the results. Section IV makes a conclusion.

II. THE PROPOSED BOOSTING ALGORITHM

A. Motivation

The motivation of adopting SOBoost as the weak learner can be summed up with three points: (i) SOBoost is a novel boosting algorithm which follows the similarity rules and outperforms Adaboost; (ii) the confidence value of decision tree has only two values, 1 and -1. However, the confidence-rated classifier of SOBoost can produce a confidence value between -1 and 1; (iii) Logitboost demands strictly on the component classifier. However, the re-weighting and potential function of SOBoost is similar to Logitboost's (see Fig. 1), hence we infer that there exists a way to blend the two of them.

 ISBN: 978-1-61208-139-7

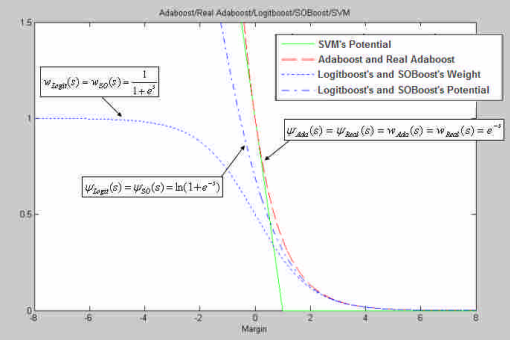


Figure 1. Comparison of the potential and weight function of the machine learning algorithm.

B. The Logitboost-SO Learning Algorithm

Though many people use Adaboost, there is a problem with it. And that is when you have the samples of enormous classification errors, the weights will increase excessively. In the result, this phenomenon will destroy the whole training system. Friedman *et al.* [11] (2000) applied log-likelihood loss function to replace the exponential loss function used in Adaboost and limited the weight distributed over the "miss" samples. Moreover, Logitboost fits the weak learner by a weighted least-square regression, which means not only the miss samples, but also the "too-correct" classifications are punished. The confidence scores of the real Adaboost are generated from the natural logarithm of the ratio of $P_w^+(\phi_t)$ to $P_w^-(\phi_t)$, where ϕ_t is the best feature of t iteration, $P_w^+(\phi_t)$ and $P_w^-(\phi_t)$ are the probability distribution of positive and negative samples. Hence the samples which are more similar may have lower confidence score, and this is a little contradictive. Thus in 2008, He *et al.* [12] proposed a similarity-oriented boosting (SOBoost) algorithm. In SOBoost they use the ratio of the bidirectional cumulative probability distribution to construct the confidence function. And therefore they ensure the monotonous of the hypotheses.

The main idea of Logitboost-SO machine learning algorithm is to integrate SOBoost concept into the Logitboost scheme. In the Logitboost-SO algorithm, the confidence function of SOBoost is used to fit an additive logisitic regression model. Besides, the weights used for calculating the confidence function are separated from the weights used for fitting logistic model and update independently according to the re-weighting rule of the SOBoost. The flowchart of Logitboost-SO is shown in Fig. 5. The pseudo-code is listed and described briefly as follows:

- **Training set :** $(x_1, y_1),...,(x_m, y_m)$ where $x_i \in X$, $y_i = \{-1,+1\}$, feature $\phi_k \in \Phi$.
- **Initialize** the weights

$$w_{same}(i) = D_{same}(i) = \frac{1}{2m_{same}},$$

$$w_{different}(i) = D_{different}(i) = \frac{1}{2m_{different}};$$

committee function $F(x_i) = 0$;

probabilites $p_{same}(i) = p_{different}(i) = \frac{1}{2}$.

- **For** $t = 1,2,...,T$

1. For $k = 1,2,...,q$

a. Divide X into J parts $X_1,...,X_J$.

b. Compute

the probability

$$P_w^l(j) = P(x_i \in X_j, y_i = l) = \sum_{i:x_i \in X_j, y_i = l} D_t(i), l = \pm 1;$$

cumulative probability

$$C_w^+(\phi_t) = \int_{-\infty}^{\phi_t} P_w^+(\phi_t) d\phi_t, \quad C_w^-(\phi_t) = \int_{\phi_t}^{\infty} P_w^-(\phi_t) d\phi_t$$

the weights

$$w_{same}(i) = p_{same}(i)(1 - p_{same}(i)),$$
$$w_{different}(i) = p_{different}(i)(1 - p_{different}(i));$$

working responses

$$z_{same}(i) = \frac{1 - p_{same}(i)}{p_{same}(i)(1 - p_{same}(i))},$$

$$z_{different}(i) = \frac{-p_{different}(i)}{p_{different}(i)(1 - p_{different}(i))};$$

the confidence function

$$h(\phi_k) = 2\,\mathrm{sigmf}(C_w^+(\phi_k) - C_w^-(\phi_k), \alpha, c) - 1$$
$$= \frac{2}{1 + \exp(-\alpha(C_w^+(\phi_k) - C_w^-(\phi_k)) - c)} - 1.$$

2. Fit the function $h(x_i)$ by a weighted least-squares regression of $z(x_i)$ to x_i with weights $w(x_i)$ and get

$$h_t(\phi_t) = \arg\min_{h:\phi \in \Phi} \sum_{i=1}^{m} w(x_i)\left(z(x_i) - h(x_i)\right)^2.$$

3. Update

$$D_{t+1}(i) = \frac{D_t(i)\,\mathrm{sigmf}(-y_i h_t(\phi_t(x_i)), \beta, 0)}{Z_t}$$

$$= \frac{D_t(i)}{Z_t(1+\exp(-\beta(-y_i h_t(\phi_t(x_i)))))}$$

where Z_t is a normalization factor

$$Z_t = \sum_{i=1}^{N} D_t(i)\,\mathrm{sigmf}(-y_i h_t(\phi_t(x_i)), \beta, 0);$$

and

$$F_{t+1}(x_i) = F_t(x_i) + \frac{h_t(\phi_t(x_i))}{2},$$

$$p_{t+1}(x_i) = \frac{1}{1+e^{-2F_{t+1}(x_i)}}.$$

- **Output** the final classifier $H(x) = \mathrm{sign}\sum_{t=1}^{T} h_t(\phi_t(x))$

The framework can be constructed by the following steps. First of all, we divide the dataset into *J* parts, and *J* depends on how sensitively you want about this machine. Second, we calculate the probability of each decided level. Based on decided levels we can get the weights which were generated from Logitboost idea. These probabilities are used to link the cumulative probabilities and confident scores which are applied in the SOBoost idea. Finally, these parameters work as the input of weak classifiers and the reweighting functions of Logitboost, and then we can obtain this machine. Concisely speaking, the flowchart of Logitboost-SO algorithm (Fig. 5) may expound this integrated concept.

III. EXPERIMENTAL RESULTS AND DISCUSSION

Verification experiments are carried out on UBIRIS.v1, to obtain a threshold separating false rejection rate (FRR) and false acceptance rate (FAR). For the case of FRR, we obtain the distribution of matching distance between the unknown classes and the registered classes. For the case of FAR, we also obtain the distribution of matching between the unknown classes for impostors and the registered classes. We use the equal error rate (EER) to evaluate feature extraction performance.

Embedded Based on the proposed learning algorithm, an iris recognition system, as shown in Fig. 2, is designed. The experiments of verification and identification are performed on UBIRIS.v1 Session 1, which has at least 5 pictures in each class. Moreover, a four-fold cross-validation is carried out to confirm the performance. In each iteration of the cross-validation, three of four partitions are used as the training set, and the other one partition is used as testing set.

As mentioned previously, we divide the iris region and feed in the boosting algorithm as different feature candidates. In Fig. 3, we apply colors in order of importance (sum of the weights) to illustrate the weights distributed by Discrete Adaboost on 8×4 non-overlapping blocks. The brighter color means that the selected region is more important.

Furthermore, in the sake of comparing the number of divisions, we also partition the iris region into 32×13 overlapping blocks with size 64×32 and feed them as different "features" in Discrete Adaboost. Since the experimental result of 32×13 divisions is much better than 8×4 divisions when using Discrete Adaboost, all the other boosting algorithms are only conducted under 32×13 divisions. Besides, since the weights of confidence-rated predictor cannot be summed, we only present the first four regions selected by Logitboost-SO in Fig. 4. From Figs. 3 and 4 we can see obviously that the lower half part of the iris region is more discriminative.

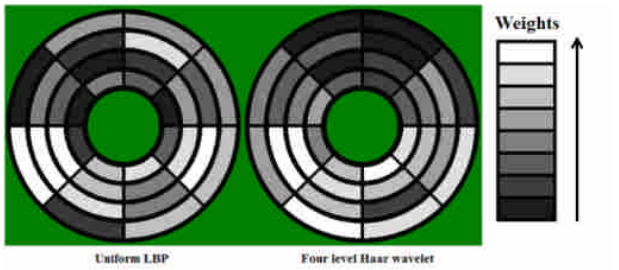
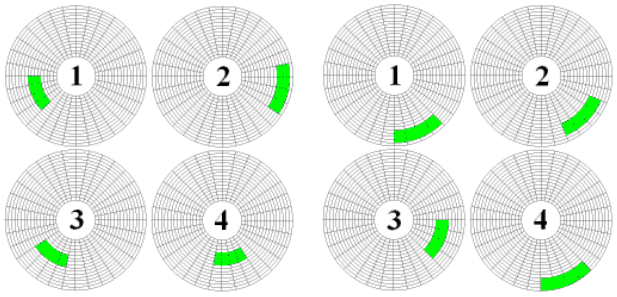


Figure 3. Weight distribution obtained from Discrete Adaboost after 500 iterations. The brighter color means that the selected region is given more weight.



(a) 4-level Haar wavelet feature (b) ULBP feature

Figure 4. The first four sub-regions selected by Logitboost-SO with different feature types.

With the levels of Haar wavelet getting higher more global information is extracted and more local information is lost. According to the characteristics of the input images, the suitable feature type for recognition is different. In Table 1, we compare FAR and FRR of different levels of Haar wavelets with discrete Adaboost, and discover that the use of four-level Haar wavelet is the best choice for our experiments.

From Tables 2 and 3 it is clear that the results of 32×13 candidates are much better than those of 8×4 candidates when using discrete Adaboost. Hence in the experiments of other boosting algorithms, the iris images are all divided into 32×13 candidates.

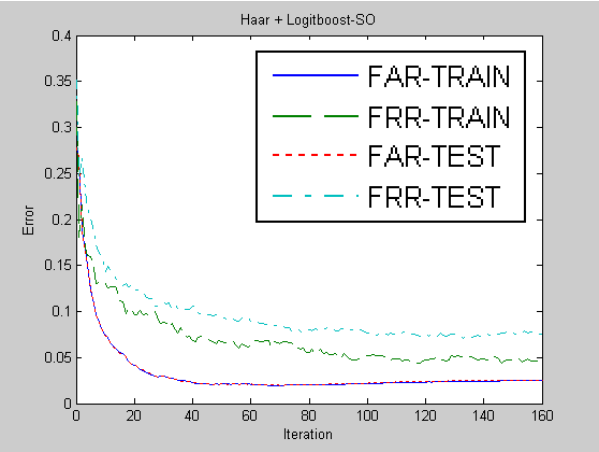
Moreover, real Adaboost performs incredibly well on training set but worse on testing set. This reveals that real Adaboost tends to be overfitting. However, as mentioned in [5], SOBoost has no such a problem. Besides, since Logitboost-SO learns the SOBoost rules, we can also avoid this problem. The results in Tables 2 and 3 confirm this argument.

Experimental results also show that the proposed boosting is better than existing boosting algorithms, whether for verification or identification applications. The

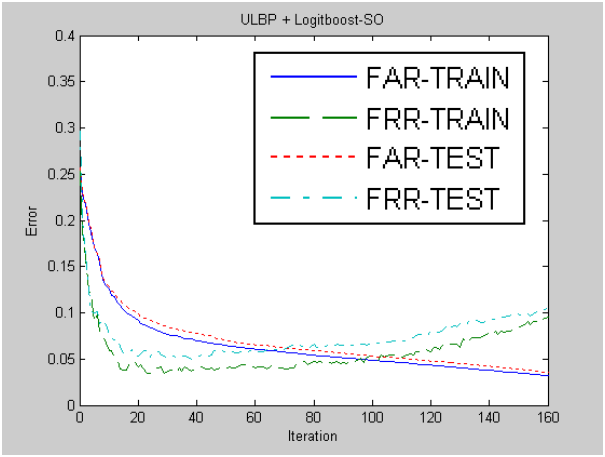
 ISBN: 978-1-61208-139-7

FAR and the FRR curves of Logitboost-SO on training and testing set are shown in Fig. 6.

In our identification experiments, the *k*-nearest neighbor algorithm is cascaded after the boosting methods. From Table 3 and Fig. 7 we observe that 3NN performs worse than 1NN. The reason is that the boosting has trained the classification system completely, and hence the larger value of *k* only makes the boundary less distinct.



(a) 4-level Haar wavelet feature



(b) ULBP feature

Figure 6. FAR and FRR curves of Logitboost-SO on training and testing set.

The performances of using ULBP feature are worse than those of four-level Haar wavelet features. This result agrees with most experiments of related iris recognition literature in which the local features always perform better than global features. That is because, although the global features are invariant to rotation, scale and translation, global features are statistical information. Hence the extracted statistical features can only give indistinct information.

TABLE 1. COMPARISON OF DIFFERENT LEVELS OF HAAR WAVELET

Haar wavelet		
Training set	***FAR***	***FRR***
2-level	0.273151	0.313725
3-level	0.154558	0.184874
4-level	0.084045	0.088235
Testing set	***FAR***	***FRR***
2-level	0.272485	0.443969
3-level	0.155489	0.233405
4-level	0.083933	0.115632

TABLE 2. VERIFICATION PERFORMANCE COMPARISON OF BOOSTING

4-level Haar wavelet		
Training set	***FAR***	***FRR***
Discrete Adaboost (8×4)	0.084045	0.088235
Discrete Adaboost (32×13)	0.023358	0.039216
Real Adaboost	0.000000	0.000000
Logitboost	0.000000	0.077031
SOBoost	0.104189	0.061625
Logitboost-SO	**0.024844**	**0.037815**
Testing set	***FAR***	***FRR***
Discrete Adaboost (8×4)	0.083933	0.115632
Discrete Adaboost (32×13)	0.023368	0.113490
Real Adaboost	0.000891	0.251249
Logitboost	0.000322	0.194147
SOBoost	0.104293	0.054961
Logitboost-SO	**0.024880**	**0.070664**
ULBP		
Training set	***FAR***	***FRR***
Discrete Adaboost (8×4)	0.056420	0.064426
Discrete Adaboost (32×13)	0.034094	0.054622
Real Adaboost	0.010614	0.000000
Logitboost	0.000051	0.152661
SOBoost	0.078475	0.047619
Logitboost-SO	**0.049455**	**0.046218**
Testing set	***FAR***	***FRR***
Discrete Adaboost (8×4)	0.059885	0.103498
Discrete Adaboost (32×13)	0.033032	0.126338
Real Adaboost	0.029831	0.121342
Logitboost	0.001057	0.360457
SOBoost	0.083388	0.053533
Logitboost-SO	**0.053907**	**0.064954**

TABLE 3. IDENTIFICATION PERFORMANCE COMPARISON OF BOOSTING

Recognition Rate (%)		
Algorithms	***1NN***	***3NN***
4-Haar + Discrete Adaboost (8×4)	77.3019	77.5161
4-Haar + Discrete Adaboost (32×13)	90.1499	89.2934
4-Haar + Real Adaboost	91.8630	92.0771
4-Haar + Logitboost	93.3619	93.3619
4-Haar + SOBoost	91.8630	91.2206
4-Haar + Logitboost-SO	**94.4325**	**94.2184**
Recognition Rate (%)		
Algorithms	***1NN***	***3NN***
ULBP + Discrete Adaboost (8×4)	75.8030	75.8030
ULBP + Discrete Adaboost (32×13)	83.0835	81.5846
ULBP + Real Adaboost	84.3683	83.2976
ULBP + Logitboost	84.1542	84.3683
ULBP + SOBoost	88.2227	87.1520
ULBP + Logitboost-SO	**89.7216**	**87.7944**

Since we have known that Logitboost-SO performs better than the existing boosting algorithms, the next step of our experiment is to optimize the performance of Logitboost-SO. Since Logitboost-SO uses the confidence rule of SOBoost, we can also improve the performance by adjusting the gradient of the confidence function α. From Fig. 8 we can see clearly that when α is getting smaller the speed of convergence is getting slower. There is a trade-off between growth rate and recognition accuracy.

After the most suitable parameter α has been found, and is applied to replace the initial default setting $\alpha = 1$, we use a two-stage classification scheme to make a

 ISBN: 978-1-61208-139-7

decision. In practice, if the confidence score obtained from one of the feature type is too close to 0 and the magnitude is lower than the threshold, another feature type is introduced to provide confidence score instead. Tables 4 and 5 provide comparisons of individual and combined feature type for verification and identification applications respectively. It is clear that cascading system outperforms the system of single feature type.

TABLE 4. VERIFICATION PERFORMANCE COMPARISON OF INDIVIDUAL AND COMBINED FEATURE TYPES

Testing Error		
Algorithms	***FAR***	***FRR***
4-Haar + Logitboost-SO	**0.018772**	**0.073519**
ULBP + Logitboost-SO	**0.063165**	**0.052106**
(4-Haar + Logitboost-SO) →(ULBP + Logitboost-SO)	**0.021419**	**0.030692**
(ULBP + Logitboost-SO) →(4-Haar + Logitboost-SO)	**0.020296**	**0.034975**

TABLE 5. IDENTIFICATION PERFORMANCE COMPARISON OF INDIVIDUAL AND COMBINED FEATURE TYPES

Recognition Rate (%)	
Algorithms	***Accuracy***
4-Haar + Logitboost-SO	**94.6467**
ULBP + Logitboost-SO	**92.5054**
(4-Haar + Logitboost-SO) →(ULBP + Logitboost-SO)	**97.0021**
(ULBP + Logitboost-SO) →(4-Haar + Logitboost-SO)	**94.8608**

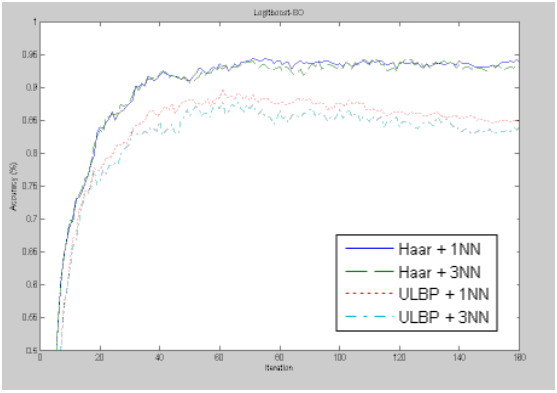


Figure 7. Recognition accuracy of Logitboost-SO on training and testing set.

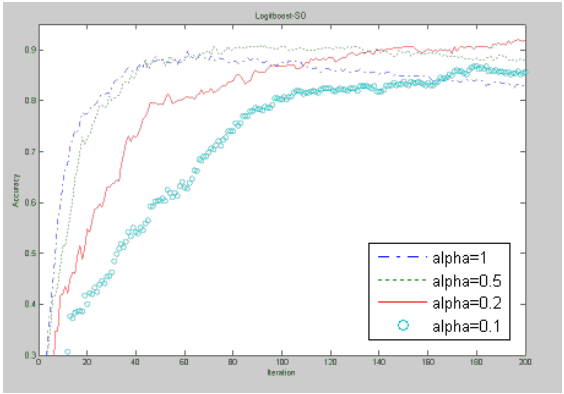


Figure 8. Recognition accuracy of Logitboost-SO using different α.

IV. CONCLUSION

In this paper, we proposed a novel iris recognition method based on Logitboost-SO and the cascading strategy. Unlike existing variants of Logitboost, Logitboost-SO combines Logitboost and SOBoost by using multi-weight update rule. To optimize the performance further, we adjust the gradient of the confidence function. Moreover, the local and global features are cascaded to provide complementary information. Our experimental results present evidence that Logitboost-SO outperforms former boosting algorithms and the cascading system can improve the performance further.

REFERENCES

[1] J. G. Daugman, "High confidence visual recognition of persons by a test of statistical independence," IEEE Trans. on Pattern Analysis and Machine Intelligence, vol. 15, no. 11, pp. 1148-1161, 1993.

[2] W. W. Boles and B. Boashash, "A human identification technique using images of the iris and wavelet transform," IEEE Trans. on Signal Processing, vol. 46, no. 4, pp. 1185-1188, Apr. 1998.

[3] L. Ma, T. Tan, Y. Wang and D. Zhang, "Personal identification based on iris texture analysis," IEEE Trans. on Pattern Analysis and Machine Intelligence, vol. 25, no. 12, pp. 1519-1533, Dec. 2003.

[4] Z. Sun, Y. Wang, T. Tan and J. Cui, "Improving iris recognition accuracy via cascaded classifiers," IEEE Trans. on System, Man and Cybernetics, vol. 35, no. 3, pp. 435-441, Aug. 2005.

[5] J. G. Daugman, "Probing the uniqueness and randomness of iriscodes: Results from 200 billion iris pair comparisons," Proceedings of the IEEE, vol. 94, no. 11, pp. 1927-1935, Nov. 2006.

[6] C.-T. Chou, S.-W. Shih, W.-S. Chen V. W. Cheng and D.-Y. Chen, "Non-orthogonal view iris recognition system," IEEE Trans. on Circuits and Systems for Video Technology, pp. 417-430, vol. 20, no. 3, Mar. 2010.

[7] M. Dettling and P. Buhlmann, "Boosting for tumor classification with gene expression data," Bioinformatics, vol. 19, no. 9, pp. 1061-1069, 2003.

[8] Y. D. Cai, K. Y. Feng, W. C. Lu and K. C. Chou, "Using Logitboost classifier to predict protein structural classes," Journal of Theoretical Biology, vol. 238, no. 1, pp. 172-176, Jan. 2006.

[9] S. Kotsiantis, E. Athanasopoulou and P. Pintelas, "Logitboost of multinomial Bayesian classifier for text classification," International Review on Computers and Software, vol. 1, no. 3, 2006.

[10] Y. Freund and R. E. Schapire, "Experiments with a new boosting algorithm," Machine Learning: Proceedings of the Thirteenth International Conference, pp. 148-156, 1996.

[11] J. Friedman, T. Hastie and R. Tibshirani, "Additive logistic regression: a statistical view of boosting," The Annals of Statistics, vol. 38, no. 2, pp. 337-407, Nov. 2000.

[12] Z. He, Z. Sun, T. Tan, X. Qiu, C. Zhong and W. Dong, "Boosting ordinal features for accurate and fast iris recognition," Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition, Jun. 2000.

 ISBN: 978-1-61208-139-7

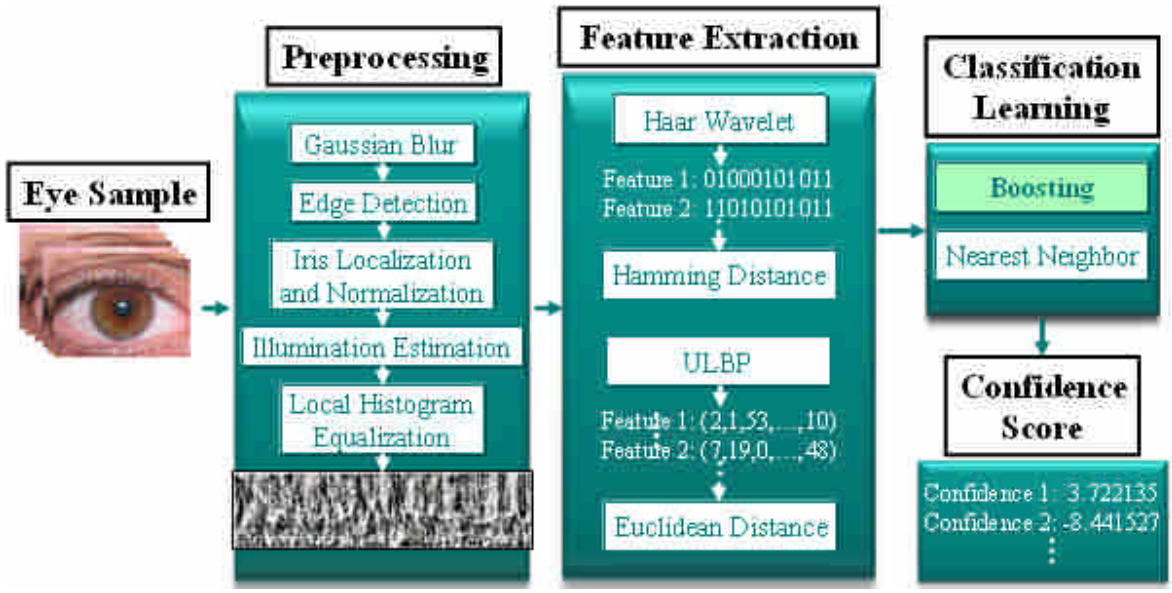


Figure 2. The system architecture.

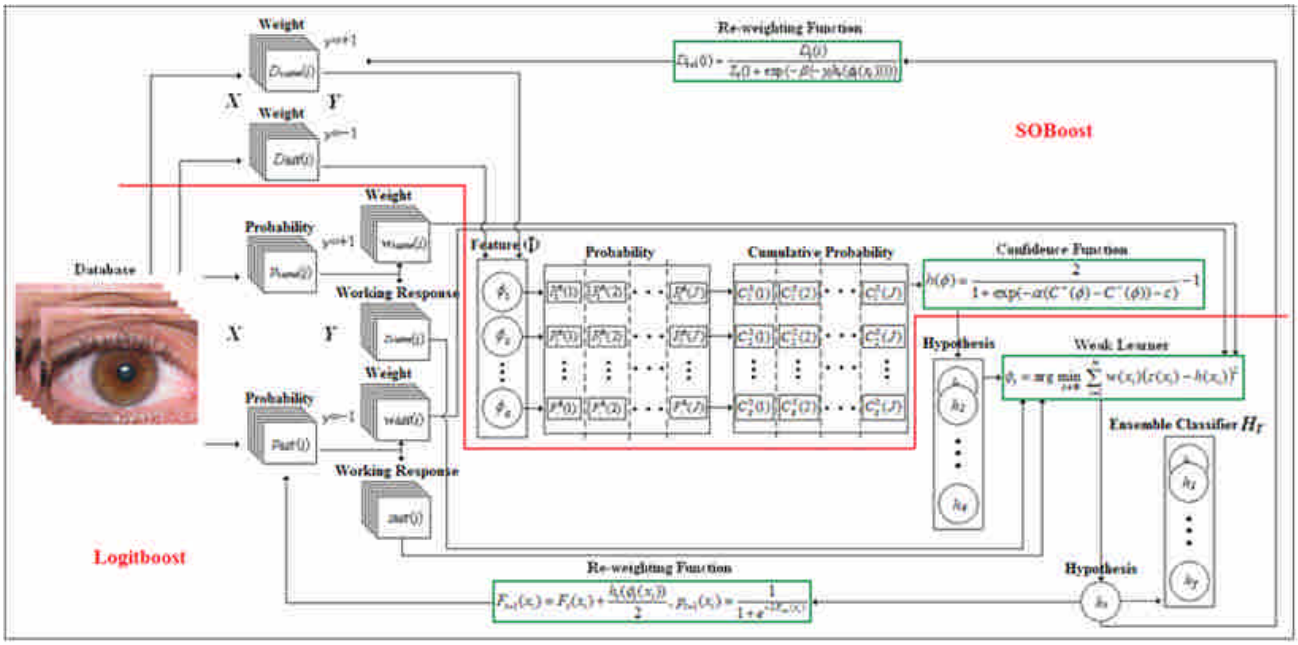


Figure 5. The flowchart of Logitboost-SO machine learning algorithm.

 ISBN: 978-1-61208-139-7

A Webshop for Digital Assets in Virtual Worlds Supported by a 3D Object Representation.

Michael Erwin Steurer
Institute for Information Systems and Computer Media
Graz University of Technology, Austria
michael.steurer@iicm.tugraz.at

***Abstract*—Virtual worlds became very popular over the past years and users spend more and more time in it. One of its main characteristics is the creation of content and a working economy that allows users to buy, sell, and even trade these objects. Existing multi-vendor platforms for virtual items lack of user friendliness for customers and merchants. In this paper we present the design of a multi-vendor webshop for virtual items enhanced by a 3D representation within the virtual world. Merchants can directly add items from their avatar's inventory to the webshop and promote it with additional meta-information. Customers can choose items from the webshop and additionally get a 3D preview prior to buying them. With the provided solution merchants do not need in-world stores to offer their objects because they are directly added to the webshop from their inventory. Customers can find products at a centralized webshop to search, compare and rate objects but use existing in-world mechanisms to buy them. The presented design has been implemented as a prototype to prove the concept.**

***Keywords*-3D Virtual Worlds, Selling Platform, OpenSimulator, Second Life, Open Metaverse**

I. INTRODUCTION

Virtual Worlds as depicted in Figure 1 are computer generated three dimensional environments and where first mentioned in Neil Stevenson's novel "Snow Crash" in 1992 [1]. The human representation within this environment (*i.e.*, residents of the virtual world), are referred to as avatars. In contrast to Massively Multiplayer Online Games virtual worlds are not quest oriented but focus on creativity, user interaction and are open-ended [2], [3]. According to Sivans 3D3C properties a 3D world simulates the real world, allows avatars to create social networks, create content, and trade with objects [4]. The virtual world tries to imitate the real world but avatars have additional capabilities like flying or teleporting to far distanced places. They can move around in this virtual world freely and interact with other avatars by using text chat, voice chat and limited forms of gestures. Users can create objects within the virtual world that base on primitive shapes like boxes or cylinders. Objects can either embed other objects or simple scripts to do autonomous tasks (*e.g.*, send embedded items to the avatar touching the object). Avatars store objects they own in their inventory. It is organized by folders and can not be accessed by other avatars. Users can "rezz" (in virtual world's jargon)

Figure 1. A typical scene from the virtual world of Second Life.

objects from their avatar's inventory which means that they make the objects appear in the virtual world. A requirement for building and creating objects is the permission of the landowner. [5]

An important factor for economy in a virtual world is a virtual currency that can be exchanged to real currency [6]. Virtual money like the Linden Dollar (L$) for Second Life or the Open-Metaverse-Cent (OM¢) for OpenSimulator based virtual worlds is directly bound to real currency. Therefore, avatars can buy, sell and even trade with virtual objects for real money [7].

Todays most successful virtual worlds are Second Life (www.secondlife.com) with 800.000 average monthly repeated logins, OpenSimulator (www.opensimulator.org) based virtual worlds with over 13.000 active users in January 2011 and Habbo Hotel (www.habbo.com) with 18 million unique logins per month [8], [9], [10], [11].

Most virtual worlds are designed as server-client architecture and users have to download a special viewer software to connect the servers. To access a virtual world users register and provide username and password. After logging in with these credentials the viewer software renders and displays the virtual world on the user's computer and provides an interface for interaction with mouse and keyboard.

The virtual world of Second Life is maintained by Linden Labs but due to it's closed source servers developers can not add any functionality. Basing on the communication protocol between the world of Second Life and the viewer

 ISBN: 978-1-61208-139-7

software a project named OpenSimulator project was founded in 2007 by Darren Guard. It is an open source 3D application server with an extensible and modular architecture. Every application server represents a particular area of virtual land, also referred to as *region*, that form a virtual world if connected to each other. Due to this network structure a virtual world it is also referred to as *grid*. Due to the modular design and the open source of the application server developers can implement new modules and extend it's capabilities. The framework presented in this paper can be applied to other open source virtual worlds but we focus on OpenSimulator based virtual worlds.

The remainder of the paper is organized as follows. In Section II we give a overview of existing selling platforms that are either web based or completely in-world. Section III describes the usage of the system and lists the necessary requirements. Basing on this Section IV points out the design and the prototypical implementation of the webshop. Finally, Section V concludes the paper and gives ideas for further work.

II. Existing Selling Platforms

In open ended virtual environments with a high degree of freedom and user interaction vanity has a large influence on the behavior of users [12]. They want to be entertained when spending time in-world and a lack of usability or complex usage can prevent users from buying and selling objects [12]. A presentation by the founders of Second Life states that 75% of all Second Life users were buyers and 25% of them were sellers[13]. Due to this it's even more important to bring the ease of use to both parties. In 2010 Second Life had on average 485.000 monthly economy participants. This number are residents with at least one activity on their virtual money account. The sales volume of their webshop was approximately 2.6 Million Euro in the fourth quarter of 2010 [9].

Merchants can offer their virtual items either in in-world stores or in webshops that are connected to the virtual world.

A. In-World Stores

In order to sell items merchants have to rent a salesroom to present their products. See Figure 2 for an in-world store that offers flowers in the OpenSimulator based virtual world of *German Grid*. Depending on the location of the salesroom the costs for renting can vary. The prices depend on the avatar traffic at the particular location and supply and demand of stores. Merchants with a small portfolio of products hesitate to invest in a such a shop because of the little income if compared to the effort to decorate and promote these items. Although items can be modified in a limited way through object-embedded scripts in-world stores lack of automated mechanisms for modification. Objects are

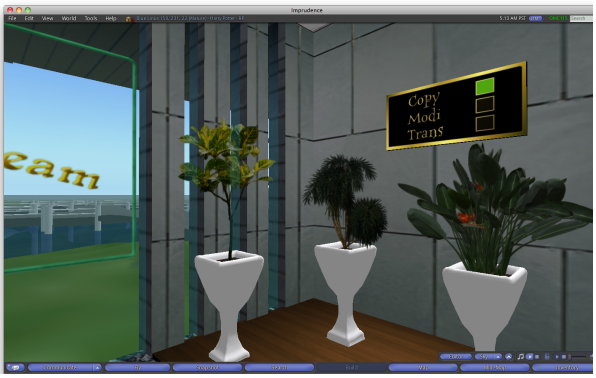


Figure 2. In in-world stores all objects are already present in the virtual world and can be bought by customers by touching the object.

created statically in the store and a users have to delete them manually. Customer buy object in in-word stores by clicking on the desired item and use grid internal mechanisms for payment and consignment. All bought objects are immediately delivered to the avatar's inventory.

Shops in virtual worlds are quite similar to shops in the real world. Avatars have to visit them to see and finally buy items. There is no automatic recommendation system but customers are attracted by other items in the shop.
Searing for items is important for a positive shopping experience but in OpenSimulator based virtual worlds there are no mechanisms to search for sellable objects or items [14], [3]. Interested users can only walk or fly through the virtual world and find objects and items accidentally. The combination of a missing search function and a locally limited market make the comparison of similar objects nearly impossible. [12] and [15] focus on these in-world stores and present studies about the buying behavior of users within a virtual world.

In-world stores are more similar to real world shops and users immediately see which objects they buy. Unfortunately, these stores also have drawbacks for customers and merchants.

B. External Webshops

There are several webshops for items in virtual worlds. SL Marketplace (http://marketplace.secondlife.com) for Second Life, and GridBay (http://www.gridbay.de) and Avatar Marketplace (http://avatarmarketplace.com) for Second Life and OpenSimulator based virtual worlds. As depicted in Figure 3 merchants put the items $m_{1..n}$ into an in-world object (also referred to as "magic box") that acts as a vault and all items in it are indexed by the webshop, $index(C_M, m_{1..n})$. Merchants immediately see their added items on the webpage and can add a more detailed textual description, some photos and extra keywords for the search engine.

 ISBN: 978-1-61208-139-7

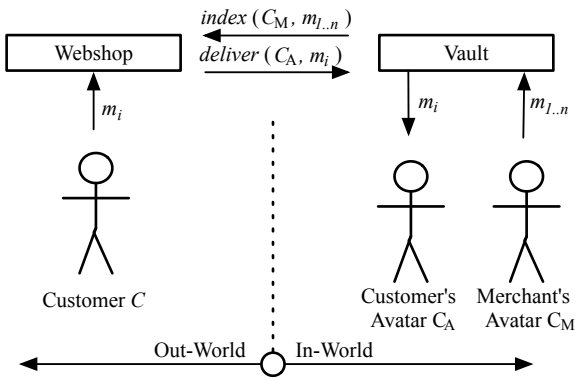


Figure 3. In existing webshop Merchants put items $m_{1..n}$ into an in-world vault which are automatically added to the webshop. A customer buys and object m_i from the webshop and the request to the vault initializes the consignment.

Customers put the desired items m_i from the webshop into a virtual trolley and do the checkout. All payments are made on the webshop's side in order to deliver the objects. As all items remain in-world the webshop sends a request with identifiers for avatar and bought objects to the vault to initialize the consignment of the objects, *deliver*(C_A, m_i) [16]. This requires the webshop to maintain a separate payment system but on the other hand it can serve different virtual worlds with different currencies.
External webshops benefit from well known techniques like browsing through items, filtering, or even comparing them but only provide the meta-information to their customers because the actual object remains in the virtual world. A common technique to motivate customers for further purchases are recommendation system that base on other users buying behavior.

Webshops are well known by users and they gain more information if compared to the plain 3D object. Webshops can provide recommendation systems for similar objects and customer reviews from other customers.

III. Functionality and Requirements

The basic idea of this paper is to describe a webshop enhanced by the benefits of an in-world store.

A. Usage and general description

Merchants have to register their avatar with the webshop and connect an in-world avatar to the account. After logging into the webshop they are provided with a list of objects retrieved from their avatar's inventory and can select the objects to be added to the offered items in the webshop. Due to security and privacy reasons merchants can only access a predefined subfolder in their inventory and are only allowed to add self-created objects to the webshop. To promote objects merchants can add additional product meta-information for the customers. The selling price is taken from the in-world properties of the object and can not be changed.

To gather data about the buying behavior and to improve the shopping experience customers are also required to register with the service and connect an avatar to this account in order to buy items. Then they can browse the webshop to find items upon the given meta-information or keywords. In existing webshops customers put their items in a virtual trolley and checkout to complete the purchase. In the proposed solution all items added to the trolley are automatically rezzed from the merchants inventory to the virtual world. The avatar that is connected to the actual account gets a notification with the accurate location of the rezzed objects and teleport to this location to inspect the created object. From this point the object behaves just like an object from an in-world store and customers can use the same mechanisms to buy the object. The purchase is initialized in-world and the entire payment and consignment mechanism is done in-world. Figure 4 gives an overview of the functionality of the webshop.

B. Requirements

The functionality of such a selling platform can be tracked down to the following requirements:

- Ease of use for merchants. Merchants are able to access their inventory from the webshop to select items to be listed on the selling platform and do not need a salesroom or vault in the virtual world. They can promote their objects by adding keywords, a better description, and pictures.
- 3D representation. Customers get a 3D preview of the objects within the virtual world prior to buying them with in-world mechanisms.
- Meta-information. Besides the three dimensional representation customers can retrieve more information about objects. This includes a detailed textual description, pictures, and information about the creator.
- Payment system and consignment. All purchases are done in-world. There is no external payment and consignment system but existing in-world mechanisms to transfer the money and send objects to the customers are used instead.
- Multivendor search and recommendations. Customers can search the entire webshop upon the item's names, their description and the set keywords. For every item the customer is provided with with a list of similar and related objects in the webshop.
- Security and trust. Merchants can define the permissions of the objects to be rezzed as a users's preview. Further, all objects created for customers have a certain time to life. If users do not buy the items they are

 ISBN: 978-1-61208-139-7

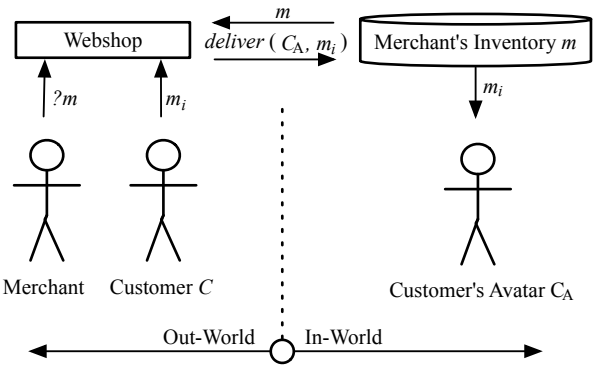


Figure 4. Merchants can request their avatar's inventory items m through the webshop ($?m$). A customer buys an object m_i and the webshop directly accesses the merchant's inventory to consign the actual object, $deliver(C_A, m_i)$.

periodically purged from the virtual world. Instead of deleting these objects they are return to the merchants inventory to prevent from data loss.

To demonstrate all these features we have implemented a prototype selling platform.

IV. Implementation

Users can interact with the virtual world using the client viewer. It detects user actions and sends the according requests to the application server (*e.g.*, rezz objects from the avatar's inventory in the virtual world and delete them). The design of the proposed solution has the same client-server architecture but the graphical client viewer is replaced by the webshop. Instead it bypasses the standard login procedure with the avatars standard log-in credentials and directly connects to the 3D application server. Hence, the webshop can send requests to the application server upon a user's request. As depicted in Figure 5 the webshop provides a web interface for users and sends these requests to the application server.

A. OpenSimulator Module

OpenSimulator based virtual worlds are a network of connected application servers and developers can extend a simulators functionality by adding modules. To provide this additional functionality to the entire virtual world every application server needs the module. It can detect requests sent by the webshop, extract the passed parameters and process them.
The developed module has an XML-RPC interface but does not support encryption. Hence, all requests sent to the module are not protected against wiretapping and forgery. For a prove of concept the unencrypted communication is sufficient but the productive use requires a more secure protocol as described in [2]. The module processes the XML-RPC requests and responds data in JSON format. In the prototypical implementation there are request to get a user's inventory, rezz items from the inventory to the virtual world and also delete these objects from the virtual world. In the following we describe it's basic functionality:

1) Access Inventory: Avatars as well as items in the avatar's inventory are identified by a unique identifier. Users can specify wether these objects are sellable, the price for the customers, and the permissions for the next owner. An XML-RPC request to get the list of an avatar's inventory contains the identifier of the avatar and replies with the data in JSON format. This response includes the folder structure of the entire inventory, unique identifiers, names, description, and the permissions of the objects. Due to privacy reasons merchants can specify a subfolder for the request to limit the response.

After fetching the inventory merchants can specify which items to be listed on the webshop. Hence, all items in the webshop are linked to virtual objects in the inventory of an avatar. If a merchant deletes one of the listed items in the avatar's inventory but not in the webshop the link is broken. To prevent from this scenario and to keep the items in the webshop up to date there is an additional request to the avatar's inventory. It contains identifiers for the avatar and a certain item in the inventory to check if it still exists. This request is executed periodically to purge all zombie records in the webshop.

2) Create and Delete Objects: Existing client viewer allow users to rezz items from their avatar's inventory only if the region owner allows it and the introduced framework does not bypass this system. After retrieving the list of inventory items users can request the 3D application server to make an object from an avatar's inventory appear in the virtual world. Besides the identifiers for the avatar and the particular item this request also requires information about the location of the new object. This location can be chosen freely but requires the user to have sufficient privileges to create items. The object appears in-world as if it was directly rezzed from the merchant's inventory by using the client viewer.

De-rezzing an object requires the identifier of the actual item and the creator of the object. In order to prevent from data loss we do not delete the object directly but move it automatically into a folder of the merchant's inventory.

3) Notify avatars: To inform avatars about rezzed objects we can send message to them. The parameters passed with the request are an identifier to specify the avatar and the actual message.

 ISBN: 978-1-61208-139-7

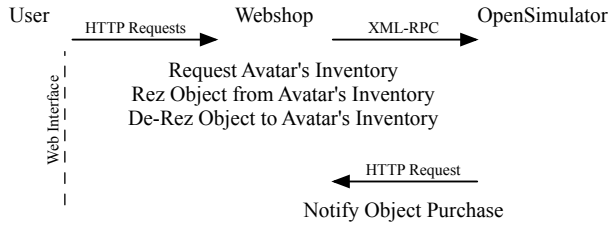


Figure 5. Users access the webshop by web interface and can request their inventory, rezz objects from the inventory and also de-rezz these items. After a successful purchase the simulator sends a notification to the webshop.

B. Webshop

It contains the database and the control logic of the entire architecture. Customers can access the webshop through web-interface and the webshop can access the OpenSimulator module through XML-RPCs. In order to use the service, both customers and merchants have to register with the webshop and link their avatars to the actual account. The registration requests the user to enter a username, a password and the name of the related in-world avatar. The webshop creates a secret confirmation PIN and sends it as private message to the specified avatar within the virtual world. If the name of the avatar is correct and the user has the credentials to access the avatar's inbox it can complete the registration and enter the secret PIN in the webshop's registration form. This mechanism proves the existence of the avatar and confirms the link between webshop account and in-world avatar [17]. Both account types are identical and so merchants can buy objects and vice versa.

Users have to log in and can either access their inventory to add items to the webshop or browse the existing catalogue. After fetching the inventory the merchant is provided with information of the avatar's inventory (*i.e.*, name, description, permissions), and can add these items to the webshop. To promote them users can add meta-information to the selected items that is stored locally in the webshop's database but directly linked to the inventory items. Meta-information can contain a more detailed description, keywords for classification, and some pictures for the customers to get a first glance of the offered items.

We have implemented a simple search function that bases on the keywords of the listed objects. Hence, customers can search the webshop for desired items, get a list of matching items choose objects from the list upon the name, detailed description and pictorial information.

There is no checkout as in existing webshops but a customer can initialize to rezz the object in-world instead. To do so, the webshop sends a request to the OpenSimulator module with identifiers for merchant and actual item. No matter whether the merchant's avatar is logged in or not the item appears at a certain position in-world. To inform the customer about this location an instant message is sent to it's avatar with the detailed information. Now, the customer logs into the virtual world to access the message and locate the rezzed object. The customer can either buy the objects in-world with in-world purchase mechanisms or continue shopping. If the customer decides to continue shopping and select a new object to be rezzed in-world the webshop requests the simulator to remove the old object and replace it with the new one. Otherwise the rezzed object will be automatically purged after a certain period of time again by webshop request. The OpenSimulator module can not only detect XML-RPC from the gateway but also trigger on object purchases. As a customer decides to buy an in-world object the module detects this event and sends a HTTP request to the webshop. With this method the webshop detects all purchases and obtains additional information for customers (*e.g.*, recommendation systems).

V. Results and Conclusion

Existing solutions that list virtual 3D objects in a 2D webshop are complicated to use. The proposed idea has the following benefits if compared to existing solutions:

- Registration Process. The registration process proves the link between the webshop account and the in-world avatar. This is required in order to access an avatar's inventory to add items to the webshop and to inform customers about rezzed objects.
- Access Inventory. Instead of "magic boxes" users can easily access their inventory from the webshop to offer items. They can add additional meta-data to give customers a more detailed description of the items.
- Purchasing Objects. Customers can employ all features of a webshop to search and compare items. The use of keywords and knowledge about previous payments can be used as a recommendation system.
- Consignment and Payment. The webshop just provides meta-information about the objects and the actual object remains inside the virtual world. Within the webshop users can request the selected object to be rezzed in-world to use in-world mechanisms for purchasing objects. This implies that the webshop does not need a separate payment module.

The complete framework of the merchant system has been implemented as a prototype without any security measures to prevent a users privacy. Obviously, this system can not be used for productive use but is a simple proof of concept.

For future work we will add the necessary functionality for security and test the shop with a limited number of merchants.

 ISBN: 978-1-61208-139-7

REFERENCES

[1] N. Stephenson, *Snow Crash.* Bantam Books, 1992.

[2] F. Kappe and M. Steurer, "The open metaverse currency (omc) - a micropayment framework for open 3d virtual worlds," in *EC-Web*. ACM, 2010.

[3] M. Q. Tran, "Understanding the influence of 3d virtual worlds on perceptions of 2d e-commerce websites," in *Proceedings of the 2nd ACM SIGCHI symposium on Engineering interactive computing systems - EICS '10*, 2010.

[4] Y. Sivan, "3d3c real virtual worlds defined: The immense potential of merging 3d, community, creation, and commerce," *Journal of Virtual Worlds Research*, vol. 1, no. 1, 2008.

[5] M. Rymaszewski, W. J. Au, M. Wallace, C. Winters, C. Ondrejka, B. Batstone-Cunningham, and P. Rosedale, *Second Life: The Official Guide*. Alameda, CA, USA: SYBEX Inc., 2006.

[6] H. Yamaguchi, "An analysis of virtual currencies in online games," http://ssrn.com/abstract=544422, 2004, [accessed on 19-April-2011].

[7] P. Crowther and R. Cox, "Building content in second life issues facing content creators and residents," in *Computer-Mediated Social Networking*, ser. Lecture Notes in Computer Science. Springer Berlin / Heidelberg, 2009.

[8] Sulake, "Habbo hotel – where else?" http://www.sulake.com/habbo/, 2011, [accessed on 19-April-2011].

[9] Linden Lab, "The second life economy in q4 2010," http://blogs.secondlife.com/community/features/blog/2011/01/26/the-second-life-economy-in-q4-2010, 2010, [accessed on 19-April-2011].

[10] M. Korolov, "June 2010 opensim grid statistics," http://www.hypergridbusiness.com/statistics/june-opensim-grid-statistics/, 2010, [accessed on 19-April-2011].

[11] M. Korolov, "December 2010 opensim grid statistics," http://www.hypergridbusiness.com/statistics/december-opensim-region-counts/, 2010, [accessed on 19-April-2011].

[12] Y. Guo and S. Barnes, "Virtual item purchase behavior in virtual worlds: an exploratory investigation," *Electronic Commerce Research*, pp. 77–96, 2009.

[13] S. Papagiannidis, M. Bourlakis, and F. Li, "Making real money in virtual worlds: MMORPGs and emerging business opportunities, challenges and ethical implications in metaverses," *Technological Forecasting and Social Change*, pp. 610–622+, 2008.

[14] OpenSimulator, "Opensimsearch," http://opensimulator.org/wiki/OpenSimSearch, 2011, [accessed on 19-April-2011].

[15] P. R. Messinger, E. Stroulia, K. Lyons, M. Bone, R. Niu, K. Smirnov, and S. Perelgut, "Virtual worlds past, present, and future: New directions in social computing," *Decision Support Systems*, vol. 47, no. 3, 2009.

[16] M. E. Steurer and F. Kappe, "A micropayment enabled webshop for digital assets in virtual worlds," in *Proceedings of the 14th International Academic MindTrek Conference: Envisioning Future Media Environments*, ser. MindTrek '10. ACM, 2010.

[17] C. Trattner, M. E. Steurer, and F. Kappe, "Socializing virtual worlds with facebook: a prototypical implementation of an expansion pack to communicate between facebook and opensimulator based virtual worlds," in *Proceedings of the 14th International Academic MindTrek Conference: Envisioning Future Media Environments*, ser. MindTrek '10. ACM, 2010.

 ISBN: 978-1-61208-139-7

A Hybrid and Auto-adjusted Spam Filter

Shu Bin Chen
Department of Computer Science and Communication Engineering
Providence University
Taichung, Taiwan, R.O.C.
g9873007@pu.edu.tw

Hsiao Ping Lee
Department of Applied Information Sciences
Chung Shan Medical University,
Taichung, Taiwan, R.O.C.
ping@csmu.edu.tw

Tzu-Fang Sheu
Department of Computer Science and Communication Engineering
Providence University
Taichung, Taiwan, R.O.C.
fang@pu.edu.tw

Abstract—**The spam E-mail problem has become more and more serious today. Enterprises and users have to spend lots of time on filtering out useful messages from spam. A variety of spam filtering mechanisms had been proposed, including list-based method, behavior-based filter, content-based method, and cocktail filtering mechanisms. In order to improve the accuracy of spam filters, this paper proposes a novel spam detection system, which combines a behavior-based method and a blacklist method. The proposed system will analyze spam behaviors and then generate behavior patterns. Additionally, the propose system will use an auto-updated blacklist mechanism which collects the updates from anti-spam organizations. The proposed system also uses a feedback mechanism that adjusts the behavior patterns based on users' responses. Therefore, the proposed spam detection system can perform more efficiently and accurately.**

Keywords- spam; list-based method; behavior-based filter; content; Internet.

I. INTRODUCTION

Spam (unsolicited bulk mail, or spam mail) was defined as sending a lot of E-mail without recipients' authorization [1]. A large number of spam mails always waste not only users' time but also network/system resource. With the rapidly growing Internet, spam problem has become more and more critical. According to the security report from Symantec MessageLabs in 2010, it says that the average ratio of the global volume of spam E-mail is up to 89.1% [2]. Spam has become a critical research issue. A variety of spam filtering mechanisms had been proposed, including list-based method, behavior-based filter, content-based method, and cocktail filtering mechanisms.

More and more researches apply hybrid filtering technologies to spam filters to obtain higher accuracy rate. However, these methods are more complex than basic filtering technologies and so that result in low computing performance. Moreover, when the weights of the combined technologies are not configured appropriately, the system may perform badly. In this study, considering the efficiency and accuracy of the spam filter, we combine a Behavior-based method and a blacklist method as a hybrid and auto-adjusted spam filter.

This study is divided into four sections. Section I briefs the motive of the study, the faced problems and the goal of the study. Section II shows related background, describing the previous mechanisms of spam filter. Section III explains the system architecture. Section VI describes the expected result of this research.

II. BACKGROUND

A. *List-based filter*

The common list-based methods are based on blacklists and white-lists [6]. Blacklist spam filter blocks the messages based on senders' IP addresses or domains that listed in the database. Some anti-spam organizations release free blacklists, e.g., open relay blacklist. Contrarily, white-list filter only accepts the messages which senders' IP addresses or domains are listed in the database. The list-based spam filters have to update and maintain the lists regularly to keep the lists correct. However, the sender of spam E-mails are mostly fake, and thus the accuracy of list-based spam filter is not good. In general, list-based spam filters usually cooperate with other kind of filtering mechanisms [6].

B. *Content-based filter*

Content-based spam filters define keywords or signatures of the spam, and search the keywords in the E-mail content. While the keyword is found in message content, the message will be marked as spam. Content-based filter are divided into two categories based on the generation of keywords and the decision methods: heuristic filter and statistical filter. The heuristic-based filter defines the keyword manually. The statistical-based filter divides E-mail content into many small tokens, and uses statistical formulas to calculate the probability that the message is a spam based on the tokens. The content-based spam filter has good detection rate. However, content-based filter has to scan the entire message that results in low performance and high latency.

Jiansheng et al. [8] proposed a statistical-based Bayesian filter for spam detection. This method collects a large number of spam and non-spam messages, and then analyzes the messages. After the preprocessing and training phase, two spam token databases are constructed based on the level of damage of spam: ordinary spam database (including commercial and educational) and malignant spam database (including religious, political and pornography). The appearance probability of each token is calculated and stored in three hash tables: hash_good for non-spams, hash_bad for ordinary spam and hash_very_bad for malignant spam. For each incoming message, it has to be divided into small tokens. Then, based on Bayes Theorem and the appearance

 ISBN: 978-1-61208-139-7

probability of pre-sampled tokens in the hash tables, the probability of the incoming message being a spam can be calculated.

Jungsuk et al. [10] used heuristic feature selection method to artificially measure the availability of each feature. This system was composed of several phases. Firstly, it collects E-mail content and then extracts URLs in the content. Then, the URLs are forwarded to a crawler machine and a feature extractor. The crawler machine will download html content based on the URLs' locations. Then a cluster recognizer will calculate the similarity between the extracted URLs and the downloaded html content from the crawler machine, and then recognize spam mail clusters. A feature extractor will extract twelve features from spam mail and the extracted URLs. Finally the selector will use the obtained results to selected significant features. With the selected features, spam can be properly classified.

TABLEI COMPARISON OF COMMON FILTERING TECHNIQUES

Filter technology	Advantage	Disadvantage
List-based filter	Faster execution rate.	High false positive rate.
Content-based filter	Good accuracy.	Must scan full content of E-mails, extremely expensive system resource.
Behavior-based filter	Fast execution rate, fewer network delays, and no regular maintenance database.	High false positive
Anti-spam Cocktail	If properly adjusted, the accuracy rate is the highest.	If use improper weights for the techniques, the system performance will be bad.

C. Behavior-based filter

Usually, spam mails contain similar characteristics. By analyzing the behaviors of spam, some researches find that some characteristics, such as the sending source, sending time, or transmission frequency, can be used to determine whether the sender is a spammer. The behavior-based spam filter can detect spam by just part of E-mail information, instead of downloading full content of an E-mail, which results in low latency. Although behavior-based spam filter does not have to scan full content of messages and thus gets good efficiency, the accuracy of this kind of filters is not well. In order to improve the accuracy, the behavior-based filter often cooperates with other filtering methods.

Meizhen et al. [3] used mining techniques to analyzecollected E-mail log, which includes server IP, frequency, content length, etc. After analysis, the less relevant information will be omitted, and only the high-related features are used to describe the behavior of spam.

Because spam features may be described in form of discrete values or continuous values, the values have to be pre-processed by data conversion and data compression. Data conversion is done by Fuzzy-ID3 algorithm, and then Fuzzy decision tree rule base is obtained. According to the generated rules, the behavior pattern of spam can be understood clearly.

Meizhen et al. [4] observed the behavior pattern of outgoing messages to obtain E-mail features, such as message size, the number of attachment file, etc. Then the authors used these features to create user sending model. They analyzed the E-mails, and found that most spam messages had URLs. Hence URL model was created. The user sending model and the URL model were combined into the Bayesian filtering system to detect spam. The experimental result showed that the detection rate of this research [4] was higher than 85%.

Gert Vliek [9] collected suspicious IP addresses that have suspicious behavior, i.e., a suspicious spammer may send a lot of network traffic, but receive only single message or zero network traffic. To detect a spam machine, firstly the detection system has to find out suspicious machines. Then, the probability that the suspicious one is a spammer is calculated. When the value of probability is higher than a preset threshold, the suspicious machine may be spam machine.

D. Anti-spam Cocktail

A variety of filtering techniques are used together in the anti-spam cocktail filter, i.e., Bayesian filtering technique combines with blacklist to filter out spam E-mails [7]. In general, the cocktail-based approach has lower false positive, but has a big drawback. That is, if the weights of different filtering technologies are not deployed appropriately, the system will perform badly. Additionally, as the system combines different approaches, the cocktail-based spam filter becomes more complicated and expenses more system resources, and leads to relatively slow execution speed.

The spam filters mentioned previously have different advantages and disadvantages. The comparisons are listed in the Table I. Considering the efficiency and accuracy of the spam filter, this study will propose a hybrid spam filter that combines behavior-based and blacklist-based filtering techniques and also user's feedback mechanism achieve efficient performance and accuracy.

III. A HYBRID SPAM FILTER

Using two or more different filtering mechanisms usually achieves higher accuracy than using single filtering technique [5]. Therefore, this study develops a novel hybrid spam filter, combining behavior-based techniques and auto-updated blacklist. Because behavior-based method and blacklist-based method are criticized for their high false positive rate, a feedback mechanism is involved in the proposed system, which keeps adjusting the behavior model to make the system more accurate.

The proposed hybrid spam filter consists of three major mechanisms: a blacklist filtering mechanism, a behavior filtering mechanism, and a feedback mechanism. The blacklist mechanism will automatically renew the black IP addresses or domains by referencing the updates from anti-spam organizations. The behavior mechanism will analyze the collected messages and information to obtain useful features, and based on Bayes Theorem, the probability that

 ISBN: 978-1-61208-139-7

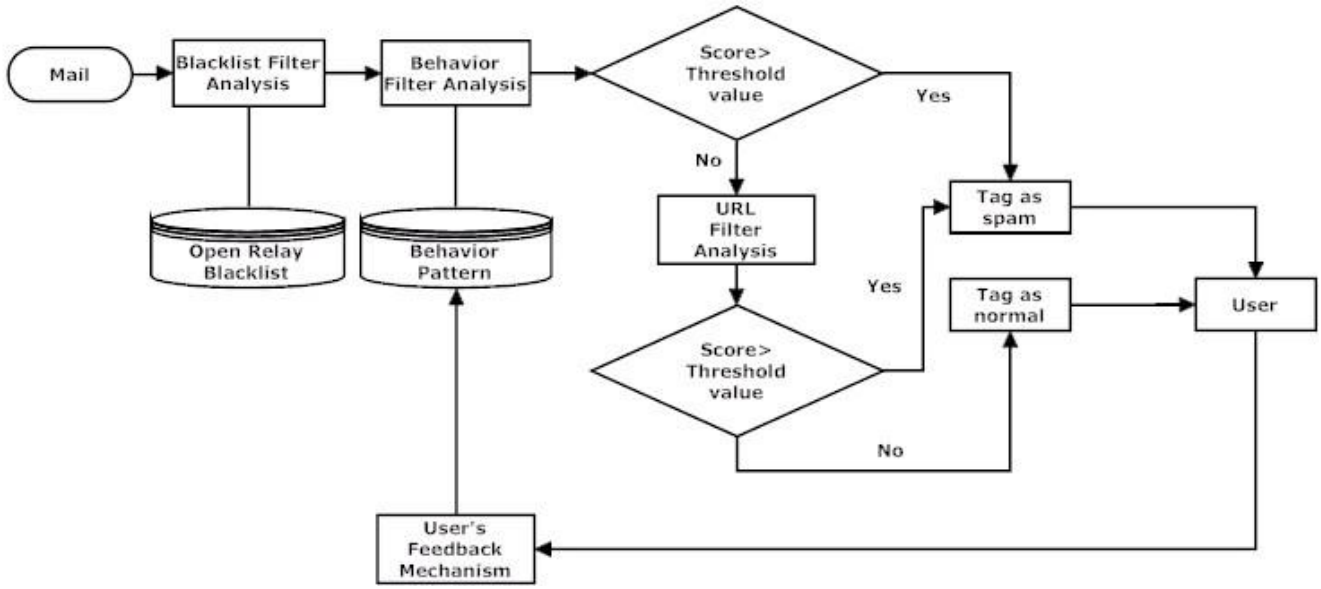


Figure 1. THE ARCHITECTURE OF THE HYBRID SPAM FILTER.

the incoming message is a spam is calculated. The behavior model is automatically adjusted based on users' feedback. Because the behavior-based and blacklist-based models are know as with the advantage of efficient execution speed, and the feedback mechanism can improve the precision, the proposed hybrid filter will achieve better performance.

The hybrid filtering system can be divided into four phases, and the system architecture is shown in Figure1.

Phase 1: With anti-spam organization provided blacklist, filter out the sender which is in the blacklist.

Phase 2: Using behavior model to detect the rest of the mail. Based on the behavior model and suspicious probability, each message will get a score. If the score is higher than a preset threshold, the incoming message is treated as a spam.

Phase 3: For the messages that passes the behavior-based filter, it will be put into a URL filter to check whether the message contains any URL that often considered as a spam URL, to determine whether the message is a spam. If its score is higher than the threshold, it is considered as a spam. On the contrary, it is considered as a normal message.

Phase 4: When users receive the message, they can check whether the decision is correct or not. The system provides a feedback mechanism for the users to respond their corrections. Base on the feedbacks from users, to the system will modify the behavior pattern to improve the accuracy of the filtering system.

IV. EXPECTED RESULTS

This study still works in progress. The proposed system will be implemented based on an open-source spam filter in the Apache SpamAssassin project [11]. The training data will be collected from the Providence University campus network and then used to establish the behavior model. While the blacklist-based filtering technique and behavior-based filtering do not require to scan full message and thus have the advantage of fast detection speed, and the feedback mechanism used in the proposed system can automatically adjust the behavior model to improve the detection accuracy, the proposed hybrid spam filter would have efficient detection speed and better accuracy.

ACKNOWLEDGEMENTS

The authors would like to thank the National Science Council of the Republic of China, Taiwan, R.O.C, for financially supporting this research under Grants 98-2218-E-126-001-MY2.

REFERENCES

[1] The Definition of Spam, http://www.spamhaus.org/definition.html.

[2] Symantec Announces MessageLabs Intelligence 2010 Annual Security Report, http://www.symantec.com/about/news/release/article.jsp?prid=20101207_01.

[3] W. Meizhen, L. Zhitang, and Z. Sheng , "A Method for Spam Behavior Recognition Based on Fuzzy Decision Tree," IEEE, Ninth International Conference on Computer and Information Technology, pp. 236-241, 2009.

[4] W. Meizhen, L. Zhitang, X. Ling, and Z. Yunhe, "Research on Behavior Statistic Based Spam Filter," IEEE, First International Workshop on Education Technology and Computer Science, pp. 687-691, 2009.

[5] Y. Hassan and E. Tazaki, "Rule Extraction Based on Rough Set Theory Combined with Genetic Programming and Its Application to Medical Data Analysis," IEEE, Intelligent Information Systems Conference, the Seventh Australian and New Zealand, pp. 385-390, 2001.

[6] C. Zhi Chen, "A Two Stage Spam Mail Filtering Method Based on Personal Mail," National Taiwan University of Science and Technology, Department of Computer Science and Information Engineering, 2007.

[7] S. Shih Neng, "Design and Implementation of A Personalized Chinese Spam E-mails Filtering System," National Dong Hua University, Department of Computer Science and Information Engineering, 2005.

 ISBN: 978-1-61208-139-7

[8] W. Jiansheng and D. Tao, "Research in Anti-Spam Method Based on Bayesian Filtering," IEEE, Computational Intelligence and Industrial Application, pp. 887-891, 2008.

[9] G. Vliek, "Detecting spam machines a Net-flow data based approach," Faculty of Electrical Engineering Mathematics and Computer Science, 2009.

[10] S. Jungsuk, E. Masashi, K. Hyung Chan, I. Daisuke, and N. Koji, "A Heuristic-based Feature Selection Method for Clustering Spam Emails," 17th International Conference on Neural Information Processing, pp. 290-297, 2010.

[11] Apache SpamAssassin project, http://spamassassin.apache.org/index.html.

 ISBN: 978-1-61208-139-7