Fifth International Conference

INFORMATION RESEARCH AND APPLICATIONS

26-30 June 2007, Varna

P R O C E E D I N G S

Volume 2

FOI ITHEA

SOFIA, 2007

Kr. Markov, Kr. Ivanova (Ed.)

Proceedings of the Fifth International Conference “Information Research and Applications” i.TECH 2007, Varna, Bulgaria

Volume 2

Sofia, Institute of Information Theories and Applications FOI ITHEA – 2007

ISSN: 1313-1109

First edition

Printed in Bulgaria by Institute of Information Theories and Applications FOI ITHEA

Sofia -1090, P.O. Box 775

e-mail: info@foibg.com

www.foibg.com

ISSN: 1313-1109

PREFACE

The Fifth International Conference "***Information Research and Applications***" (**i.TECH 2007**) is organized as a part of "ITA 2007 - Joint International Scientific Events on Informatcs".

ITA 2007 as well as the i.TECH 2007 is supported by

International Journal on Information Theories and Applications (IJ ITA)

and

International Journal on Information Technologies and Knowledge (IJ ITK)

i.TECH 2007 is dedicated to:

- 60th Anniversary of the Institute of Mathematics and Informatics of Bulgarian Academy of Sciences;
- 15th Anniversary of the Association of Developers and Users of Intelligent Systems (Ukraine);
- 10th Anniversary of the Association for Development of the Information Society (Bulgaria).

The aim of the conference is to be one more possibility for contacts for scientists. The usual practice of IJ ITA and IJ ITK are to support several conferences at which the papers may be discussed before submitting them for referring and publishing in the journals. Because of this, such conferences usually are multilingual and bring together both papers of high quality and papers of young scientists, which need further processing and scientific support from senior researchers.

We would like to express our thanks to all who support the i.TECH 2007 and especially to the *Natural Computing Group* (NCG) (http://www.lpsi.eui.upm.es/nncg/) of the Technical University of Madrid, which is leaded by prof. Juan Castellanos.

Let us thank the Program Committee of the conference for referring the submitted papers. Special thanks to prof. Viktor Gladun, prof. Alexey Voloshin, prof. Avram Eskenazi and prof. Luis Fernando de Mingo.

i.TECH 2007 Proceedings has been edited in the *Institute of Information Theories and Applications FOI ITHEA* in collaboration with the leading researchers from *Institute of Cybernetics "V.M.Glushkov", NASU (Ukraine)*, *Kiev University "T.Shevchenko" (Ukraine)*, *Institute of Mathematics and Informatics, BAS (Bulgaria)*, *Institute of Information Technologies, BAS (Bulgaria)*, *University of Calgary (Canada)*, *VLSI Systems Centre, Ben-Gurion University (Israel)*.

The i.TECH 2007 Conference found the best support in the work of Organizing Committee Chairman Ilia Mitov.

To all participants of i.TECH 2007 we wish fruitful contacts during the conference days and efficient work for preparing the high quality papers to be published in the International Journal "Information Theories and Applications" or in the International Journal "Information Technologies and Knowledge".

Varna, June 2007 Kr Markov, Kr. Ivanova

i.TECH 2007 has been organized by:

- Institute of Information Theories and Applications FOI ITHEA (Bulgaria)
- V.M.Glushkov Institute of Cybernetics of National Academy of Sciences of Ukraine
- Institute of Mathematics and Informatics, BAS (Bulgaria)
- Institute of Information Technologies, BAS (Bulgaria)
- Technical University of Madrid (Spain)
- Taras Shevchenko National University of Kiev (Ukraine)
- Kharkiv National University of Radio Electronics (Ukraine)
- Association of Developers and Users of Intelligent Systems (Ukraine)
- Association for Development of the Information Society (Bulgaria)
- International Journal “Information Theories and Applications”
- International Journal "Information Technologies and Knowledge"

Program Committee:

Victor Gladun (Ukraine)

Alexey Voloshin (Ukraine) — Krassimir Markov (Bulgaria)

Avram Eskenazi (Bulgaria) — Luis Fernando de Mingo (Spain)

Adil Timofeev (Russia)	Juan Castellanos (Spain)	Orly Yadid-Pecht (Israel)
Alexander Gerov (Bulgaria)	Koen Vanhoof (Belgium)	Peter Stanchev (USA)
Alexander Kuzemin (Ukraine)	Krassimir Manev (Bulgaria)	Radoslav Pavlov (Bulgaria)
Alfredo Milani (Italy)	Krassimira Ivanova (Bulgaria)	Rumiana Kirkova (Bulgaria)
Anna Kantcheva (Bulgaria)	Laura Ciocoiu (Romania)	Stanimir Stoyanov (Bulgaria)
Arkady Zakrevskij (Belarus)	Levon Aslanyan (Armenia)	Stefan Dodunekov (Bulgaria)
Ilia Mitov (Bulgaria)	Martin Mintchev (Canada)	Stoyan Poryazov (Bulgaria)
Ivan Popchev (Bulgaria)	Nelly Maneva (Bulgaria)	Vladimir Ryazanov (Russia)

Organizing Committee:

Ilia Mitov	Krassimira Ivanova	Stoyan Poryazov
Emilia Saranova	Rositsa Ovcharova	Todorka Kovacheva
Tsvetanka Kovacheva	Valeria Dimitrova	Vera Markova

Papers of i.TECH 2007 are collated in following sections:

- Multimedia Semantics and Cultural Heritage
- Knowledge Discovery and Engineering
- Transition P Systems
- Neural Nets
- Decision Making
- Software Engineering
- Advanced Technologies
- Distributed and Telecommunication Systems
- Cyber Security

Official languages of the conference are English and Russian.

General sponsor of the i.TECH 2007 is FOI BULGARIA (www.foibg.com).

TABLE OF CONTENTS – VOLUME 1

Transition P Systems

Neural Nets

TABLE OF CONTENTS – VOLUME 2

Advanced Technologies

Distributed and Telecommunication Systems

Cyber Security

About:

INDEX OF AUTHORS

Decision Making

SELF-ORGANIZING MAP AND CELLULAR AUTOMATA COMBINED TECHNIQUE FOR ADVANCED MESH GENERATION IN URBAN AND ARCHITECTURAL DESIGN

Álvaro Castro Castilla, Nuria Gómez Blas

***Abstract**: This paper presents a technique for building complex and adaptive meshes for urban and architectural design. The combination of a self-organizing map and cellular automata algorithms stands as a method for generating meshes otherwise static. This intends to be an auxiliary tool for the architect or the urban planner, improving control over large amounts of spatial information. The traditional grid employed as design aid is improved to become more general and flexible.*

***Keywords**: self-organizing map, cellular automata, CAD, CAAD, architectural computation.*

***ACM Classification Keywords**: J.6. Computer Applications - Computer-aided design*

Introduction

Architectural and urban design brings up a kind of problems inherently different from those found in engineering or science. In contrast with generalized belief, specially from the engineering and scientific domains, the main difficulties in the development of consistent CAD tools are more related to simultaneity of constrains and complexity of their parametrization, rather than in the topics of creativity and inspiration. Currently these tools are aimed to do little more than mimic the abilities of pen and paper, with the addition of basic copy-related commands.

Different methods for coping with that complexity are to be investigated from artificial intelligence and other research areas such as L-systems or fractals. In this paper a technique for supporting the designer's work in the early stages of the process is proposed through the combination of two models: self-organizing maps and cellular automata .

A self-organizing map (SOM) is a new, effective software tool for the visualization of high-dimensional data. In its basic form it produces a similarity graph of input data. It converts the nonlinear statistical relationships between high-dimensional data into simple geometric relationships of their image points on a low-dimensional display, usually a regular two-dimensional grid of nodes [Kohonen, 2001].

Cellular automata (CA) are mathematical and computational models for systems in which the global behavior is reached through the collaboration of multiple simple parts.

Motivations for cellular automata study can be found in different fields as vehicles for studying pattern formation and complexity. CA can be treated as abstract discrete dynamical systems embodying intrinsically interesting and potentially novel behavioral features [Ilachinski, 2001].

Problem description

Urban planning deals mainly with land units and classification, taking decisions at the large scale that affect each one smaller portion. The basic atom for urban design are those pieces of ground that introduce an extensive list

of parameters, such as area, use, cost per measuring unit, etc [Colonna, Di Stefano, Lombardo, Papini, Rabino, 1998]. A variable amount of them, acting together and usually seen statistically, are used by the designer to constrain and design others. In fact, those layers of information usually come one after the other but depending on upcoming ones. That implies cyclic processes and introduces the need for induction thinking.

Land units manifest their geometrical phenotype: the land divisions map and the set of rules that should govern its physical execution. The former is commonly worked out using a mesh for spatial structuration. Of course, there several other techniques, even unique ones depending on the project. This is a matter of process decision, however we can't -and should not- avoid some sort of deviation at most stages of the urban project. We will aim for the most general approach, in this case, mesh-based design.

Meshes introduce two issues; these are the limitations we want to solve with this proposal:

Fixed topology: basic design meshes have fixed topology, as they aim for simplicity and ease of use for the designer. In contrast with the usability, the information this meshes use to be built from is inherently dynamical in its topology, as one consequence of complexity in real-life originated data. In other words, spatial relationships are not kept over time in urban reality. So our tool should be able to assume that feature and develop the framework to process such variations.

Shape constrains: refers to the land division methods and geometrical structures that conform the actual plan. Historical and cultural diversity show us several different ways to accomplish this goal. However, any taken decision at this aspect would limit the method's generality. This suggests that building a mesh where basic units are not limited by enclosures, thus conforming shapes, but turning its basic units into something shapeless would be desirable. This is quite obvious but not a common design approach, as it implies a higher level of abstraction and a bigger leap towards materialization of a design idea.

The objective is to build a new mesh that improves on the basis of these two limitations for the specific task of helping the urban planner. After that we should be able to use it in order to evaluate its first results.

Orthogonal grid approaches

Urban behavior modeling based on cellular automata has been widely studied in recent years. Many of those models are based on the typical orthogonal mesh, the archetypical grid. The CA orthogonal lattice is commonly used as the most neutral -thus general- geometrical base. It turns variables which are spatially extensive into their density-intensity equivalents and this immediately means that comparisons can be made [Batty, Xie, Sun, 1999]. The geometric configuration of the spatial units used to represent the spatial data can have a profound effect, explaining why using spatial systems which neutralize the effect of configuration remove any bias caused by convoluted or distorting geometries.

The regular grid has other advantages, as the additional ability to work in layers without additional efforts to make different ones fit into the same system, and the significant work which proves its success even with highly refined CA rules involving cultural and human factors [Portugali, Benenson, 1997].

For such reasons current models are centering their interest in *orthogonal isotropic shape-constrained meshes* (the regular grid); they are essentially analytic. Nevertheless, most of them function with a certain degree of deviation from classic CA [Zhongwei, 2003], redefining cell space, neighborhood, lattice and time concepts, which is necessary for some degree of flexibility.

Proposed technique

We are going to describe a technique for building meshes as a generalization of these grids, more flexible and adaptable entities where orthogonality, isotropy, and constant topology are specific cases. Fur such purpose, a combination of Kohonen's self-organizing map and a general CA algorithms set lead us to satisfactory results.

1. Motivations for SOM and CA combined approach

We detected interesting properties in both models that acting indepently solve different aspects of the grid (see Table 1). Basic SOM algorithms work reconfiguring their neurons' weight in a competitive basis, resulting in a problem-specific distribution while preserving topology of the map. However, CA algorithms work efficiently calculating relationships among cells, with just an initial configuration and no needed input data. On the other hand, classic CA rely on fixed-topology lattices, being highly suitable for massively parallel calculations, although they are not limited to be rectangular and uniform [Abdalla, Setoodeh, Gürdal, 2006].

Self-Organizing Maps	Cellular Automata
preserves topology of the map	relies on topologically static lattice
works on input data	works from seed and sets of rules
problem-specific adaptation (outer generation)	cell relationships (inner generation)

Table 1. SOM and CA for design.

These features suggest the possibility of dividing the design problem into two conceptually complementary parts:

- *Outer generation*: refers to the ability of the design system to assimilate external information and reconfigure itself depending on the input.
- *Inner generation*: concerning the inner properties of the system, not directly depending on the external constrains of the design problem. Although it can introduce external variables to the system, the only potential effect on the global qualities could be emergent, thus unpredictable.

2. Kohonen's self-organizing map

SOM are Artificial Neural Networks that carry out their adjustment process through the unsupervised learning paradigm, and the input data are called unlabeled data. As a simplified definition, we can say that, in a topology-preserving map, units located physically next to each other will respond to classes of input vectors that are likewise next to each other. Although it is easy to visualize this in two-dimensional array, it is not so easy in a high-dimensional space. N-dimensional input vectors are projected down on the two-dimensional map in a way that maintains the natural order of the input vectors. This dimensional reduction could allow us to visualize easily relationships among the data that otherwise might go unnoticed [Freeman, Skapura, 1991].

A basic SOM is able to reconfigure its weights distribution depending on the training data. Weight vectors correspond to the classes the algorithm is finding in the training set. If we graph the weights vectors together with the input vectors we get a similarities diagram where the neuron units are shown as representatives of a data class from that input. We can describe the learning process by the equation

$$w_i = \langle(t)(x - w_i) U (y_i)$$

Where w_i is the weight vector of the *i* unit and *x* is the input vector. The function $U(y_i)$ is zero unless $y_i > 0$ in which case $U(y_i) = 1$ so only active units will learn. The factor α(t) is a function of time to allow its modification as the learning process progresses.

As a competitive structure a winning unit is determined for each input vector based on the similarity between the input and the weight vectors. The winning unit is evaluated by

$$|| x - w_c || = \min_i \{ || x - w_i || \}$$

Finally the updated weights are those of the winning unit plus a defined neighborhood with certain decay. The cycle is repeated, adjusting the neurons' weight until a number of iterations is reached or a fixed condition is satisfied. The neurons are gradually learning as they win for a specific class of input sample. New sample vectors

will fall into a previous class or excite a neuron that wasn't previously excited, tending to be a representative of a new class. Hence, the algorithm -in its simple form- is often used as a generic classifier.

The output, seen as a *weights map*, is generally used as a representation of the classes. If you graph the input vectors on top of it, you would see the relationships easily.

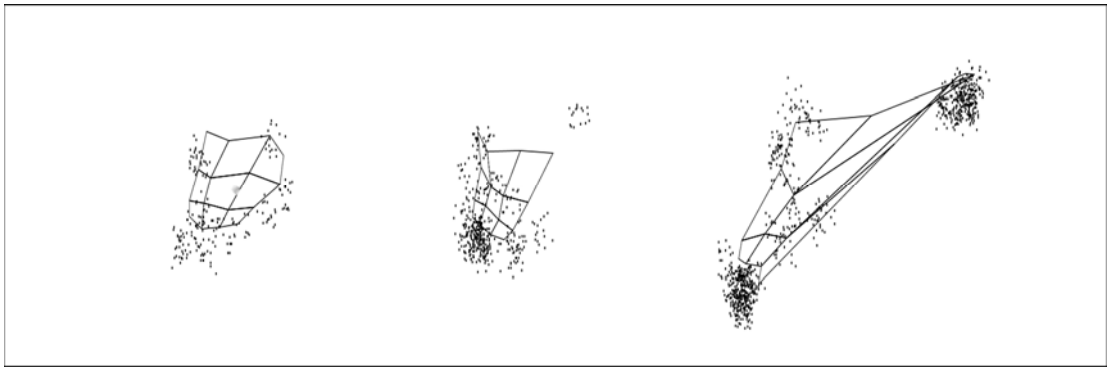


Figure 1. Graphing input vectors (point clouds) and weights (nodes of the mesh).

Nevertheless, our interest is mainly focused on that output graph. The principle we are going to follow is very simple: If we use spatial data as input, we will get spatial data output. Thus *the output graph of spatial information is spatial information itself*; and so the graph turns into a visualized mesh of points.

As information required by the designer is visual, the evaluation of the results could be done entirely with this mesh, until the next step (CA) is applied. Moreover, the mesh is adaptable to the new samples, which means that the designer could learn how to introduce new inputs an re-evaluate the results on the fly.

The SOM was built according to the following features:

- 2-dimensional input vectors
- Variable map resolution
- A Gaussian neighborhood.
- 2 learning phases: At the beginning the learning factor is tuned for exploration of the space and after 100 cycles it optimizes weights. The second phase stops after 10000 iterations or the variation is small enough.
- The weights are graphed in a 2-dimensional map of points, connected by lines representing the immediate neighborhood

It was implemented in Java and ran as an independent application, isolated from the rest of the experiment. Networks of 50x50, 100x100, 150x50 neurons were used, all of them running at a reasonable speed for its manipulation. The option of exporting the neurons' weights as a list of point coordinates was implemented in order to extract the desired configurations out of the program.

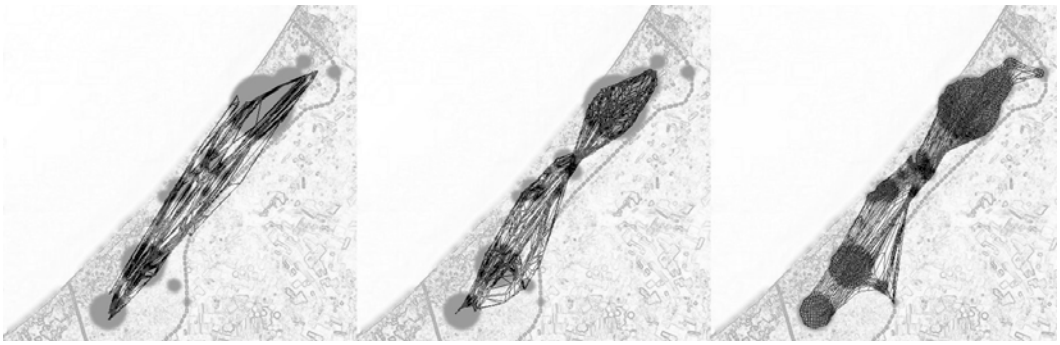


Figure 2. Evolution of the SOM algorithm applied on spatial data, in a vast land region.

3. Cellular Automata: topology reconfiguration

Cellular Automata are a very widespread method for modeling complex systems, such as urban growth. There have already been much work using this models in architecture and urbanism and even specialized ones have been developed [Torrens, 2000]. The discussion of these goes beyond the scope of this paper.

We have seen an algorithm that yields and adaptive mesh, continuously changing over time until we decide to halt it and extract the output. This mesh is an ordered set of points, which interpretation depends on the supplied data.

On the other hand, CA use a lattice of regularly spaced cells with no other geometrical information than the relationship among them. They are built from individual cells that contain all the information needed to update the state. Furthermore, the only external information to the cell comes directly from the adjacent cells, which along with it forms this neighborhood.

The proposal is to run a CA algorithm on top of the previously generated mesh, matching each neuron to a CA cell. The only prerequisite is that the CA dimension and the lattice resolution match exactly to the mesh.

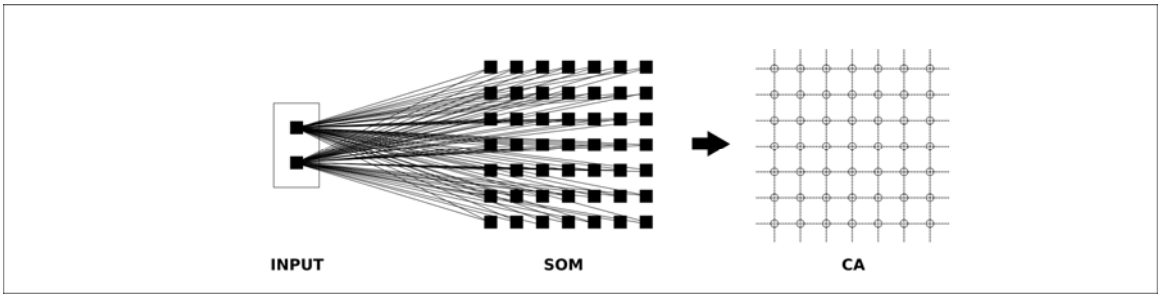


Figure 3. SOM-CA mapping

The CA task is to activate or deactivate a neuron for the next weights adjustment (modifying the neighborhood) or just for the output. By this way we obtain two different possibilities:

- *Filtered flow:* selecting the output units, a linear approach.
- *Feedback flow:* reconfiguring the neighborhood through the suppression of some units, a cyclic approach.

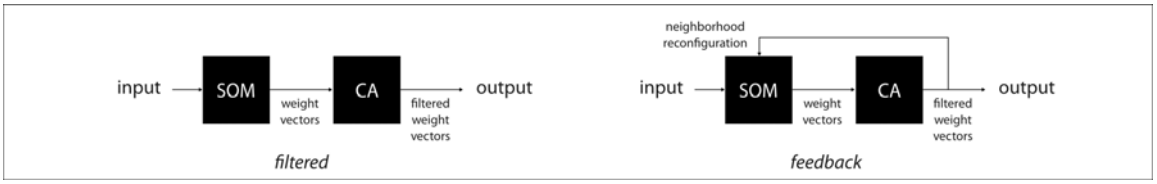


Figure 4. SOM-CA data flow methods.

Both of them result in the *topology reconfiguration* of the mesh, modifying the neurons space. This neurons space acts as the lattice of the CA, which itself has a constant cartesian topology (coming out from the SOM). That is necessary for a regular study of patterns formation [Marr, Hütt, 2005]. These patterns are then used for the reconfiguration of the SOM. It implies that there is always a virtual topology where the CA lives, and an actual topology that is being constantly regenerated by the cellular automata. The use of the CA paradigm for addressing the topology design has recently been demonstrated successful with several approaches [Abdalla, Setoodeh, Gürdal, 2006]. This activate/deactivate behavior could also be extended with other parameters from more complex, even n-dimensional models [Castro, Cabañero, 2007].

The tests were made using the simplest option (filtering), so the separated programs approach was reasonable. A more detailed study should be dedicated exclusively to the feedback configuration.

This second program was also implemented in Java providing the following features:

- Time control: start/stop
- Direct interaction with the CA cells: manually add/erase
- A set of implemented Classic CA, plus traffic-pedestrian simulation and other interesting rules to test
- The ability of changing rules on the fly
- A stack of rule processes, allowing to apply more than one rule in the same run, in a linear and ordered manner

As the previous program, the mesh generator, it is able to export the data at a specific time and extract it from the program.

Merging the two algorithms: results

Both programs were implemented with the ability to export their results, so we can use the frozen data externally. In this last section it is described briefly how this information has been introduced into the designer's workflow, the last step of the technique. The most important thing that has to be performed is to process that information into a 3-dimensional geometry inside a design tool. The chosen platform was the open-source 3d modeling software called Blender, running an embedded Python interpreter.

The python algorithm is straightforward. These steps are to be followed:

1. Read the SOM mesh data points
2. Read the CA cells status
3. Match the SOM neurons with the CA cells, activating or deactivating the neurons
4. Project the SOM in the 3d space, on the XY plane
5. Optional use of an interpretative/generative algorithm to create volumes or additional point properties. It is an interesting step, in which relies the final aspect of the urban plan as it introduces the actual phenotype directly. It is important to notice that although it will introduce great differences in the system, the underlying structure will remain dependent on the previous steps. If this step is avoided, the designer will work directly with the imported 2-dimensional mesh.

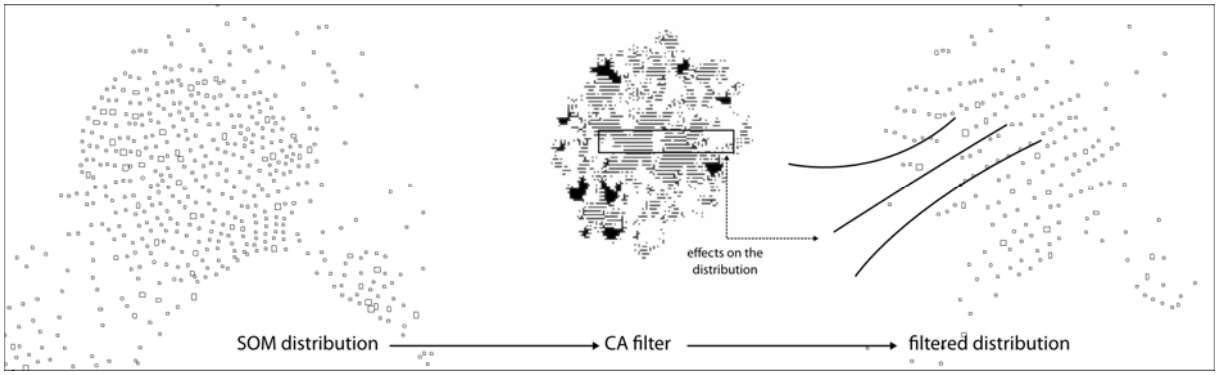


Figure 5. Example application: effects CA-filtered SOM.

Conclusion

The degree of success this technique could have in the design process is a matter of discussion both in the practical and the theoretical domains. However, different applications have already shown that it can be used as an aid tool with more general ambit than the regular grid. Becoming more than an aid tool, capable of synthesizing part of a design is a promising but ambitious objective.

Future work will develop this issues:

- Input data preprocessing
- Feedback flow as a more powerful and complex mechanism
- Multidimensionality of the input data
- Creation of layered or 3-dimensional meshes

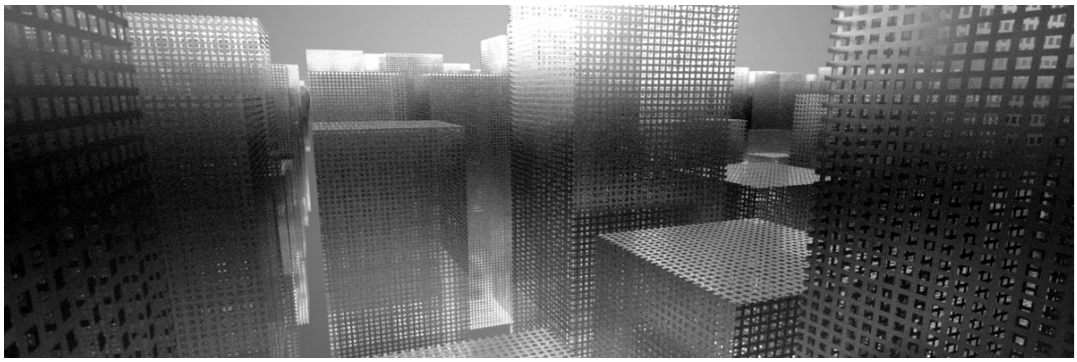


Figure 6. View of a prototype of synthesized city generated with this system.

Bibliography

[Ilachinsky, 2001] Andrew Ilachinsky. Cellular automata - A Discrete Universe. World Scientific Publishing Co. Pte. Ltd., Singapore, 2001.

[Kohonen, 2001] Teuvo Kohonen. Self-Organizing Maps. 3rd Edition. Springer-Verlag, Berlin, 2001.

[Colonna, Di Stefano, Lombardo, Papini, Rabino, 1998] Colonna, Di Stefano, Lombardo, Papini and Rabino. Learning Cellular Automata: Modelling Urban Modelling. In: Proc. of the 3rd International Conference on GeoComputation, University of Bristol, 1998.

[Batty, Xie, Sun, 1999] M. Batty, Y. Xie, Z. Sun. Modeling urban dynamics through GIS-based CA. In: Computers, Environment and Urban Systems. Elsevier, 1999.

[Portugali, Benenson, 1997] J. Portugali, I. Benenson. Individuals' cultural code and residential self-organization in the city space. In: Proceedings of GeoComputation '97 & SIRC '97. University of Otago, New Zealand, 1997.

[Zhongwei, 2003] Sun Zhongwei. Simulating urban growth using cellular automata. Intl. Inst. for Geo-information science and Earth observation. The Netherlands, 2003.

[Abdalla, Setoodeh, Gürdal, 2006] M. M. Abdalla, S. Setoodeh, Z. Gürdal. Cellular automata paradigm for topology optimisation. In: IUTAM Symposium on Topological Design Optimization of Structures, Machines and Materials: Status and Perspectives. Springer, Netherlands, 2006.

[Freeman, Skapura, 1991] J. A. Freeman, D.M. Skapura. Neural Networks. Algorithms, Applications and Programming Techniques. Addison-Wesley Publishing, USA, 1991.

[Torrens, 2000] P. M. Torrens, How Cellular Models of Urban Systems Work. CASA Working paper series, paper 28. CASA, London, 2000.

[Marr, Hütt, 2005] C. Marr, M. Hütt. Topology regulates pattern formation capacity of binary cellular automata on graphs. Physica A 354 (2005). Elsevier, 2005.

[Castro, Cabañero, 2007] Álvaro Castro, Carlos Cabañero. IC_ emergent processes particles projector. In: AMinima, Barcelona, 2007.

Authors' Information

Álvaro Castro Castilla - *Universidad Politécnica de Madrid, Escuela Técnica Superior de Arquitectura, Madrid, Spain; e-mail: alvaro@endosymbionts.com*

Nuria Gómez Blas - *Dept. Organización y Estructura de la Información, Escuela Universitaria de Informática, Universidad Politécnica de Madrid, Crta. De Valencia km. 7, 28031 Madrid, Spain; e-mail: ngomez@dalum.eui.upm.es*

A FUZZY CONTROL APPROACH FOR VOTE ESTIMATION

Jesús Cardeñosa, Pilar Rey

***Abstract**: This document presents the results of applying fuzzy control methods for the estimation of the lack of answer for the vote variable in the opinion polls carried out by the Sociological Research Centre (SRC).*

***Keywords**: Fuzzy systems, voting systems*

***ACM Classification Keywords**: J4 Social and Behavioral Sciences; I.2.3. Deduction and Theorem Proving*

Introduction

This document presents the results of applying fuzzy control methods for the estimation of the lack of answer for the vote variable in the opinion polls carried out by the Sociological Research Centre (SRC). The first section describes the data used, emphasizing some problems which could be solved by means of classification procedures. The second section shows the designed fuzzy control system for classification, describing the procedures to generate the fuzzy rules and sets and their operation. Finally, we present the results and some conclusions and possible ways of improvement.

Type of Data

One of the missions assigned to the SRC by the law regulations is to carry out surveys and opinion polls to know the Spanish social reality; in particular, to learn about the Spaniards' vote intention for the general elections. In the opinion poll periodically carried out the first month of every trimester (known as "Opinion Barometer"), a direct question is posed about this vote intention. In addition, the survey includes other questions about the socio-demographic variables (sex, age, study level, labour situation, profession,…), and about other subjects related to vote behaviour that can contribute to improve knowledge about the reasons a person may have to vote for a specific political party.

In the surveys carried out by public institutions, the opinion polls are especially easy due to the simplification of some of the most technical phases of the survey processing. One of the reasons is the use of variables of qualitative type with a limited number of answering categories. In addition, only proportion estimates are almost exclusively used, and they originate smaller sampling errors than other kind of estimates. Another peculiarity of these surveys is the treatment given to the partial lack of answer (when the interviewed answers one or some of the questions, but not all of them): the category "*Don't know* / *Don't answer"* is added, and it is considered as any other one. By this procedure, the proportion estimated for the rest of the categories are skewed downwards, but the precision level demanded for the opinion surveys does not seem to be really great. In the case of the variable measuring the proportion of vote intention for every party ("*vote*"*,* from now on*),* the high precision is demanded and it is precisely in this question where the partial lack of answer is usually bigger. To deal with this problem, the SRC usually make some *a posteriori* treatment (incorporation of expert's opinion, econometric modelling…). In any case, after elections are held, the media frequently points out critics and comments about the poor results obtained by these procedures regarding vote forecasting. Thus, when testing fuzzy systems, we will use "*vote"* as classification variable, with the purpose of using the classification results as an alternative procedure for the proportion estimates. The lack of answer would be replaced by classified values and the new proportions would be calculated. Since this work is only a first approach to the use of fuzzy controllers for this type of data and we tried to build a simple system with few rules, the answers will be grouped in four categories: "*IU*", "*PP*", "*PSOE*" and "*OTHER*". (Being IU, PP, and PSOE acronyms of the main Spanish Political parties)

As input variables we will use the answer to four questions related to the vote: the assessment of the government management, the assessment of the first opposition party performance, the memory of the party voted for in the

last national elections, and the ideological position (this latter resulting from asking the interviewed about his/her ideological position ranging from 1 to 10: 1 being the extreme left, and 10 the extreme right). The answer categories for the questions about the assessment of government management and the first opposition party performance are "*Very good*", "*Good*", "*Regular*", "*Bad"* and "*Very bad*". For memory of party voted for, they are also grouped in "*PP", "PSOE", "IU*" and "*OTHER*". This may lead to an excessive simplification, because besides grouping very different political parties (as in the case of vote intention variable), it also groups other categories that may show very different behaviours as "*Did not have age*", "*Does not remember*", "*Vote in white*" and "*Did not vote*". We have selected the data from the study number 2640 of the SRC Data Bank Catalogue (April 2006 Barometer), with a sample size of 2.500 interviews. The micro data corresponding to people residing in Autonomous Communities with other great parties (Catalonia, Galicia, and Basque Country) have been eliminated, and also those having partial lack of answer in some of the four input variables or in the output variable, resulting in 1.216 micro data to test the procedure.

Design of the Fuzzy Controller

An attractive point of the fuzzy control systems is the option of using simple rules that do not require special efforts for its design. As we have not experts in vote motivation, the classification rules are built by supposing that the interviewed answered the survey questions with some consistency. On the other hand, we are going to apply fuzzy rules in which each one will describe one of the possible classification categories. The antecedent part of the rules will be expressed by means of defined fuzzy sets in the answer categories sets of the four input variables, whereas the consequent part will be a crisp class label in the set of the classification categories. The general expression of these rules is:

R_i : If x_1 is A_{i1} and x_2 is A_{i2} and x_3 is A_{i3} and x_4 is A_{i4} then $y = y_i$, $i=1, 2, 3, 4$

where: x_1 = *assessment of government management*

x_2 = *assessment of first opposition party performance*

x_3 = *ideological position*

x_4 = *memory of the party voted for*

y_1= *IU*

y_2 = *OTHER*

y_3 = *PP*

y_4 = *PSOE*

In a detailed form, the rules are:

– R_1: If the assessment of government management is **negative**, the assessment of first opposition party is **negative**, the ideological position is **low** and the memory of the party voted for is "***IU***", then the vote is "***IU***".

– R_2: If the assessment of government management is **negative**, the assessment of first opposition party is **negative**, the ideological position is **average** and the memory of the party voted for is "***OTHER"***, then the vote is "***OTHER***".

– R_3: If the assessment of government management is **negative**, the assessment of first opposition party is **positive**, the ideological position is **high** and the memory of the party voted for is "***PP***", then the vote is "***PP***".

– R_4: If the assessment of government management is **positive**, the assessment of first opposition party is **negative**, the ideological position is **low average** and the memory of the party voted for is "***PSOE***", then the vote is "***PSOE***".

It is important to point out that, although we have defined for "*memory of the party voted for"* the same answer categories than those for the "*vote"*, in the rules the meaning is very different, since for the "*memory of the party*

voted for" variable we will define a fuzzy set for each party, whereas for the "*vote"* variable it regards to crisp class labels. After that, we move on to build the A_{jj} fuzzy sets of the antecedent part of the rules. In the first place, we built the "*Positive Assessment*" (*PA*) and "*Negative Assessment*" (*NA*) sets for the variables x_1 and x_2, which will be identical for both. As of the answer categories, their membership functions will respectively be in a natural way:

$$\mu_{PA}(x)=\begin{cases}0.00 & if\ x=very\quad bad\\ 0.25 & if\ x=bad\\ 0.50 & if\ x=regular\\ 0.75 & if\ x=good\\ 1.00 & if\ x=very\quad good\end{cases}\qquad and\quad \mu_{NA}(x)=1-\mu_{PA}(x)$$

where a set is the complementary of the other, taking for the complementary the strong standard negation (Figure 1).

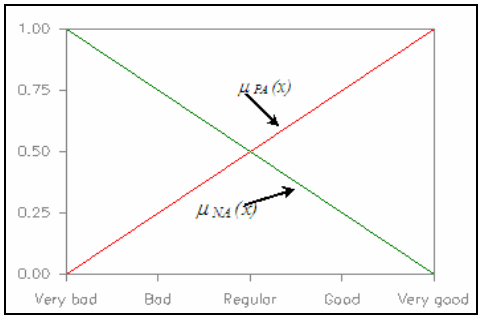


Figure 1

Now, in order to build the "*Low*" (*L*), "*Low Average*" (*LA*), "*Average*" (*A*) and "*High*" (*H*) fuzzy sets for the universal set of answers to the question about the ideological position, we will base on the average and the standard deviation for each classification category obtained from the sample. For later evaluation of the behaviour of the fuzzy classifier, we took into consideration the first 916 observations to estimate, being left the other 300 for the classification test. Table 1 shows the obtained values:

Party	Average	Standard deviation
IU	*3.00*	*1.06*
OTHER	*4.53*	*1.40*
PP	*6.67*	*1.31*
PSOE	*3.71*	*1.25*

Table 1

From this one, we built the Gaussian membership functions, in the form:

$$\mu_L(x)=e^{-\frac{(x-3{,}00)^2}{2.1{,}06^2}},\ x=1,2,\ldots,10\quad ,,\ \mu_{LA}(x)=e^{-\frac{(x-3{,}71)^2}{2.1{,}25^2}},\ x=1,2,\ldots,10$$

$$\mu_A(x)=e^{-\frac{(x-4{,}53)^2}{2.1{,}40^2}},\ x=1,2,\ldots,10\quad ,,\ \mu_H(x)=e^{-\frac{(x-6{,}67)^2}{2.1{,}31^2}},\ x=1,2,\ldots,10$$

that appears in Figure 2. They are taken as symmetrical functions around the average because that is how the studied frequencies in the micro data sample seem to behave.

Finally, we built the fuzzy sets for the results of the "*memory of the party voted for*" variable also from the information provided by the sample of the first 916 micro data, as the vote frequency (distribution by categories of the classification variable) for each group of "memory of the party voted for", as it appears in Table 2:

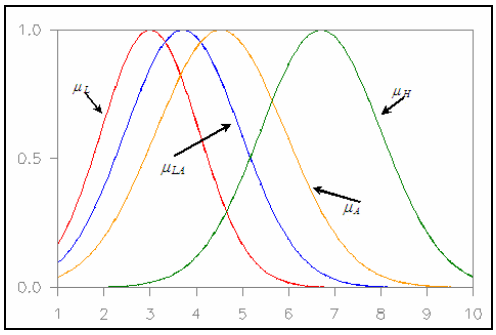


Figure 2

	Vote				
Memory of the party voted for	*IU*	*OTHER*	*PP*	*PSOE*	*Total*
IU	66.1	6.8	3.4	23.7	100.0
OTHER	2.9	37.5	23.0	36.6	100.0
PP	0.7	7.2	75.9	16.2	100.0
PSOE	1.0	7.0	6.1	85.9	100.0

Table 2

This gives rise to the membership functions in Figures 3 to 6:

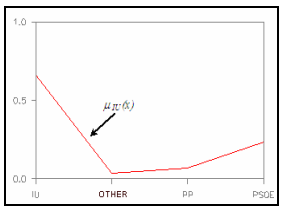


Figure 3

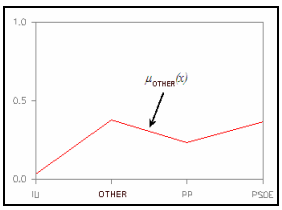


Figure 4

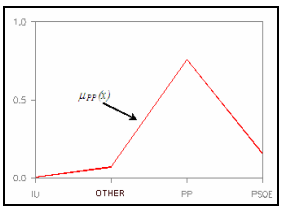


Figure 5

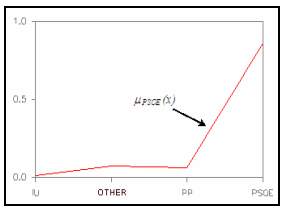


Figure 6

Once the fuzzy sets for the input variables were built, the next action was to follow the steps to design the system of fuzzy control.

Operation of the System

It will be necessary to choose the implication functions, t-norms, and so on, that allow the development of the system logout, once an input takes place. It has already been pointed out that the standard negation is used for the complementary. As implication function we chose the Mamdani implication, based on the t-norm of the minimum. For the t-norms (connective y) we used the product, and as t-conorm the standard fuzzy union (maximum operator). It has also been studied the use of the t-norm of the minimum instead of the product one: the first one showed the problem that frequently the maximum value to select the class was the same for different parties. The use of the t-norm of the product allows avoiding it, making possible the interaction among the antecedent proposals of each rule. On the other hand, we applied the Zadeh's compositional rule of inference locally to each fuzzy relation generated by the rules, coming later to combine the resulting sets in a disjunctive way.

If we suppose now that a numerical input to the system takes place, that is to say, the fact $x = x_0$ takes place or $x \in P^* = \{x_0\}$, where $x = [x_1, x_2, x_3, x_4]^T$ is the vector of input of the four variables, it will be $\mu_{P^*}(x) = \begin{cases} 1, & si\ x = x_0 \\ 0, & si\ x \neq x_0 \end{cases}$

We will then have, to make the inference, the four rules and the previous fact. In the first place, we will combine the fuzzy sets of the inputs of each rule applied to $P^* = \{x_0\}$ in a multiplicative way according to the selected t-norm, getting the activation rank of each rule as:

$$\beta_i(x_0) = \prod_{j=1}^{4} \mu_{ij}(x_{j0}), i = 1,2,3,4$$

where $x_0 = [x_{10}, x_{20}, x_{30}, x_{40}]^T$ and μ_{ij} are the corresponding membership functions. As the output of each rule is a crisp set (the class label of the corresponding party), it will have a membership function that will be, in fact, a characteristic function, properly:

$$\mu_{y_i}(y) = \begin{cases} 1, & si\ y = y_i \\ 0, & si\ y \neq y_i \end{cases}, \quad i = 1,2,3,4$$

In order to apply Zadeh's compositional rule to each rule, it would be necessary to make:

$$\mu_{Q_i^*}(y) = \mathcal{J}(\beta_i(x_o), \mu_{y_i}(y)) = \min[\beta_i(x_o), \mu_{y_i}(y)] = \begin{cases} \beta_i(x_o), & si \quad y = y_i \\ 0, & si \quad y \neq y_i \end{cases}, \quad i = 1,2,3,4$$

where J is the Mamdani implication. Applying now the disjunctive combination of each rule outputs, the output of the classifier will be determined by the rule with the highest activation degree, that is to say,

$$y = y_{i^*}, i^* = \arg \max_{1 \le i \le 4} \beta_i(x_0)$$

Results

In order to test the classifier, we applied it to obtain the value of the *vote* for the 300 observations that have been left in the sample with this aim. If we compared the vote thus obtained with the real value provided by the interviewed person, we found that there was coincidence in 240 observations (80% of the data). In order to make the experiment more representative, we repeated all the steps four times, leaving 300 different observations every time for verification. Although the test number is not enough to extract significant consequences, it is observed that all the parameters taken into account stayed quite stable from a repetition to another one, which provides certain confidence in the reliability of the procedure. The results of the four iterations are shown in Table 3.

Iteration	***% of coincidence***
1	80.00
2	80.76
3	80.76
4	80.70
Average	80.56

Table 3

On the other hand, tests have also been made changing the membership functions obtained for other simpler ones, of trapezoidal and triangular type. It has been found that it is very simple to refine the parameters of those functions to significantly improve the results, but those refined parameters do not continue giving good results in other segments of sample. Although results may seem a little unexciting, if we compared them with 51%, at the most, that has been obtained with data of a similar survey, using classifiers based on Bayesian networks, it is rather positive. (These results are not totally comparable since the other survey used different input variables).

Conclusions

The experiment made here is only a first approach to obtain classifiers for the vote in opinion polls based on fuzzy controllers; therefore, instead of conclusions we are making some comments.

The final mission has been to improve the imputations of the "vote" lack of answer in the barometers of the SRC. It is necessary to point out that this experiment is rather more modest than to improve the estimations of the vote for the following elections. Also, it is necessary to remember that the interviewed people can be incongruous when they are asked about the vote and when they go to vote.

The shown results are quite encouraging, mainly considering that the design of the control system has not required of exhaustive previous analysis, but it has been obtained with a few rules based on common sense.

The system is simple and with very few rules to group the parties in only four groups, but it would necessarily become more complicated if it were tried out with all the political parties.

In the barometers there also are other questions included that can be used like input variables in the system: for example, questions about the assessment of the main leaders of the political parties, the confidence in the President of the government, etc. It is quite possible that its inclusion could allow improving results.

The tests made with other simpler functions of property indicate that it is also possible to improve the results using that route.

Bibliography

[Chen 2006] Chia-Chong Chen. Design of PSO-based Fuzzy Classification Systems. In: Tamkang Journal of Science and Engineering, Vol.9 nº1 pp 63-70. http://www2.tku.edu.tw/~tkjse/9-1/9-1-7.pdf

[Diez 2005] F.J. Diez. Introducción al razonamiento aproximado. Dpto. Inteligencia Artificial, UNED.

[Halkidi 2003] Vazirgiannis, Michalis, Halkidi, Maria, Gunopulos, Dimitrious. Uncertainty Handling and Quality Assessment in Data Mining. 2003, IX, 226 p., ISBN: 978-1-85233-655-4

[Yuan 1994] GJ Klir, B Yuan Fuzzy sets and fuzzy logic: theory and applications. Prentice-Hall, Inc. Upper Saddle River, NJ, USA1994

[Roubos 2002] Johannes A. Roubos, Magne Setnes, and Janos Abonyi. Learning fuzzy classification rules from labeled data. IN: Information Sciences. Vol 150. pp77-93. 2003

[Tanaka 1997] K. Tanaka An introduction to Fuzzy for Logic Practical Applications. Springer New York. 1997

Author's information

Jesús Cardeñosa – *Group of Validation and Industrial Applications. Facultad de Informática. Universidad Politécnica de Madrid; Madrid 28660, Spain; e-mail: carde@opera.dia.fi.upm.es. http://www.vai.dia.fi.upm.es*

Pilar Rey *– Banco de Datos. Centro de Investigaciones Sociológicas. C/ Montalban ,8; 28014 Madrid (Spain). e-mail: prey@cis.es; http://www.cis.es*

PROBABILISTIC AND MULTIVARIATE ASPECTS OF CONSTRUCTION OF THE MODELS AND PROCEDURES FOR PREDICTION OF THE AVALANCHE-DANGEROUS SITUATIONS INITIATION

Alexander Kuzemin, Vyacheslav Lyashenko

Abstract: *The interpretation model for analysis of the avalanche-dangerous situations initiation based on the definition of probabilities of correspondence of studied parameters to the probabilistic distributions of avalanche-dangerous and non-avalanche-dangerous situations is offered. The possibility to apply such a model to the real data is considered. The main approaches to the use of multiple representations for the avalanche-dangerous situations initiation analysis are generalized.*

Key words: *avalanches, probability, set, situation, model.*

Introduction

Avalanche-dangerous regions occupy 6% of the land area. But despite this the problem of such phenomena investigations is rather important and urgent as analogous phenomena can become the cause of people's death and considerable destructions [1, 2]. Among the most essential and important problems in the given aspect one should note substantiation of the utility to use the corresponding mathematical apparatus intended both for investigation of the avalanche-dangerous situations development dynamics and for development of methods for estimation of the potential avalanche cells, prediction of avalanches volumes and descent frequency. This concerns the fact that every avalanche can be regarded as a unique phenomenon of nature with its specific peculiarities. At the same time despite its uniqueness it is possible to single out the climatic conditions variations characteristic ranges which are prerequisites to prediction of the feasible avalanche descent. Eventually, the totality of these two factors defines the presence those approaches to prediction and warning of avalanches descent, at present these approaches are used in geoinformation systems (GIS) which make it possible to accumulate continuously meteorological information, carry out various calculations, reveal regularities and realize spatial tie of the obtained results [3, 4].

Available methods of the avalanche-dangerous situations initiation prediction

There is no doubt that the foundation of avalanche-dangerous situations initiation prediction consists in the procedure of the preliminary analysis of such events. As a rule the solution of the formulated problem is based on the statistical analysis methods. In particular, the approaches of such analysis make it possible to substantiate the most significant system of the facts which is expedient to use in the avalanche-dangerous situations prediction procedures. Such approaches found their development in the predictions of snow avalanches descent based on application of the images similarity method (or the nearest neighbors method) [5] or through the application of the regression equations [6]. The data of nomograms which in a general case extend the interconnection of such indices as temperature, value of snow cover and precipitations are also used for estimation of the avalanches descent probability. But in spite of this the remaining non-predictive nature of the avalanche-dangerous situation doesn't always allow to prevent negative consequences of emergencies caused by their descent. This is associated with that the available procedures of the avalanche-dangerous situation initiation prediction are not sufficiently precise. At the same time the severity of the problem and variety of ways to solve it motivate the necessity to search alternative methods which can give more argued answers.

Thus, as the main aim of the given investigation one can single out consideration of the approaches, alternative to the available methods, to the avalanche-dangerous situations initiation prediction. First of all, in this case the

consideration of the conceptual foundations of the models and procedures of such a prediction is significant in our opinion.

The probabilistic aspects of the avalanche climate initiation medium analysis

Analysis of different characteristics of the avalanche climate initiation medium makes generally the foundation of the avalanches descent prediction. Among such characteristics the most abundant ones are: the air temperature, humidity, atmospheric precipitations volume, wind velocity, angle of the slope of surface (descent angle) where the avalanche descent is possible. In general, the variation dynamics of both individual of the above characteristics the avalanche climate initiation and their totality can, with some probability, characterize initiation either avalanche-dangerous or avalanche-non-dangerous event. As this takes place, a feasible range of the studied avalanche climate initiation characteristics variations describes a definite region of avalanche-dangerous and avalanche-non-dangerous situation. In the conceptual plan the essence of the probabilistic aspect of the avalanche climate initiation analysis can be reduced to the definition of the probability to assign some point as the considered medium current characteristics either to the region of the avalanche initiation or to the initiation of avalanche non-dangerous situation. Otherwise the given approach can be treated as a correspondence of the current characteristics of the avalanche climate initiation medium; parameters of these characteristics define some region using probabilistic distribution of avalanche dangerous or avalanche non-dangerous situations preceding this. Consequently, it is possible to speak about so-called probable conformity of the researched characteristics of the avalanche-dangerous climate environment to probabilistic distribution of the avalanche-dangerous or avalanche non-dangerous situations.

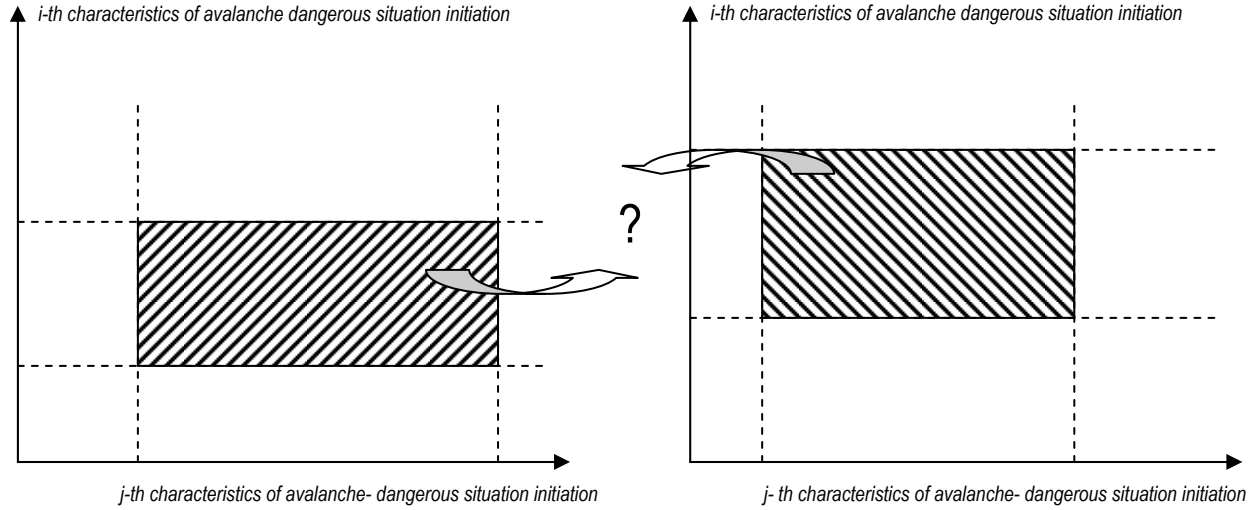


Fig.1 On the explanation of the probabilistic aspect of the avalanche climate initiation medium analysis

In particular, procedure of such analysis can be considered proceeding from the pairwise analysis of various characteristics of the avalanche climate initiation. The advisability of such a transition is related to the fact that at the stage of the preliminary analysis it is possible to omit less significant factors of impact on the avalanche-dangerous situation initiation. Thus, the base element of the analysis procedure being considered is estimation of the probability of the avalanche climate initiation current parameters to fall within the regions typical and atypical for the avalanche climate initiation. The given regions can be presented in the plane in the form of the rectangle;

its metric values correspond to definite parameters of variation of the avalanche- dangerous and avalanche-non-dangerous situations initiation medium characteristics (Fig.1).

To explain the offered aspect and substantiate its significance let us analyze the probabilistic aspects of some characteristics of the avalanche climate initiation medium using the real data of the avalanches descent in ITAGAR Chychkan region Kyrgyzstan Republic within 2001–2006, obtained in the frameworks of carrying out joint scientific and research work. The essence of such analysis is reduced to the estimation of the feasible assignment of the avalanche-non-dangerous situations to the avalanche-dangerous ones and vice versa in terms of different characteristics of their initiation.

First and foremost, it should be noted that the considered characteristics of the avalanche climate initiation medium follow the normal distribution law. This makes it possible to use this law for estimation of the corresponding probabilities. The corresponding probabilities with regard to the avalanche-dangerous and avalanche-non-dangerous situations are presented in Table 1.

Table 1.

Probabilities of correspondence of the current parameters of avalanche climate initiation medium to the avalanche-dangerous and avalanche-non-dangerous situations

Characteristics of the avalanche climate initiation medium analysis	Feasibility of correspondence
under condition of considering the avalanche dangerous situation and avalanche dangerous current parameters	
air temperature – wind velocity	0,854
air temperature – wind velocity	0,823
wind velocity – precipitations quantity	0,707
precipitations quantity – descent angle	0,809
under condition of considering the avalanche-dangerous situation and avalanche- non-dangerous current parameters	
air temperature – wind velocity	0,488
humidity – precipitations quantity	0,582
wind velocity – precipitations quantity	0,317
precipitations quantity – descent angle	0,341
under condition of considering the avalanche-non-dangerous situation and avalanche- non-dangerous current parameters	
air temperature – wind velocity	0,798
humidity – wind velocity	0,878
wind velocity – precipitations quantity	0,866
precipitations quantity – descent angle	0,939
under condition of considering the avalanche-non-dangerous situation and avalanche- dangerous current parameters	
air temperature – wind velocity	0,555
humidity – wind velocity	0,482
wind velocity – precipitations quantity	0,403
precipitations quantity – descent angle	0,591

As can be seen from the data in Table 1 the assumptions made above are reasonably justified i.e. the probability of correspondence of the like situations and parameters is essentially significant. This allows making

generalization even for estimation of probable initiation of the avalanche-dangerous as a whole. To do this one should consider

- either generalization of the obtained probabilities reasoning from the significance of different groups of characteristics of the avalanche climate initiation characteristics analysis in the assumption that the probabilities of correspondence can be considered as conditional probabilities of the concrete situations analysis;
- or a separate group of characteristics of the avalanche climate initiation medium analysis based on the greatest/least values of the correspondence probabilities.

Multivariate aspects of the avalanche climate initiation medium analysis

In any case the obtained above correspondence probabilities for some interval of time make it possible to pass to the examination and analysis of the multitude of the avalanche dangerous initiation. As an alternative to the boundary of the distribution polygons built using the correspondence probabilities in some temporal interval can be considered. The analysis of such sets allows formalizing the totality of the avalanche-dangerous and avalanche-non-dangerous situations and creating the procedure of the corresponding prediction.

Moreover, the fuzzy sets theory methods and approaches can be used as one of the specified problem solution directions. In particular, such an approach can be used for prediction of the avalanche descent time. A fuzzy-set

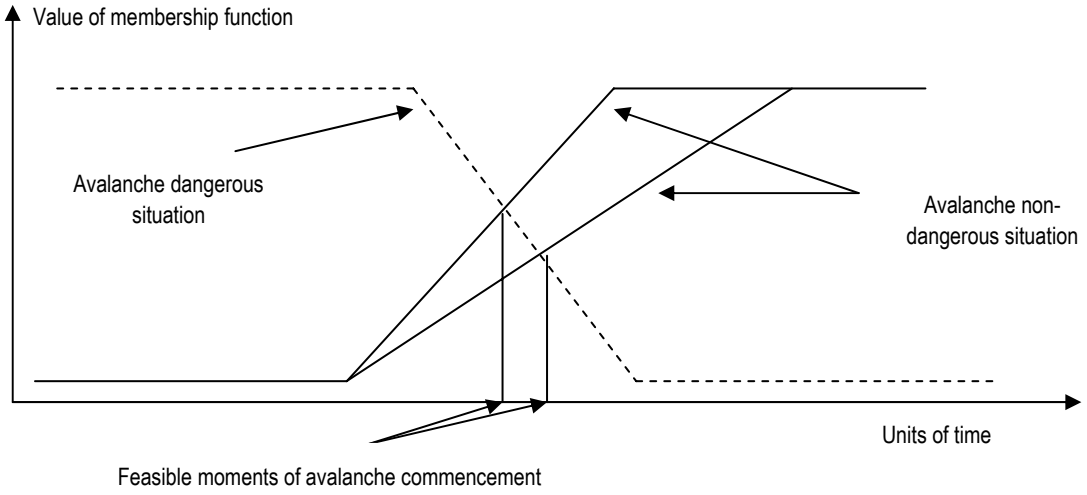


Fig.2. Fuzzy model in analysis of avalanche descent commencement

description of the avalanche-dangerous and avalanche-non-dangerous situations makes it possible to perform.

The essence of such description reduces to creation of the fuzzy model of estimates of the avalanche dangerous situations temporal characteristics. Thus, for example, it is possible to suppose that in the case of the avalanche non-dangerous situations the time till the tentative avalanche descent will be the greater the greater is the likelihood of assignment of the current parameters of the analysis of the avalanche climate initiation to such a situation. Respectively, in the case of minor probabilities of the avalanche climate initiation analysis current parameters assignment to such situation can points to insignificant reserve of time till the avalanche descent commencement moment. When considering the avalanche dangerous situations the corresponding characteristics are the converse ones. This makes it possible to introduce the avalanche descent commencement

into a definite fuzzy set functions and their generalization will just allow predicting the avalanche descent time (Fig.2).

It is possible to pass on to the distribution functions when estimating the avalanches descent time based on the analysis of the corresponding probabilities to subdivide available data into the avalanche-dangerous and avalanche-non-dangerous ones.

Thus, the model construction generalized scheme and construction of the procedure for prediction of the avalanche-dangerous situations initiation is reduced to:

- sequential obtaining of the probabilistic characteristics of the avalanche climate initiation medium;
- construction of the corresponding sets of subdivision into avalanche-dangerous and avalanche-non-dangerous situations;
- analysis of the avalanche descent initiation time using fuzzy models of its interpretation.

Conclusion

This work presents the general concept of building models and procedures for prediction of the avalanche-dangerous situations based on the probabilistic and multiple approaches to its interpretation. Such an approach allows, first and foremost, to take into account the range of variations of the factors acting on the avalanche-dangerous situations initiation, to build adequate procedures for their prediction. Moreover, the essential characteristics of the problem under consideration which is confirmed, in particular, by the feasibility of the probabilistic model of the general approach based on the real data.

Bibliography

1. Federici P., Tamburin A.i, Luzi G., Rott H, Schaffhauser A., Strozzi T., Bernardini G. Galahad: An EU Project for the Remote Monitoring of Glaciers, Avalanches and Landslides // IDRC. – Davos, 2006. – Vol. 2. – P. 177–180.
2. Voitkovsky K.F. Science of Avalanches. – M.: MSU, 1989. – 158 p.
3. Durand Y., Brun E., Merindol L., Guyomarc'h, Lesaffre B., Martin E. A Meteorological Estimation of Relevant Parameters for Snow Models. Ann. Glaciol.,18, 1993. – P. 65–71.
4. Kuzemin A., Toroev A. Mobile Means of Control and Prediction of Avalanche Climate Using Information Conversion in Acoustic Range. 291 // IDRC. – DAVOS, 2006. – Vol. 2. – P. 291–294.
5. Buser, O., Butler, M. and Good, W. 1987. Avalanche forecast by the nearest neighbors method // IAHS. – Publ. 162. – P. 557–569.
6. Izhboldina V.A. Aerosinoptical conditions of formation and descent of snow-storm avalanches in the Kola Peninsula// Collection of articles: Studies of snow and avalanches in the Khibini Mountains. – L.: Hidrometeoizdat, 1975. – P.51–63.

Authors' Information

Kuzemin A.Ya. *– Prof. of Information Department, Kharkov National University of Radio Electronics, Head of IMD, (Ukraine), kuzy@kture.kharkov.ua*

Liashenko V.V. *– senior scientific employee, Kharkov National University of Radio Electron (Ukraine), kuzy@kture.kharkov.ua*

INFORMATION SUPPLY OF GEO-INFORMATION SYSTEMS FOR THE FORECASTING PROBLEM OF THE AVALANCHE DANGER

Alexander Kuzemin, Olesya Dyachenko, Darya Fastova

Abstract: *This article is dedicated to the vital problem of the creation of GIS-systems for the monitoring, prognostication and control of technogenic natural catastrophes. The decrease of risks, the protection of economic objects, averting the human victims, caused by the dynamism of avalanche centers, depends on the effectiveness of the prognostication procedures of avalanche danger used. In the article the structure of a prognostication subsystem information input is developed and the technology for the complex forecast of avalanche-prone situations is proposed.*

Keywords*: GIS, prognostication, risk, situation, the avalanche danger*

Introduction

A study of the natural calamities mechanisms, the development of their connections with the climatic and ecological changes led to the development of the new specialized systems technology for control, which was called **geo-information systems (GIS)**. The basic tasks of GIS-systems are the development of the prognostication methodology for technogenic catastrophes, the estimation of risks and creation of decision making support systems. GIS-systems are intended for working and analysis of the enormous massifs of data for the definition of the characteristics of the zones of high risk, improvement in the planning, which precedes calamities and estimation of damage [1, 2]. The methods and techniques of GIS-systems make possible to evaluate strategies of reduction in the probability of the catastrophes occurrence, including the calculation of social and economic nature. This includes monitoring physical, biological and chemical parameters on the spot of calamity, control of measurement data and development of short- and extended forecast models. The developed GIS-systems make possible to continuously accumulate meteorological information, to perform different calculations, to reveal regularities, to achieve a three-dimensional tying of results. The purpose of this article consists in effectiveness increase in solution taken by GIS-system due to the development of the complex forecast technology of avalanche danger. According to the stated goal, it is necessary to solve the following subtasks:

1. to determine structure, tasks and purposes of the developed GIS-system;
2. to develop structure and method of prognostication subsystem operations;
3. to determine information supply of prognostication subsystem;
4. to build the technology of the avalanche danger complex forecast.

Structure and the task of GIS-system

GIS-system is the totality of the following elements (Fig. 1):

1. *monitoring subsystem*, intended for guaranteeing of dynamic monitoring and mapping the indices of dangerous situations on that investigated of territories in the visual (tabular, graphic, cartographic and animated) form.
2. *integration subsystem* of the different data sources for the solution of the problems of control of natural catastrophic situations.
3. *analytical subsystem*, which ensures the complex analytical processing of information for the solution of complex analytical problems.

4. *prognostication subsystem*, which ensures the multivariant scenic and purposeful prognostication of situations on the base of the complex of the interconnected models of the separate parameters.
5. *system of decision making support*, based on the development of models and methods for making of adequate decisions of operational and strategic nature.
6. *subsystem of results representation*, which ensures data presentation in the most visual tabular, graphic and cartographic form, which reflects qualitative characteristics and basic tendencies of the indices of situation.

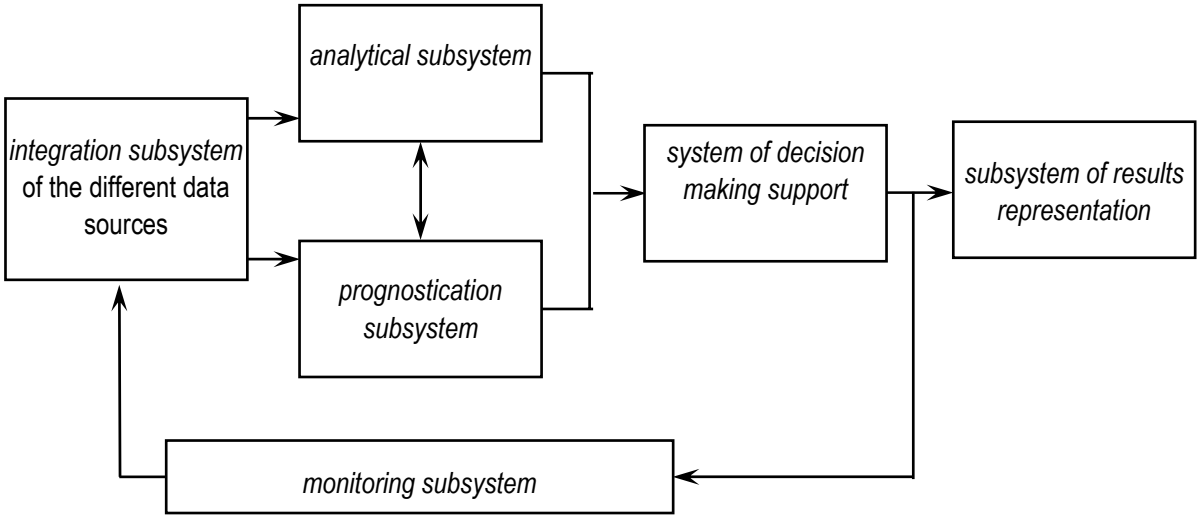


Fig. 1. GIS-system structure.

The distinctive special feature of geo-information systems is the feedback, whose function fulfills the subsystem of monitoring. In general form the role of GIS-technologies in avalanche studies is reduced to the synthesis of knowledge about the relief, the climate and the preceding events, for the purpose of possibility determination of gathering snow avalanches. For this in the GIS environment are imported already existing maps or new projects are created. The analysis of the works, dedicated to use GIS in avalanche studies, showed that the GIS-technologies at present adapt for the solution of the following problems:

- development of the zones of the origin of avalanches;
- simulation of processes and phenomena, which determine the conditions for gathering snow avalanches;
- definition of lethal areas;
- creation of the cadastral surveys of avalanche centers, data bases about the avalanches;
- forecast of avalanche danger.

Subsystem of prognostication

The technology of the prognostication of avalanche danger is the information complex, which consists of three basic blocks (Fig. 2):

1. database - is intended for collection, storage and initial processing of the data of hydrographic and weather services, snow-avalanche stations and electronic charts of surface of locality, which contain information about snow accumulations, to the underlying surface and so forth
2. Mathematical and algorithmic guarantee - is the collection of mathematical methods and approaches, on base of which is produced the simulation and the prognostication of avalanche-prone danger. The prognostication of avalanche-prone situation is characterized by four basic parameters: by place, by

type, by time and with its degree of power. Each of the characteristics has available their mathematical, algorithmic and program apparatuses.

3. block of results assignment for different levels of users - the obtained forecasts are analyzed

by experts and leaders of **Emergency And Disaster Relief Ministry**, after which they are transferred for modification to the system of decision making support for the purpose of use with the correction of anti-avalanche measures and to the elimination of the consequences of gathering avalanche.

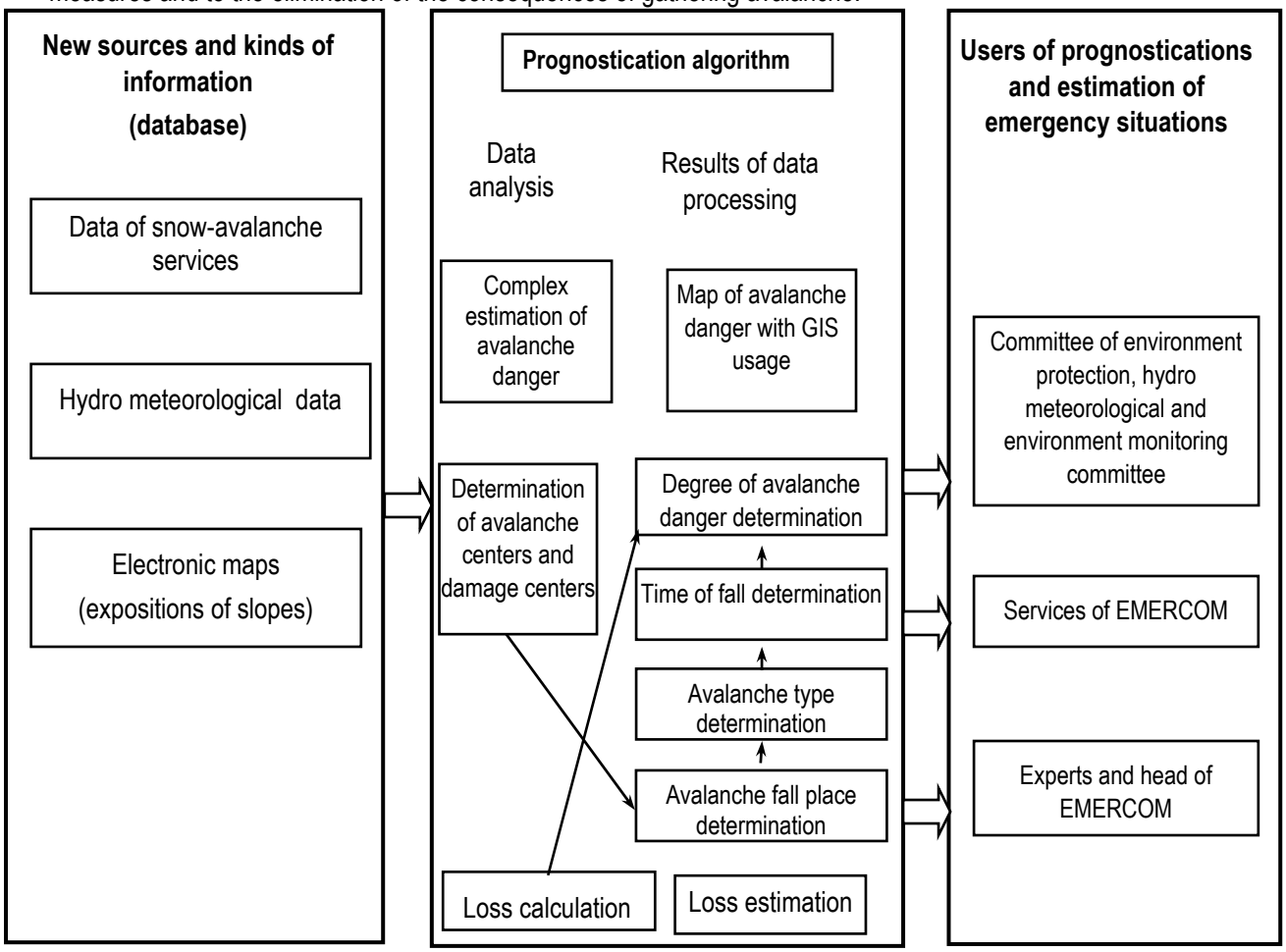


Fig. 2. Technology of avalanche danger prognostication.

For the realization of the complex forecast of avalanche-prone situations it is necessary to determine the place of gathering avalanche, to produce its classification on the genesis, to determine time and degree of the power of avalanche. The place of gathering avalanches is determined with the aid of the electronic charts of area relief, on which is determined the slope of slopes, they calculate the thickness of snow cover and the distance of ejection. The genetic type of avalanche depends the geographical and climatic special features of mountain locality. 5 basic genetic types of avalanches can be specified:

1. avalanche of the freshly fallen snow.
2. wet avalanches.
3. avalanche of snowstorm snow.
4. avalanche of temperature reduction.
5. avalanche of sublimation diaphthoresis.

This classification of avalanches is conditional, since it is very difficult to find the relation of the processes of the appearance of avalanches with the natural conditions of avalanche-prone regions. Depending on the genesis of avalanches the procedures of remaining characteristics prognostication of avalanche danger are selected. In the world practice a large quantity of power estimation scales of avalanches is used - these are the European scale, the American scale, the French scale of avalanches power evaluation [4]. In the developed subsystem of the prognostication of avalanche danger adapts the French scale of the evaluation of the power of avalanche danger, developed By F. Rapin [5] (table 1). The hazard level is evaluated by five progressively growing steps, which are described through such physical parameters as the damaged territory, the thickness of avalanche board, the volume of derailed snow, dynamic pressure, probability of caving avalanches and their effect on the vital activity in the mountains.

Table 1. Scale of the avalanches power degrees

	Degree of avalanche power	Physical parameters	Probability of avalanche caving
1	Insignificant	Damaged territory:~0.2 GA Thickness of the avalanche board:: 20 cm Amount of derailed snow: $\sim 100\,\mathrm{m}^3$ Dynamic pressure: ~2 KPa	Caving is possible only with the very significant increment loads on the separate very steep slopes. Spontaneously can occur only the motions of snow
2	Moderate	Damaged territory:~1 GA Thickness of the avalanche board:: 40 cm Amount of derailed snow: $\sim 1000\,\mathrm{m}^3$ Dynamic pressure: ~10 KPa	Caving is possible with the significant increment loads first of all on the slopes indicated, spontaneous caving of avalanches highly improbably
3	Significant	Damaged territory:~5 GA Thickness of the avalanche board:: 80 cm Amount of derailed snow: $\sim 10000\,\mathrm{m}^3$ Dynamic pressure: ~50 KPa	Caving is possible with the insignificant increment load on the slopes indicated. Is possible caving separate average and less probably large avalanches
4	Large	Damaged territory:~20 GA Thickness of the avalanche board:: 150 cm Amount of derailed snow: $\sim 80000\,\mathrm{m}^3$ Dynamic pressure: ~200 KPa	Caving is possible on the majority of slopes with insignificant increment load
5	Very large	Damaged territory:~50 GA Thickness of the avalanche board:: 250 cm Amount of derailed snow: $\geq 400000\,\mathrm{m}^3$ Dynamic pressure: ~500 KPa	A numerous spontaneous avalanche caving on any slopes are expected

A time aspect of avalanche danger forecast provides for determination of the possibility of gathering avalanches in the assigned territory into the caused time interval. Among the difficulties, connected with the calculation of the time aspect of the avalanche activity, it is possible to state [6]:

1. Provided in the scientific literature classification of forecasts on short -, middle- and long-term does not use the fixed time intervals for their separation. The analysis of works on the prognostication of avalanche danger shows that in practice the forecast can be comprised for day, 48 hours, 72 hours, for the winter season, for the long-standing interval of time.
2. Forecasts of avalanche danger are created with the use of those of specially developed for the region or the separate center procedures, which determine the algorithm of avalanche danger detection.

3. A lot of procedures provides for the forecast of avalanche-prone period - the time interval, for which it will remain the action of avalanche formation factor. Usually, this approach is used with the forecast of avalanches during the snowfalls and snow-storms. Avalanches are forecasted from the moment of achieving the critical conditions to the end of the snowfall (snow-storm), and for the period from one to two days from their end - thus far the instability of snow cover remains.

lead time (time between forecast composition and beginning of its action) of forecast, placed in many procedures of forecast is equal to zero. In practice this indicates the statement of facts of reaching the avalanches of conditions critical for the gathering. The basic reasons for this lie in the transience of avalanche-prone situation appearance (from several hours to days), a constant change in the meteorological conditions, impossibility of the continuous and general collection of necessary information. The complex forecast of avalanche danger is the necessary information, which is The basic tasks of the system of support and decision making are:

1. guarantee of the planning, planning, controlling organizations with the information about propagation of natural dangers, creation of land cadastre, selection of optimum places for building of linear and area units (Russia, USA, Switzerland, Austria and other.);
2. ecological control of region - influence of avalanches on the dynamics of landscapes, the nature and the boundaries of plant communities;
3. selection of tourist groups safe movement ways;
4. development of anti-avalanche activities;
5. study of interrelations of dangerous natural and anthropogenic phenomena (Russia, USA).

Conclusions

This article examines the geo-information system, intended for predicting of avalanche danger and decision making by the averting and overcoming of its consequences. Structure and tasks of the developed GIS-system are examined. The subsystem of prognostication is in detail represented, is described its information input, which consists of the data base, mathematical and algorithmic complexes, the block of the assignment of results. Is proposed the technology of the complex forecast of the avalanche danger, which includes the determination of the position of gathering, the classification of avalanches on the genesis, the determination of time and degree of the power of avalanches.

Bibliography

1. Kupcova A.V., Perekrest V.V.. GIS of Kabardino-Balkar republic created and working. Information bulletin of GIS Association. Moscow, 1996, № 3(5), p.24-25 (RUS)
2. Pertziger, F. 1998. Using of GIS technology for avalanche hazard mapping, scale 1:10 000. NGI, Oslo, pub. Nr.203, 210-214.
3. Bozhinskiy A.N., Losev K.S. General avalanche-caring. – Leningrad.: Hydrometeoizdat, 1987. – 280 p. (RUS)
4. Buser O., Fuhn, P., Gubler W., Salm B. Different methods for the assessment of avalanche danger. Cold. Reg. Sci. Technol., 1985, 10 (3), 199-218.
5. Rapin F. A new scale for avalanche intensity. International Snow Science Workshop., 2002, vol.2, 103-110
6. Fuhn P. An overview of avalanche forecasting models and methods. Oslo, NGI, Pub.N 203, 1998, 19-27.

Authors' Information

Alexander Ya. Kuzemin *– Prof. of Information Department, Kharkov National University of Radio Electronics, Head of IMD, Ukraine, e-mail: kuzy@kture.kharkov.ua*

Olesya Dyachenko – *phd student, Kharkov National University of Radio Electronics, Ukraine*

Darya Fastova – *phd student, Kharkov National University of Radio Electronics, Ukraine*

DEVELOPING AN EXPERT SYSTEM FOR SITUATIONAL ANALYSIS OF AVALANCHE DANGER

Alexander Kuzemin, Vyacheslav Lyashenko, Asanbek Toroyev, Ilia Klymov

Abstract: *A basic concept of an expert system for situational analysis of avalanche danger is proposed in this article. Avalanche danger classes and the possible methods of dividing microsituations in them are described.*

Keywords*: expert system, microsituation analysis, avalanche forecast*

Introduction

Conducting of constant monitoring and building of interpretation models for prediction of such situations initiation can be considered as one of emergency situations risk monitoring and, in particular, situations caused by avalanches. Hereinafter, such models make the basis of the decision-making support system; this is favorable for development of recommendations for timely performance of preventive measures directed to emergencies prevention [4].

Shortages of the traditional approaches application

Patterns similarity method and regression analysis [1, 2] are the most often chosen when considering methods and models for avalanches prediction. But the prediction results obtained with these methods are not always applicable and demonstrate a number of shortages: they require appreciable computational resources; they don't cover existing diversity of causes leading to an avalanche formation. Impossibility to define the avalanche danger degree, avalanches number and dimensions are also among shortages [3].

Avalanche danger initiation microsituations classes

Representation of diversity of their initiation factors in the form microsituations set help to increase faithfulness of analysis and avalanches slip prediction. Every such microsituation corresponds to a definite combination of the avalanche initiation medium factors. At the same time such a representation makes it possible to divide the whole set of causes acting on the avalanche initiation into two subclasses of situations. One of subclasses characterizes a set of microsituations representing the avalanche initiation and the other subclass is characteristic for avalanche non-dangerous situation in general. Then the emergency avalanche situations risks management can be presented the generalized description of the system with the set of different microsituations. Reasoning from such an interpretation the logic rules for generalization of the analyzed set o data for their further division into classes of the avalanche-dangerous and avalanche-non-dangerous situations were derived:

$$\text{“avalanche-dangerous”}=\left(\{F_{Л}^{Л}(X)\}/\{F_{Н}^{Н}(X)\}\right)\cup\left(\{F_{Н}^{Н}(X)\}/\{F_{Л}^{Л}(X)\}\cup\right)$$
$$\cup\left(\{F_{Л}^{Н}(X)\}/\{F_{Н}^{Л}(X)\}\right)\cup\left(\{F_{Н}^{Л}(X)\}/\{F_{Л}^{Н}(X)\}\right)\cup\left(\{F_{Л}^{Н}(X)\}/\{F_{Н}^{Н}(X)\}\right),$$

$$\text{“avalanche-non-dangerous”}=\cup\left(\{F_{Л}^{Н}(X)\}\cap\{F_{Н}^{Л}(X)\}\right)\cup\left(\{F_{Н}^{Н}(X)\}/\{F_{Л}^{Л}(X)\}\right),$$

where $F_{Л}^{Л}(X)$ $(F_{Л}^{Н}(X))$, $F_{Н}^{Л}(X)$ $(F_{Н}^{Н}(X))$ is a probability function of referring the avalanche-dangerous (avalanche-non-dangerous) microsituation to the avalanche- dangerous (avalanche-non-dangerous) class, respectively, with a set of the avalanche danger initiation medium factors X.

Further development of the analyzed data division into classes of the avalanche –dangerous and avalanche non-dangerous situations is an introduction of the integral measure of proximity between microsituations based on Wilcoxon test value; this makes it possible to obtain reasonable results using real data for the avalanche-dangerous Itagar Chychkan region of Kyrgyz Republic (Fig. 1).

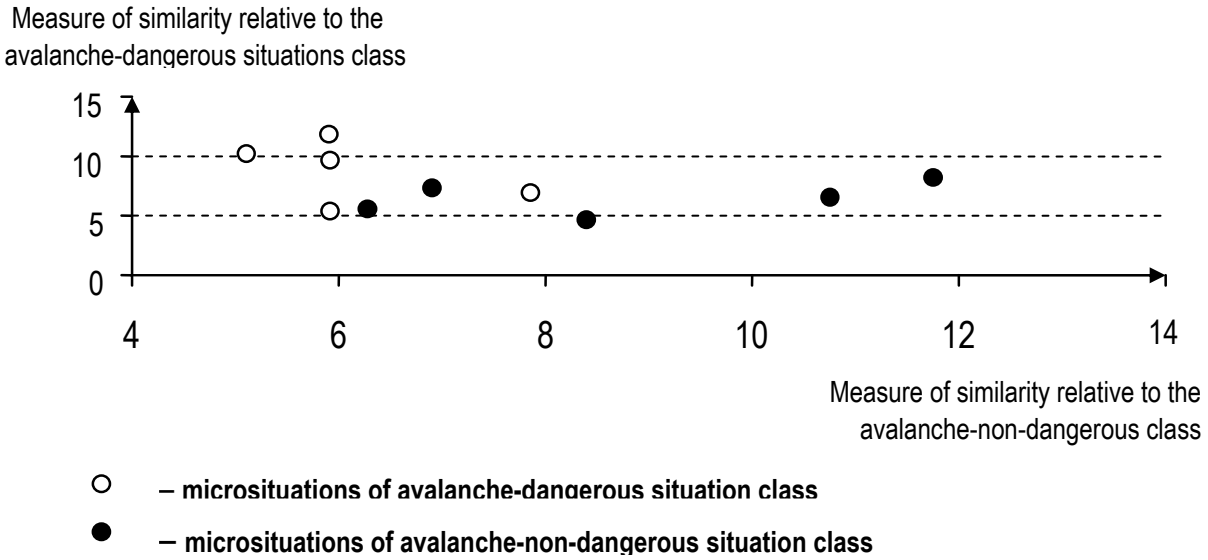


Fig.1. Distribution of microsituations in the space of features describing their relation to the avalanche-dangerous and avalanche-non-dangerous situations classes

Yet, in the conceptual plan the essence of the probabilistic aspect of analysis of the avalanche-dangerous initiation medium can be reduced to definition of the probability reference of some point as current characteristics of the considered medium either to the field of the avalanche-dangerous situation initiation or to the field of the avalanche-non-dangerous situation consideration.

The given approach can be treated also as a correspondence of the current characteristics of the avalanche climate initiation medium, whose parameters define some region, to the preceding probabilistic distributions of the avalanche-dangerous and avalanche-non-dangerous situations. Consequently, we may speak about the probability of correspondence of the investigated characteristics of the avalanche-dangerous climate initiation medium to the probabilistic distributions of the avalanche-dangerous and avalanche-non-dangerous situations.

Representation of the avalanche danger factors as the microsituations classes made it possible to obtain the objective correspondence between the probabilistic estimates of the avalanches slip and the degrees of the avalanche danger scale.

Starting from the analysis of the presented microsituations classes the positive result was obtained relative to correction of the prediction system response time to a possible avalanche slip. The essence of such an estimate consists in construction and analysis based on the fuzzy sets theory of the corresponding functions of the prediction time correction $\mu(X)$ (Fig. 2).

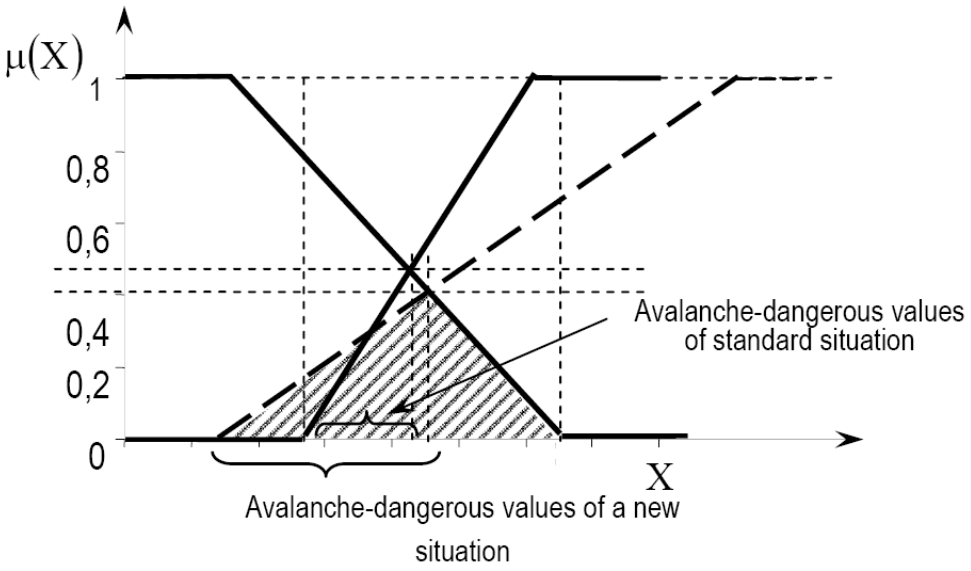


Fig.2. Methods of the theory of fuzzy sets as a basis for correction of the prediction system response time to a possible avalanche slip

Such analysis allow to make first steps in development of intelligent expert system which is able to analyze microsituations, predict possible avalance-dangerous events and track list of activites took in such situations. According to results system can propose a list of activites, which were used in previous situations alike this, helping an experts to quickly react on possible danger (see Fig. 3).

Such results can be achieved by constantly learning system. Each measurement is recorded during 'Learning mode' of the system. Special attention should be paid to each avalanche, forecasted by system or not. After each avalanche the data should be specified, defining more accurate terms of avalanche situations for this station/region (see Fig. 4.)

The effectiveness of this system was checked using statistic of avalanches provided by meteorological stations at Itagar, Chichkan district, Kyrgyz republic.

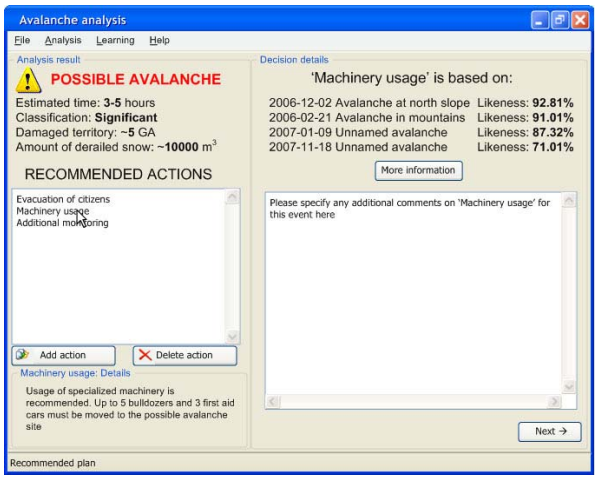


Fig 3. Analysis of possible avalanche situation

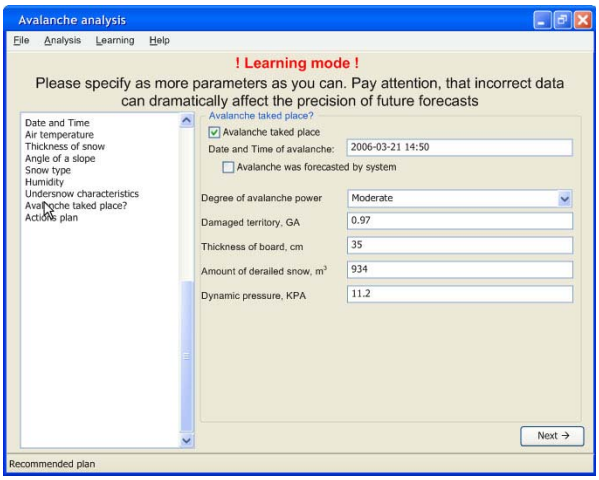


Fig 4. Learning mode of the system

Conclusion

The approach presented in this work makes it possible to receive an adequate description of the avalanche-dangerous medium initiation using a set of the corresponding microsituations and, as a result, to increase the analysis precision and prevent arising natural calamities in the form of the avalanches slip.

Bibliography

1. Buser, O., Butler, M. and Good, W. 1987. Avalanche forecast by the nearest neighbors method. IAHS Publ. 162. P. 557-569.
2. Durand Y., Brun E., Merindol L., Guyomarc'h, Lesaffre B., Martin E. A meteorological estimation of relevant parameters for snow models. Ann. Glaciol.,18, 1993, 65-71.
3. Fuhn P. An overview of avalanche forecasting models and methods. Oslo, NGI, Pub.N 203, 1998, 19-27.
4. Kuzemin A., Toroev A. Mobile means of control and prediction of avalanche climate using information conversion in acoustic range. IDRC-2006. – DAVOS, 2006. – Vol. 2. – P. 291–294.

Authors' Information

Kuzemin A.Ya.: *Prof. of Information Department, Kharkov National University of Radio Electronics, Head of IMD, (Ukraine), kuzy@kture.kharkov.ua*

Liashenko V.V. *– senior scientific employee, Kharkov National University of Radio Electron (Ukraine), kuzy@kture.kharkov.ua*

Toroyev A.A.: *director of joint-stock company "Computational techniques and automation systems", kuzy@kture.kharkov.ua*

Klymov I. M., *student of Kharkiv National University of Radioelectronics, speciality "Intelligent decision support systems", ilia.klimov@gmail.com*

USING THE BUSINESS PROCESS EXECUTION LANGUAGE FOR MANAGING SCIENTIFIC PROCESSES

Anna Malinova, Snezhana Gocheva-Ilieva

Abstract*: This paper describes the use of the Business Process Execution Language for Web Services (BPEL4WS/BPEL) for managing scientific workflows. This work is result of our attempt to adopt Service Oriented Architecture in order to perform Web services – based simulation of metal vapor lasers. Scientific workflows can be more demanding in their requirements than business processes. In the context of addressing these requirements, the features of the BPEL4WS specification are discussed, which is widely regarded as the de-facto standard for orchestrating Web services for business workflows. A typical use case of calculation the electric field potential and intensity distributions is discussed as an example of building a BPEL process to perform distributed simulation constructed by loosely-coupled services.*

Keywords*: BPEL, scientific workflows, Web services, SOA.*

ACM Classification Keywords*: D.2.12 Interoperability - Distributed objects*

Introduction

There is growing interest in the use of Web services infrastructures for scientific computing. Web services provide interoperability of various applications running on heterogeneous platforms. They enable dynamic connections and automation of business processes within and across scientific collaborations for application integration, reusability, and flexibility which is motivated mainly by the loosely coupled nature of the Web services. Web services became the preferred way to realize Service Oriented Architectures (SOA). SOA is architecture that represents software functionality as discoverable services on the network. All program functions and methods are exposed as services described by some universal description language (WSDL [1] in the case of Web services-based SOA). These interfaces can be invoked by other services to perform business processes.

The Business Process Execution Language (BPEL) [2] is an XML-based language used for integration of a number of Web services into more complex composite services. Thus BPEL enables the top-down realization of Service Oriented Architecture through composition, orchestration, and coordination of Web services. The BPEL composite services are called business processes and are managed by a workflow engine. BPEL processes orchestrate the interactions between the Web services using standard XML (SOAP) messages for communication. BPEL processes can be executed on any platform or product that compiles with the BPEL specification. BPEL supports the Web services technology stack, including SOAP, WSDL, UDDI, WS-Reliable messaging, WS-Coordination, and WS-Transaction.

This paper describes the use of BPEL for managing scientific workflows. The complex, unpredictable, and inter-dependent nature of the components in a scientific workflow leads to such requirements, concerning workflow language, as exception handling, recovery from uncertain situations, user interactions to facilitate interactive steering and monitoring, flexibility to support dynamic selection of services at runtime, etc. In the context of addressing these requirements, the BPEL specification features are discussed in the following sections.

We use BPEL to orchestrate Web services in order to perform simulation of metal vapor lasers. This includes providing of Web services wrappers of legacy scientific applications. We begin our workflow solution by defining a number of inter-related patterns that match the basic requirements of the users of the system. A typical use case of the calculation the electric field potential and intensity distributions is also provided as an example of building a BPEL process to perform distributed simulation constructed by loosely-coupled services. The BPEL processes we build are realized with the Oracle BPEL Process Manager and Designer and are deployed to the Oracle Application Server.

Scientific Workflows and BPEL4WS

The spectrum of what might be called scientific workflow is wide and includes scientific discovery workflows, workflows that automate manual procedures or reengineer custom tools, and data and compute-intensive workflows. Scientific workflow support is needed for practically all information-oriented scientific disciplines, including bioinformatics, chemistry, ecology, geology, physics, etc. In this section we provide a number of common requirements of scientific workflows: service composition and reuse, scalability, detached execution, reliability and fault tolerance, user interaction, monitoring, “smart” re-runs, data provenance, etc. ([5], [6], [7]).

In general, the BPEL vocabulary is tailored more to the requirements of business processes, which often have different requirements compared to scientific workflows. For example, in [5] is outlined that business workflow approaches focus on control-flow patterns and events, whereas dataflow is often a secondary issue. Scientific workflow systems, on the other hand tend to have execution model that are much more dataflow-oriented.

In this section we provide an analysis of the BPEL specification in the context of the above listed requirements. We do this also in the context of the implementation technology we have adopted, particularly the Oracle's BPEL Process Manager as part of the Oracle SOA Suite.

Service composition and reuse.

Web services can be combined in two ways: orchestration and choreography. In orchestration a central process (which can be another Web service) takes control of the involved Web services and coordinates the execution of different operations. The involved Web services do not "know" (and do not need to know) that they are taking part in a composition process. Choreography, in contrast, does not rely on central coordinator and occurs in pear-to-pear style workflows where communications occur directly between partners. All participants in the choreography need to be aware of the business process, operations to execute, messages to exchange, and the timing of message exchanges.

BPEL supports two different ways of describing business processes that support orchestration and choreography: Executable processes - they follow the orchestration paradigm and can be executed by an orchestration engine; Abstract business protocols - they allow specification of the public message exchange between parties only. They do not include the internal details of process flows and are not executable. They follow the choreography paradigm.

From the perspective of composing Web services to execute scientific processes, orchestration is a more flexible paradigm and has the following advantages over choreography:

- The coordination of component process is centrally managed by a known coordinator.
- Web services can be incorporated without their being aware that they are taking part of a larger process.
- Alternative scenarios can be put in place in case faults occur.

Scalability.

Some scientific workflows involve large volumes of data and/or require high-end computational resources, e.g. running a large number of parallel jobs on a cluster computer. To support such data-intensive and compute-intensive workflows, suitable interfaces to Grid middleware components are necessary.

Parallel flows enable a BPEL process to perform multiple tasks at the same time, which is useful when we need to perform several time-consuming and independent tasks. Concurrency is provided with the <flow> activity which causes all the activities nested within it to be executed concurrently. Control exits from <flow> when all nested activities terminate.

BPEL also provides features to support handling multiple requests. Multiple customer interactions can be handled concurrently by creating multiple instances of the process, one for each interaction. This is not a problem if the interaction consists of a single, synchronous invocation of an operation on a server, and the server does not invoke other servers in the process of handling the request. Scientific processes, however, often require long running conversations and transactions between partners. For workflows, this involves the addition of process identifiers that are embedded and exchanged between partners during a conversation. For handling such situations, BPEL allows a process to declare a *correlation set* local to a scope. This is a set of properties such that all messages having the same values of all the properties in the set are part of the same interaction and hence are handled by the same instance. Thus, a correlation set identifies a particular instance of among a set of instances of that process, and a correlation set and a port together uniquely identify a process instance among all process instances at a host machine.

Detached execution.

Long running scientific workflows require an execution node that allows the workflow control engine to run in the background on a remote server, without necessarily staying connected to a user's client application that has started and is controlling workflow execution.

In a BPEL process a web service can be invoked as a synchronous or asynchronous operation. Synchronous web services provide an immediate response to a query, and block the BPEL process for the duration of the operation. Asynchronous web services do not block the BPEL process, and are useful for environments in which a service can take a long time to process a client request. Asynchronous services also provide a more fault-tolerant and scalable architecture than synchronous services.

Reliability and fault tolerance.

A scientific workflow might incorporate a service that often "fail", change its interface, or just become unacceptably slow. Thus the workflow definition should support the definition of failure handling mechanisms.

BPEL provides a flexible structure for dealing with failures. Fault and compensation handlers are used to reverse the effects of partially completed interactions. The execution of these handlers is tied in with the concept of scopes. A scope serves to define the execution context of an activity. A scope can have a name and local declarations and encloses (possibly complex) activity to be executed. Declarations include, among other items, local variables, fault handlers, and a compensation handler. Compensation applies only to scope's external effects – the effects it has invoked at other sites. On entry, a scope's compensation handler is given a snapshot of the process's state at the time control exited (normally) from the scope. Since it can access only the snapshot and not the variables themselves, compensation cannot affect the state of the process and applies only to external activities. Faults signal failure and start the process of reversing the effects of an interaction. A fault might be raised in a process if it gets a fault response to an operation that it has invoked synchronously. Alternatively, a process might explicitly execute a <throw> activity if it recognizes that an anomalous situation has arisen.

User interaction.

Many scientific workflows require user decisions and interactions at various steps. An interesting challenge is the need for user interaction in a detached execution. Using a notification mechanism the user might be asked to reconnect to the running instance and make a decision before the paused (sub-) workflow can resume.

BPEL 1.1 and 2.0 do not include human interactions and are limited to service orchestration. Oracle BPEL Process Manager provides manual task Web service to integrate people and manual tasks into BPEL processes [8]. By implementing this as a true BPEL service, the interface to the task service is described with WSDL and people can be included in 100% standard BPEL processes – to the BPEL process, the person/manual task looks like any other asynchronous Web service. User notification is also supported – anything that can be done in BPEL (invoking a Web service, sending an e-mail message or JMS message, executing some Java code, and so on) can be done to notify a user of a task-related event.

Monitoring.

Scientific workflows are potentially long running activities and it is of importance to scientists to be able to observe and monitor the ongoing execution of a workflow.

Oracle BPEL Manager provides sensors to monitor BPEL activities, variables, and faults during runtime [8]. The following types of sensors can be defined, either through the BPEL Designer or manually by providing sensor configuration files:

- Activity sensors: Used to monitor the execution of activities within a BPEL process. For example, activity sensors can be used to monitor the execution time of an invoke activity or how long it takes to complete a scope. Along with the activity sensor, the variables of the activity might be also monitored.
- Variable sensors: Used to monitor variables (or parts of a variable) of a BPEL process. For example, variable sensors can be used to monitor the input and output data of a BPEL process.
- Fault sensors: Used to monitor BPEL faults.

The following two requirements are much desirable for scientific workflow systems, although difficult to implement:

"Smart" re-computations.

A special kind of user interaction is the change of a parameter of a workflow. A "smart" re-computation (re-run) would not execute the workflow from scratch, but only those parts that are affected by a parameter change. Another useful technique in this context is the ability to backtrack (in the case or parameter change or even a system failure) to a previously saved state without starting over from scratch.

Data provenance.

Computational experiments and runs of scientific workflows should be reproducible and indicate which specific data products and tools have been used to create a derived data product. A scientific workflow system should be able to automatically log the sequence of applied steps, parameter settings and intermediate data products. A related requirement is automatic report generation: The system should allow the user to generate reports with all relevant provenance and runtime information, e.g., in XML format for archival and exchange purposes, and in HTML (generated from the former, e.g., via a XSLT script) for human consumption.

While the above list of requirements for scientific workflow systems is by no means complete, it should be sufficient to capture many of the core characteristics. Other requirements include the use of an intuitive GUI to allow the user to compose a workflow visually from smaller components to animate workflow execution, to inspect intermediate results, etc., although these requirements are not related with the BPEL specification itself. Rather, they depend on the BPEL Designer and engine vendor.

Basic Patterns of the Workflow Solution

In this section we define a number of inter-related patterns that match the basic requirements of the users of the Web services-based system for simulation of metal vapor lasers we are in process of developing. Our approach is to decompose the simulation processes into a collection of basic patterns that can be fully automated. These are hierarchically organized, as they are co-dependant, as presented in Figure 1:

- Simulation execution and monitoring pattern: This is the most basic workflow pattern. It presents the ability to submit a simulation, monitor its execution, and handle any potential faults (which may include resubmitting the job if needed).
- Data retrieval and storage pattern: Adds to the above the capability of retrieving data from the data storage before job submission and uploading results to the storage upon completion.
- Metadata management: Generated metadata for the created file, and uploading these to the metadata storage.

This hierarchical organization maps well to BPEL-base workflows. Because these are themselves presented as Web services, it is possible for a workflow to incorporate other workflows into its structure. Not all capabilities presented, such as metadata capture, may be required for every job of the scientific process that the user wishes

to execute. This hierarchical approach also favors reusability and enables the potential addition of new services and its incorporation into larger workflows.

A data/ metadata management tools will be integrated with the workflow tools using web services technology. This integration is essential to allow the automatic collection and publishing of metadata relating to the simulations along with efficiently handling the data files themselves.

A Typical Use Case

In this section we describe an example workflow for calculation of the electric field potential and intensity distributions of the radio-frequency discharge in He-Cd laser. The obtained results are necessary for further investigation of the dependence between different laser characteristics, oriented for practical engineering purposes in order to improve the total efficiency of the real He-Cd laser.

We have wrapped legacy FORTRAN codes as Java modules trough the use of the Java Native Interface [3]. Then we have transformed these modules into Web services. These are:

- "Electrode" Web service – for a predefined geometry of the 2D cross-section of the laser design this module generates the appropriate mesh for discretization and classifies the points in the different sub-regions: outer electrodes, laser tube, etc. Basic input parameter is the applied to the electrodes voltage, which is then used to determine the boundary conditions.
- "Poisson" Web service – This module provides numerical solution of the Poisson equation with which we model the potential equation.

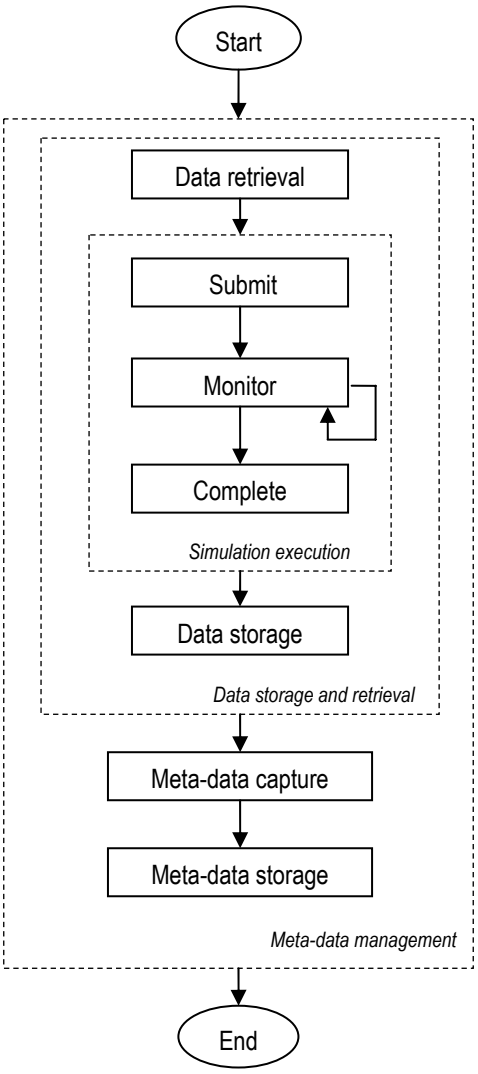


Figure 1. Basic simulation workflow patterns.

The BPEL process first invokes a synchronous "GetData" service to retrieve data from the data storage. The data retrieved is a number of parameters predefined for the chosen laser type. These parameters are stored natively in XML files. An XML schema was created to describe these parameter sets. It consists of two types: *parameterset* and *parameter*. The type *parameter* is an abstract type which all parameters extend. Thus we can manage different parameter sets relevant to different laser types, as well to have different parameter set instances of a particular laser type. We have also created an Oracle's BPEL Manager Human Workflow Task service to incorporate a user task in the BPEL process and particularly the parameter approval. This enables the user to change some of the predefined values of physical constants, geometrical parameters, etc.

Then the "Electrode" service is invoked synchronously. The service creates output file which is next used as input for the "Poisson" service to calculate the electric field intensity and to save the output in a file.

Conclusion

From our work so far we conclude that BPEL is fully applicable for orchestrating scientific services. Furthermore BPEL could serve as a standard representation for scientific workflows and hence aid reproducibility. In addition the Oracle BPEL Process Manager, which we use, has built in support for the use of Web Services Invocation Framework (WSIF) [4]. Thus a direct Java code injection into a workflow script is enabled. This is a possible way to overcome some of the limitations of BPEL specification.

Acknowledgments

This work is supported by the National Science Fund of the Bulgarian Ministry of Education and Science, Project **VU-MI-205/2006**.

Bibliography

[1] Web Services Description Language (WSDL). http://www.w3c.org/TR/wsdl

[2] Business Process Execution Language (BPEL). http://www-128.ibm.com/developerworks/library/specification/ws-bpel/

[3] Java Native Interface (JNI). http://java.sun.com/j2se/1.5.0/docs/guide/jni/index.html

[4] Web Services Invocation Framework (WSIF). http://ws.apache.org/wsif

[5] B.Ludascher, I.Altintas, C.Berrkley, D.Higgins, E.Jaeger, M.Jones, E.Lee, J.Tao Y.Zhao. Scientific workflow management and the Kepler system. In: Concurrency and Computation: Practice & Experience.Vol. 18, 2006., p. 1039 - 1065.

[6] A.Akram, D.Meredith, R.Allan. Application of Business Process Execution Language to scientific workflows. In: International Transactions on Systems Science and Applications (accepted 2006).

[7] N.Joncheere, W.Vanderperren, R.Straeten, Requirements for a Workflow System for Grid Service Composition. In:Proceedings of the 2nd International Workshop on Grid and Peer-to-Peer Based Workflows (GPWW 2006), Vienna, Austria, September 2006. LNCS Springer-Verlag.

[8] D.Bradshaw, M.Kennedy. Oracle® BPEL Process Manager Developer's Guide10g (10.1.3.1.0). http://download-east.oracle.com/docs/cd/B31017_01/integrate.1013/b28981/toc.htm

Authors' Information

Anna Malinova – *University of Plovdiv "Paisii Hilendarski", 24 Tzar Asen St., Plovdiv-4000, Bulgaria; e-mail: malinova@pu.acad.bg*

Snezhana Gocheva-Ilieva – *University of Plovdiv "Paisii Hilendarski", 24 Tzar Asen St., Plovdiv-4000, Bulgaria; e-mail: snow@pu.acad.bg*

AUTOMATED SYSTEM FOR EFFECTIVE INTERNET MARKETING CAMPAIGN (ASEIMC)

Todorka Kovacheva

Abstract: *The purpose of the paper is to present an automated system for realization of effective internet marketing campaign (ASEIMC). The constantly growing number of websites available online brings more problems for the contemporary enterprises to reach their potential customers. Therefore the companies have to discover novel approaches to increase their online sales. The presented ASEIMC system gives such an approach and helps small and medium enterprises to compete for customers with big corporations in the Internet space.*

Keywords: *internet marketing, internet strategy, marketing strategy, search engine optimization, web promotion*

Introduction

Internet is constantly and rapidly growing network. Everyday thousands of new websites become available online. Many businesses try to sell their products and services through Internet. The result: millions of website promoting and selling the same products and services. The competition is at high level. The companies from one business branch compete for visitors and customers online with other companies from the same business branch.

The constantly growing number of websites available online brings more problems for the contemporary enterprises to reach their potential customers. Therefore the companies have to discover novel approaches to increase their online sales.

The scope of the Internet Marketing

Internet marketing is a combination of activities directed to increasing of the rating of the site of the campaign in Internet, increasing of its traffic and as a result attracting of new clients and growing of the campaign One of the major activity of the Internet marketing is the search engine positioning. The major practices for this goal realization are:

- Code optimization
- Content optimization
- Technical processing
- Link popularity

The Internet marketing is realized profound analysis and market situation research (competitor activity, position of the branch, trends in the demand and supply), evaluation of the possibilities for using Internet for realization of competitive goods and services etc. It includes all activities related to the advertising activity of the company in Internet space: from design of the company's site to the sales realization.

Internet Marketing combines methods of using the Internet to promote and sell products and services. It includes:

- Information management – the management of information from various sources to one or more audiences, control over the structure, processing and delivery of information.
- Public relations – building sustainable relations with all publics in order to create a positive brand image.
- Customer service – the provision of service to customers before, during and after a purchase.
- Sales – the act of meeting buyers and providing them with a service for a negotiated compensation. Selling is a practical implementation of marketing.

Internet Marketing is important for all companies selling their products and services online. More than 2/3rds of all customers visit the web on regular basis and search for different types of products or services. Many companies compete for that customers and pay a high amount of money for developing of the effective internet marketing campaign.

Therefore to overcome the competition we need a competitive solution. The ASEIMC is designed as such solution.

Description of ASEIMC

The main purpose of the Automated System for Effective Internet Marketing Campaign is to increase the amount of sales of the company. At the end of the campaign we want to have more income than costs. The sales amount must be higher than the costs for the campaign. As bigger the positive difference between the profit and the invested costs as more effective is the internet marketing campaign.

Because of the big number of sites published in World Wide Web and big amount of repeating information it is very difficult for the modern enterprises to develop effective adverting strategy and which to be conformed to the limited budget of small and medium enterprises. The purpose of the automated system developed by us is to provide considerable competitive advantages of small and medium enterprises in the Internet space doing small investments.

The system works by doing preliminary survey (gathering of information) for the behavior and the specifics of the competing companies and branches and the data are gathered in a database which specially designed for this purpose. On the basis of these data an analysis of the strong and weak assets of the target company, whose adverting strategy is developed, is made. The information gathered in the database is used for analysis of the trends in the developing of close branches and working out forecast for probable innovations which will be made and which will affect the company activity directly or indirectly.

The system for gathering of information is a Web spider, which is developed by the author and which similarly to the traditional web robots travels over the Internet space and collects diverse data from the information published in World Wide Web.

The Architecture of ASEIMC

To realize its goals the AISEIMC is divided in the following modules:

1. Module for gathering of information from the global net. It is a web robot, designed for the purpose.
2. Database, in which the information gathered from module 1 is stored.
3. Module for analysis of the data from the database and working out of forecast of trends in the behavior of the competitors and competing branches.
4. Goal definition module. Before continuing with the planning of the company's advertising campaign in Internet it is necessary to define goals, which we wish to achieve through this advertising campaign. These goals may be:
    a. To impose trade mark of the company;
    b. To increase the number of the sales;
    c. To enter new markets;
    d. To promote a new product of service of the company;
    e. To inform the society for changes in the company's policy etc.

    This feature is realized through set of questions q which are logically ordered one after another. As a result the proper target zone is defined. These goals have not to contradict to the strategic company management.

5. Analysis of current state. Before continuing with planning and realization of the advertising campaign it is necessary to specify the position from which the company starts. This is necessary not only to be able to develop effective advertising strategy but also to evaluate the achieved results at the end of the advertising campaign.
6. Budget validation. A significant point in developing the company's advertising campaign is specifying of the financial funds which we have. Their amount is a limit, which defines the boundaries of carrying out our advertising campaign.
7. Keyword selection module. Next stage is to choose the key words and expressions which will help for advertising campaign realization. These words have to be the ones used by the users. This module provides information about which key words we have to use taking into consideration the searching for them by the users and the content of the site.
8. Competition analysis module. After the key words of campaign are specified, it is important to evaluate the chances of the company to become a leader in the most popular search engines by key words used by users. Therefore the competitors' web sites are analyzed according to these keywords. Their advantages and disadvantages are evaluated. Their behavior is forecasted. As a result the module prints a report containing instructions for the developing search engine optimization strategy.
9. Developing search engines optimization strategies module. It takes the report from the previous and builds the optimization strategy, which the company has to follow.
10. Website optimization and promotion module. The ability to realize the actual optimization of the web site based on the strategy developed in item 6. The promotion of the site in the search engines and Internet space is carried out.
11. Result evaluation module. At the end of each advertising campaign it is important the achieved results to be evaluated and effectiveness of the campaign to be specified.
12. Improving of the company's advertising strategy. This is a continuous activity directed to increasing the results.

Conclusion

The system designed by the author helps for incresing of the popularity of the company's web site and as a result of this it leads to increasing of the number of sales and the amount of the realized profit from the company activity. It can be used effectively not only by small and medium enterprises but also by transnational corporations. It contributes to achieving signifficant competative advantages in Internet space at comparatively small investments.

Bibliography

1. Kovacheva T., Extended Executive Information System, International Journal "Information Theories & Applications", Vol.11, Number 4, pp.394-400, 2004
2. Kovacheva T., Toshkova D., Neural Network Based Approach For Developing The Enterprise Strategy, International Journal "Information Theories & Applications", Vol.13, Number 1, pp.139-145, 2006

Author's Information

Todorka Kovacheva *– Gluon Technologies Ltd, Varna, Bulgaria, www.gluontechnologies.com, e-mail: todorka_kovacheva@yahoo.com, phone: +359899920659*

APPLICATION OF METHOD OF THE WEIGHED TOTAL FOR DIAGNOSTIC INDEX SIGNIFICANCE CALCULATION IN DIFFERENTIAL DIAGNOSTICS OF DERMATOLOGICAL DISEASES

Anatoly Bykh, Elena Visotska, Olga Kozina, Anna Tikhonova, Andrey Porvan, Alexander Zhook

Abstract: *The basic methods of decisions making in multi-criterion conditions are considered, from which the method of the weighed total for calculation of diagnostic indexes significance in differential diagnostics of dermatological diseases is chosen.*

Keywords: *dermatology, differential diagnostics, method of the weighed total, multi-criterion task.*

ACM Classification Keywords*: Decision Making*

Introduction

The external cover of human body – skin, is constantly exposed to influence of various factors of external environment that is why pathological processes are developing in skin. In everyday clinical practice the dermatologists quite often run into a situation, when discerning of reliable clinical diagnosis is difficult. It is explained by both a plenty of possible displays of dermatological pathologies and many symptoms are characteristics of different diseases. A different importance of the symptom is not equal for different diseases, that still more complicates differentiation of diagnoses [Ананьев, 2005]. This is the reason why task of differential diagnostics of dermatological diseases can considered like a multicriterion task.

Determination of significance of diagnostic indexes

Normative (formal) methods, which assume that an expert know the definite rational method of correct decision choice, are basic methods of decisions making in multicriterion conditions [Брахман, 1984]. Depending on the role of an expert in forming and grounding of importance of alternative diagnostic decision all formal methods divide by axiomatic, lines, methods of compensation, methods of thresholds of incomparableness and man-machine methods. From these methods for task of differential diagnostics of dermatological diseases direct methods are most useful. Due to such group of direct methods an expert can formulate of resulting significance of the given symptom for each disease as dependence on its estimations on private criteria without the theoretical grounds, and the parameters of this dependence (weight and expressed of symptom) are formed directly by the method of expert estimations [Кац, 2004]. Because for every disease there is the set of diagnostic indexes which fully describe one, for calculation of significance of this set for differential diagnostics of the given dermatological disease the method of the weighed total can be apply expediently:

$$U_i = \sum_{j=1}^{m} \delta_{ij} f(x_j),$$

where $\sum_{j=1}^{m} \delta_{ij} = 1$, δ_{ij} – weight of *j*-th diagnostic index for the *i*-th disease;

x_j – estimation of degree of *j*-th diagnostic index.

As well as for all direct methods, for the method of the weighed total rigorists to the experts are characteristic, especially on the initial stages of work at forming of expert estimations.

Conclusion

It is possible to do a conclusion, that a plenty of methods of decision of the multicriterion tasks oriented to the concrete problem situations is presently developed. For estimation of significance of diagnostic indexes at differential diagnostics of dermatological diseases expediently to use the weighed total method. Application of this method will allow improving quality of differential diagnosis discerning.

Bibliography

[Ананьев, 2005] О. Л. Ананьев, Е. В. Анисимова, Н. В. Иваничкина и др. Кожно-венерические заболевания. Полный справочник. М.: Эксмо, 2005.

[Брахман, 1984] Т.Р. Брахман. Многокритериальность и выбор альтернативы в технике - М; Радио и связь, 1984. 287с

[Кац, 2004] М.Д. Кац. Использование искусственного интеллекта для разработки методов дифференциальной диагностики внутри групп трудно различимых заболеваний. Клиническая информатика и телемедицина. 2004. №1. С.86 – 89.

Authors' Information

Anatoly Bykh *– Doctor of Physics and Mathematics, professor, Head of Biomedical Electronics Department of Kharkov National University of Radio Electronics*

Elena Visotska *– PhD, lecturer of Biomedical Electronic Devices and Systems Department of Kharkov National University of Radio Electronics*

Olga Kozina *– PhD, lecturer of Computers and Programing Department of National Technical University 'KPI'*

Andrey Porvan *– engineer of Biomedical Electronic Devices and Systems Department of Kharkov National University of Radio Electronics*

Anna Tikhonova *– engineer of Biomedical Electronic Devices and Systems Department of Kharkov National University of Radio Electronics*

Alexander Zhook *– student of Biomedical Electronic Devices and Systems Department of Kharkov National University of Radio Electronics*

Kharkov National University of Radio Electronics, Ukraine, 61166, Lenin Avenue, 14, Biomedical Electronic Devices and Systems Department , e-mail: diagnost@kture.kharkov.ua

Software Engineering

ADVANCE OF THE ACCESS METHODS

Krassimir Markov, Krassimira Ivanova, Ilia Mitov, Stefan Karastanev

***Abstract**: The goal of this paper is to outline the advance of the access methods in the last ten years as well as to make review of all available in the accessible bibliography methods.*

***Keywords**: Access Methods, Overview of the Access Methods*

***ACM Keywords**: D.4.3 File Systems Management, Access methods*

Introduction

The Access Methods (AM) had been available from the beginning of the developing the computer peripheral devices. As many devices there exists so many possibilities for developing different AM we have. Our attention is focused only to the access methods for devices for permanently storing the information with direct access such as magnetic discs, flash memories, etc.

In the beginning, the AM ware functions of the Operational Systems Core or so called Supervisor, and ware executed via corresponded macro-commands in the assembler languages [Stably, 1970] or via corresponding input/output operators in the high level programming languages like FORTRAN, COBOL, PL/I, etc.

Establishing of the first data bases in the 60-ties years of the last century caused gradually accepting the concepts "physical" as well as "logical" organization of the data [CODASYL, 1971], [Martin, 1975]. In 1975 the concepts "access method", "physical" and "logical" are clearly separated. In the same time Christopher Date [Date, 1977] specially remarked:

"The Data Base Management System (DBMS) does not know anything about:

a) physical records (blocks);
b) how the stored fields are integrated in the records (nevertheless that in many cases it is obviously because of theirs physical disposition);
c) how the sorting is realized (for instance it may be realized on the base of physical sequence, using an index or by a chain of pointers);
d) how is realized the direct access (i.e. by index, sequential scanning or hash addressing).

This information is a part of the structures for data storing but it is used by the access method but not by the DBMS. "

Every access method presumes an exact organization of the file which it is operating with and has no relation to the interconnections between the files, respectively – between the records of one file and that in the others files. These interconnections are controlled by the physical organization of the DBMS.

So, in the DBMS we may distinguish four levels:

- access methods of the core (supervisor) of the operation system;
- specialized access methods which upgrade these of the core of the operating system;
- physical organization of the DBMS;
- logical organization of the DBMS.

During the 80-ies years the "Multi-Dimensional Access Methods" had raised. In accordance with them the corresponded "spatial information structures" and the "spatio-temporal information structures" had risen, too. These AM developed the methods of the operating systems via specializing them to the give data models. From different point of view this period had been presented in [Ooi et al, 1993], [Gaede, Günther, 1998], [Arge, 2002], [Mokbel et al, 2003], [Moënne-Loccoz, 2005].

Usually the "one-dimensional" (linear) AM are used in the classical applications, based on the alpha-numerical information, whereas the "multi-dimensional" (spatial) methods are aimed to serve the work with graphical, visual, multimedia information. Now a special attention is given to the multi-dimensional AM. Maybe one of the most popular analyses is given in [Gaede, Günther, 1998]. The authors presented a scheme of the genesis of the basic multi-dimensional AM and theirs modifications. This scheme firstly was proposed in [Ooi et al, 1993] and it was expanded in [Gaede, Günther, 1998]. An extension in direction to the multi-dimensional spatio-temporal access methods was given in [Mokbel et al, 2003].

This work continues the investigation provided in [Markov, 2006].

The main goal of this paper is to present a new variant of this scheme. It is presented on Fig.1. In it the new access methods created after 1998 are added. The methods presented in [Gaede, Günther, 1998] are marked in italics and methods presented in [Mokbel et al, 2003] are underlined. Access methods, which are given in the two surveys simultaneously, are marked in underlined italics. In the appendix of this paper the corresponded bibliography is given.

The access methods presented on Fig.1 we may classify as follow:

- One-dimensional AM;
- Multidimensional Spatial AM;
- Metric Access Methods;
- High Dimensional Access Methods;
- Spatio-Temporal Access Methods.

One-dimensional Access Methods

One-dimensional AM are based on the concept "record". Let remember that the "record" is a logical sequence of fields which contain data eventually connected to unique identifier (a "key"). The identifier (key) is aimed to distinguish one sequence from another [Stably, 1970]. The records are united in the sets, called "files". There exist three basic formats of the records – with fixed, variable and undefined length.

In the ***context-free methods*** the storing of the records is not connected to theirs content and depends only on external factors – the sequence, disk address or position in the file. The necessity of stable file systems in the operating systems does not allow a great variety of the context-free AM. There are three main types well known from 60-ies and 70-ies years: *Sequential Access Method (SAM); Direct Access Method (DAM) and Partitioned Access Method (PAM)* [IBM, 1965-68].

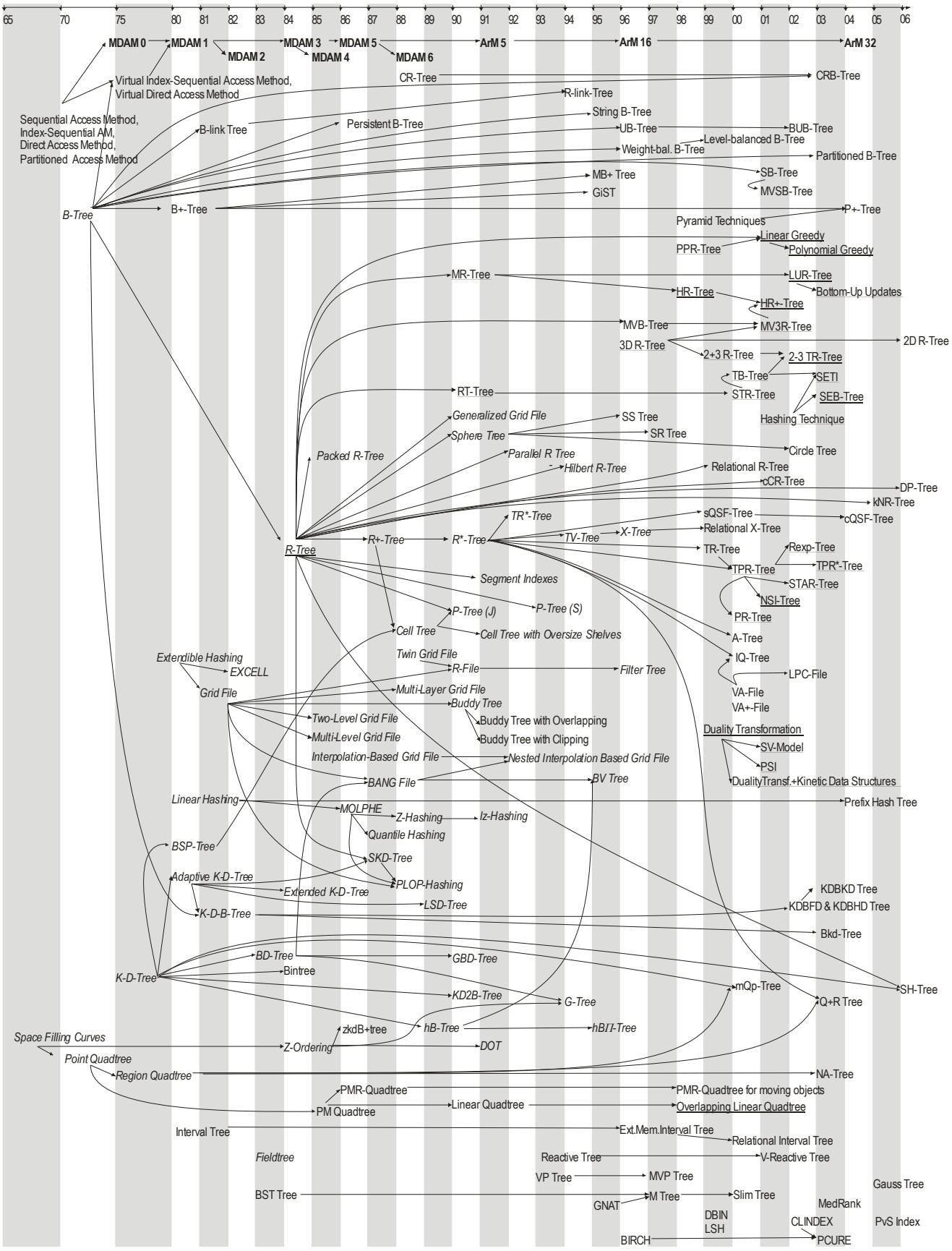


Fig. 1. Genesis of the Access Methods and their modifications
extended variant of [Gaede, Günther, 1998] and [Mokbel et al, 2003]

The main idea of the ***context-depended AM*** is that the part of the record is selected as a key which is used for making decision where to store the record and how to search it. This way the content of the record influences on the access to the record.

Historically, from the 60-ies years of the last century the attention is directed mainly to this type of AM. Modern DBMS are built using context-depended AM such as: unsorted sequential files with records with keys; sorted files with fixed record length; static or dynamic hash files; index file and files with data; clustered indexed tables [Connolly, Begg, 2002].

Multidimentional Spatial Access Methods

Multidimensional Spatial Access Methods are developed to serve information about spatial objects, approximated with points, segments, polygons, polyhedrons, etc. The implementations are numerous and include traditional multi-attributive indexing, geographical information systems and spatial databases, content indexing in multimedia databases, etc.

From the point of view of the spatial databases can be split in two main classes of access methods – Point Access Methods and Spatial Access Methods [Gaede, Günther, 1998].

Point Access Methods are used for organizing multidimensional point objects. Typical instance are traditional records, where on every attribute of the relation corresponds one dimension. These methods can be separated in three basic groups:

- Multidimensional Hashing (for instance Grid File and its varieties, EXCELL, Twin Grid File, MOLPHE, Quantile Hashing, PLOP-Hashing, Z-Hashing, etc);
- Hierarchical Access Methods (includes such methods as KDB-Tree, LSD-Tree, Buddy Tree, BANG File, G-Tree, hB-Tree, BV-Tree, etc.);
- Space Filling Curves for Point Data (like Peano curve, N-trees, Z-Ordering, etc).

Spatial Access Methods are used for working with objects which have arbitrary form. The main idea of the spatial indexing of non-point objects is using of the approximation of the geometry of the examined objects to more simple forms. The most used approximation is Minimum Bounding Rectangle (MBR), i.e. minimal rectangle, which sides are parallel of the coordinate axes and completely include the object. There exist approaches for approximation with Minimum Bounding Spheres (SS Tree) or other polytopes (Cell Tree), as well as their combinations (SR-Tree).

The usual problem when one operates with spatial objects is their overlapping. There are different techniques to avoid this problem. From the point of view of the techniques for organization of the spatial objects Spatial Access Methods can be split in four main groups:

- Transformation – this technique uses transformation of spatial objects to points in the space with more or less dimensions. Most of them spread out the space using space filling curves (Peano Curves, z-ordering, Hibert curves, Gray ordering, etc.) and then use some of point access method upon the transformed data set. For instance UB-Tree [Bayer, 1996], is variant of B-Tree, where keys are region addresses, sorted via "≤" and z-ordering;
- Overlapping Regions – here the data set are separated in groups; different groups can occupy the same part of the space, but every space object associates with only one of the groups. The access methods of this category operate with data in their primary space (without any transformations) eventually in overlapping segments. Methods, which use this technique includes R-Tree, R-link-Tree, Hilbert R-Tree, R*-Tree, Sphere Tree, SS-Tree, SR-Tree, TV-Tree, X-Tree, P-Tree of Schiwietz, SKD-Tree, GBD-Tree, Buddy Tree with overlapping, PLOP-Hashing, etc.;
- Clipping – this technique use eventually clipping of one object to several sub-objects, which will be stored. The main goal is to escape overlapping regions. But this advantage can lead tearing of the objects, extending

of the resource expenses and decreasing of the productivity of the method. Representatives of this technique are R+-Tree, Cell-Tree, Extended KD-Tree, Quad-Tree, etc.;

– Multiple Layers – this technique can be examining as variant of the techniques of Overlapping Regions, because the regions from different layers can overlap. But there exist some important differences: first – the layers are organizing hierarchically; second – every layer split primary space in different way; third – the regions of one layer never overlaps; fourth – the data regions are separated from space extensions of the objects. Instances for these methods are Multi-Layer Grid File, R-File, etc.

Metric Access Methods

Metric Access Methods deal with relative distances of data points to chosen points, named anchor points, vantage points or pivots [Moënne-Loccoz, 2005]. These methods are designed to limit the number of distance computation, calculating first distances to anchors, and then finding searched point in narrowed region. These methods are preferred when the distance is highly computational, as e.g. for the dynamic time warping distance between time series. Presentatives of these methods are: Vantage Point Tree (VP Tree), Bisector Tree (BST-Tree), Geometric Near-Neighbour Access Tree (GNNAT), as well as the most effective from this group – Metric Tree (M-Tree) [Chavez et al, 2001].

High Dimensional Access Methods

Increasing of the dimensionality strongly aggravates the qualities of the multidimensional access methods. Usually these methods exhaust their possibilities till dimensions around 15. Only X-Tree reaches the boundary of 25 dimensions, after then this method gives worse results then sequential scanning [Chakrabarti, 2001].

The exit of this situation is based on the data approximation and query approximation in sequential scan. These methods form a new group of access methods – High Dimensional Access Methods.

Data approximation is used in VA-File, VA+-File, LPC-File, IQ-Tree, A-Tree, P+-Tree, etc.

Because in high dimensional access methods the selectivity of the methods makes worse, it is allowed some answers inaccuracy. For query approximation two strategies can be used:

– examine only a part of the database, which is more probably to contain resulting set – as a rule these methods are based on the clustering of the database. Some of these methods are: DBIN, CLINDEX, PCURE;

– splitting the database to several spaces with fewer dimensions and searching in each of them. Here two main methods are used:

1) Random Lines Projection (presentatives of this approach are MedRank, which uses B+-Tree for indexing every arbitrary projection of the database, and PvS Index, which consist of combination of iterative projections and clustering);

2) Locality Sensitive Hashing, which is based on the set of local-sensitive hashing functions [Moënne-Loccoz, 2005].

Spatio-Temporal Access Methods

The Spatio-Temporal Access Methods have additional defined time dimensioning. [Mokbel et al, 2003]. They operate with objects, which change their form and/or position during the time. According to position of time interval in relation to present moment the Spatio-Temporal Access Methods are divided to:

– indexing the past, i.e. methods for operating with historical spatio-temporal data. The problem here is continuously increasing of the information over time. To overcome the overflow of the data space two approaches are used – sampling the stream data at certain time position or update the information only when data is changed. Spatio-temporal indexing schemes for historical data can be split in three categories: first category includes methods that manages spatial and temporal aspects into already existing spatial methods;

second can be explained as snapshots of the spatial information in each time instance; the third category focus on trajectory-oriented queries, while spatial dimension lag on second priority. Presentatives of this group are: RT-Tree, 3DR-Tree, STR-Tree, MR-Tree, HR-Tree, HR+-Tree, MV3R-Tree, PPR-Tree, TB-Tree, SETI, SEB-Tree;

- indexing the present. In contrast to previous methods, where all movements are known, here current positions are neither stored nor queried. Some of the methods, which answer of the questions of the current position of the objects are 2+3R-Tree, 2-3TR-Tree, LUR-Tree, Bottom-Up Updates, etc.;
- indexing the future. These methods have to answer on the questions about current and future position of moving object – here are embraced the methods like PMR-Quadtree for moving objects, Duality Transformation, SV-Model, PSI, PR-Tree, TPR-Tree, TPR*-tree, NSI, VCIR-Tree, STAR-Tree, R^{EXP}-Tree.

Conclusion

In this paper we presented a short overview of the current state in the field of development of the access methods. During the last four decades the access methods have been developed toward plenty of modifications of small number basic ideas. It is important to remark that the research has been provided on software as well as on hardware levels. For instance, in [Schlosser et al, 2005] a technology for storing of multi-dimensional data with physically preserving the multi-dimensionality of the data is presented.

The developed multi-dimensional index structures are effective for the small number of dimensions (from 2 – 5 up to 10 -15) and are uncomfortable for multi-dimensional spaces which are typical for the contemporary practical problems and the linear scanning may be preferable in many cases [Chakrabarti, 2001]. This is known as "Curse of dimensionality".

The concept of ”curse of dimensionality” was first coined by Richard Bellman [Bellman 1961]. He employed it to describe the problem caused by the exponential increase of the volume with the augmentation of the space dimension when addressing the problem of optimizing functions with several variables. Later, the term was used to indicate, more generally, non-intuitive phenomena observed when the dimension of data increases [Bouteldja et al, 2006].

The survey of the access methods suggests that the context-free multi-dimensional access methods practically are not available. One step in developing such methods is the Multi-domain Access Method introduced in [Markov, 2004].

We have no place to present all access methods in details. The main goal was to collect the basic publications of the most popular access methods. The further survey needs to be provided to present current state of the art in this area.

Appendix 1. Access Methods and Corresponded Publications

Access Method	Publicated in
2+3 R-Tree	[Nascimento, 1999] M. A. Nascimento, J.R.O. Silva, Y. Theodoridis. *Evaluation of Access Structures for Discretely Moving Points*. In Proc. of the Intl. Workshop on Spatio-Temporal Database Management, STDBM, pages 171–188, Sept. 1999.
2-3 TR-Tree	[Abdelguerfi et al, 2002] M. Abdelguerfi, J. Givaudan, K. Shaw, R. Ladner. *The 2-3 TR-tree, A Trajectory-Oriented Index Structure for Fully Evolving Valid-time Spatio-temporal Datasets*. In Proc. of the ACM workshop on Adv. in Geographic Info. Sys., ACM GIS, pages 29–34, Nov. 2002.

2D R-Tree	[Osborn, Barker, 2006] W. Osborn, K. Barker. *Searching through Spatial Relationships using the 2DR-tree*. The IASTED Conference on Internet and Multimedia Systems and Applications Honolulu, Hawaii, USA August 14-16, 2006
3D R-tree	[Theodoridis et al, 1996] Y. Theodoridis, M. Vazirgiannis, T. Sellis. *Spatio-Temporal Indexing for Large Multimedia Applications*. In Proc. of the IEEE Conference on Multimedia Computing and Systems, ICMCS, June 1996.
Adaptive K-D-Tree	[Bentley, Friedman, 1979] J. L. Bentley, J. H. Friedman. *Data structures for range searching*. ACM Comput. Surv. 11, 1979, 4, 397–409.
A-Tree (Approximation Tree)	[Sakurai et al, 2000] Y. Sakurai, M. Yoshikawa, S. Uemura, H. Kojima. *The a-tree: An index structure for high-dimensional spaces using relative approximation*. In VLDB, pages 516–526, 2000.
B+-tree	[Comer, 1979] D. Comer. *The ubiquitous B-tree*. ACM Comput. Surv. 11, 2, 1979, 121–138.
Balanced Multidimensional Extendible Hash Tree	[Otoo, 1985] E.J. Otoo. *Balanced multidimensional extendible hash tree*. In Proceedings of the fifth ACM SIGACT-SIGMOD symposium on Principles of database systems, Cambridge, Massachusetts, United States, 1985, Pages: 100 – 113
BANG File	[Freeston, 1987] M. Freeston. *The BANG file: A new kind of grid file*. In Proceedings of the ACM SIGMOD International Conference on Management of Data, 1987, pp. 260–269.
BD-Tree	[Ohsawa, Sakauchi, 1983] Y. Ohsawa, M. Sakauchi. *BD-tree: A new n-dimensional data structure with efficient dynamic characteristics*. In Proceedings of the Ninth World Computer Congress, IFIP 1983, 1983, pp. 539–544.
Bintree	[Tamminen, 1984] M. Tamminen. *Comment on quad- and octrees*. Commun. ACM 30, 3, 204–212. 1984
BIRCH	[Zhang et al, 1996] T. Zhang, R. Ramakrishnan, M. Livny. *BIRCH: an efficient data clustering method for very large databases*. pages 103–114, 1996.
Bkd-Tree	[Procopiuc et al, 2003] O. Procopiuc, P. K. Agarwal, L. Arge, J.-S. Vitter. *Bkd-tree: A Dynamic Scalable kd-tree*. In Proceedings of International Symposium on Spatial and Temporal Databases, 2003
B-link Tree	[Lehman, Yao, 1981] P. Lehman, S. Yao. *Efficient locking for concurrent operations on B-trees*. ACM Trans. Database Syst. 6, 4, 1981, 650–670.
Bottom-up Updates	[Lee et al, 2003] M. Lee, W. Hsu, C. Jensen, B. Cui, K. Teo. *Supporting Frequent Updates in R-Trees: A Bottom-Up Approach*. In Proc. of the Intl. Conf. on Very Large Data Bases, VLDB, Sept. 2003.
BSP-Tree	[Fuchs et al, 1980] H. Fuchs, Z. Kedem, B. Naylor. *On visible surface generation by a priori tree structures*. Computer Graph. 14, 3, 1980.
BST-Tree (Bisector Tree)	[Kalantari, McDonald, 1983] I. Kalantari, G. McDonald. *A data structure and an algorithm for the nearest point problem*. IEEE Trans. Software Eng., 9(5):631–634, 1983.
B-Tree	[Bayer, McCreicht, 1972] R. Bayer, E. M. McCreicht. *Organization and maintenance of large ordered indices*. Acta Inf. 1, 3, 1972, pp. 173–189.
BUB-Tree (Bounding UB Tree)	[Fenk, 2002] R. Fenk. *The BUB-Tree*. In Proceedings of VLDB Conf. Hongkong, 2002
Buddy Tree	[Seeger, Kriegel, 1990] B. Seeger, H.-P. Kriegel. *The buddy-tree: An efficient and robust access method for spatial data base systems*. In Proceedings of the Sixteenth International Conference on Very Large Data Bases, 1990, pp. 590–601.
Buddy Tree with Clipping	[Seeger, 1991] B. Seeger. *Performance comparison of segment access methods implemented on top of the buddy-tree*. In Advances in Spatial Databases, O. Gu¨ nther and H. Schek, Eds., LNCS 525, Springer-Verlag, Berlin/Heidelberg/New York, 1991, 277–296.

Buddy Tree with Overlapping	[Seeger, 1991] B. Seeger. *Performance comparison of segment access methods implemented on top of the buddy-tree*. In Advances in Spatial Databases, O. Gu¨ nther and H. Schek, Eds., LNCS 525, Springer-Verlag, Berlin/Heidelberg/New York, 1991, 277–296.
Buffer Tree	[Arge, 1995] L. Arge. *The buffer tree: a new technique for optimal I/O-algorithms*. In Proc.Workshop on Algorithms and Data Structures, pages 334–345. LNCS 955. Springer-Verlag, Berlin, 1995. [Arge, 2003] Lars Arge. *The Buffer Tree: A Technique for Designing Batched External Data Structures.* Algorithmica, Springer-Verlag New York Inc. 2003
BV Tree	[Freeston, 1995] M. Freeston. A *general solution of the n-dimensional B-tree problem*. In Proceedings of the ACM SIGMOD International Conference on Management of Data, 1995, pp. 80–91.
cCR-tree (Cache-Conscious R-Tree)	[Kim et al, 2001] K. Kim, S.K. Cha, K. Kwon. *Optimizing multidimensional index trees for main memory access*. International Conference on Management of Data Proceedings of the 2001 ACM SIGMOD international conference on Management of data, Santa Barbara, California, United States, 2001, Pp: 139 – 150
Cell Tree	[Günther, 1988] O. Günther. *Efficient Structures for Geometric Data Management*. LNCS 337, Springer-Verlag, Berlin/Heidelberg/New York. 1988.
Cell Tree with Oversize Shelves	[Günther, Noltemeier, 1991] O. Günther, H. Noltemeier. *Spatial database indices for large extended objects*. In Proceedings of the Seventh IEEE International Conference on Data Engineering, 1991, 520–526.
Circle Tree	[Moore, 2002] A.Moore. *The circle tree – a hierarchical structure for efficient storage, access and multi-scale representation of spatial data*. Presented at SIRC 2002 – The 14th Annual Colloquium of the Spatial Information Research Centre University of Otago, Dunedin, New Zealand, December 3-5th 2002
CLINDEX	[Li et al, 2002] C. Li, E. Chang, H. Garcia-Molina, G. Wiederhold. *Clustering for approximate similarity search in high-dimensional spaces*. IEEE Transactions on Knowledge and Data Engineering, 14(4):792–808, 2002.
cQSF Tree (Scalable QSF Tree)	[Orlandic, Yu, 2004] R. Orlandic, B. Yu. *Scalable QSF-Trees: Retrieving Regional Objects in High-Dimensional Spaces*. Journal of Database Management (JDM, IDEAS Group Publishing) Vol. 15☹in press), 15-page, 2004
CRB-Tree (Compressed Range B-Tree)	[Govindarajan et al, 2003] S. Govindarajan, P. K. Agarwal, L. Arge. *CRB-Tree: An Efficient Indexing Scheme for Range-Aggregate Queries*. Proceedings of the 9th International Conference on Database Theory, 2003, Pp:143-157
CR-Tree (Compressed Range Tree)	[Chazelle, 1988] B. Chazelle. *A functional approach to data structures and its use in multidimensional searching*. SIAM J. Comput., 17(3):427–462, June 1988
DBIN (Density Based Indexing)	[Bennett et al, 1999] K. P. Bennett, U. Fayyad, D. Geiger. *Density-based indexing for approximate nearestneighbor queries*. In KDD '99: Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining, pages 233–243, New York, NY, USA, 1999. ACM Press.
DOT	[Faloutsos, Rong, 1991] C. Faloutsos, Y. Rong. *DOT: A spatial access method using fractals*. In Proceedings of the Seventh IEEE International Conference on Data Engineering, 1991, pp. 152–159.
DP-Tree	[Li et al, 2006] M. Li, W.-C. Lee, A. Sivasubramaniam. *DPTree: A balanced tree based indexing framework for peer-to-peer systems*. in Proceedings of the 14th International Conference on Network Protocols (ICNP 2006), pages 12-21, November, 2006

Duality Transformation	[Kollios et al, 1999] G. Kollios, D. Gunopulos, V. J. Tsotras. On Indexing Mobile Objects. In Proc. of the ACM Symp. on Principles of Database Systems, PODS, pages 261–272, June 1999.
Duality Transformation with Kinetic Data Structure	[Agarwal et al, 2000] P. K. Agarwal, L. Arge, and J. Erickson. *Indexing Moving Points*. In Proc. of the ACM Symp. on Principles of Database Systems, PODS, pages 175–186, May 2000.
EXCELL (Extendible Cell)	[Tamminen, 1982] M. Tamminen. *The extendible cell method for closest point problems*. BIT 22, 1982, pp. 27–41.
Extended K-D-Tree	[Matsuyama et al, 1984] T. Matsuyama, L.V. Hao, M. Nagao. *A file organization for geographic information systems based on spatial proximity*. Int. J. Comput. Vis. Graph. Image Process. 26, 3, 1984, pp. 303–318.
Extendible Hashing	[Fagin et al, 1979] R. Fagin, J. Nievergelt, N. Pippenger, R. Strong. *Extendible hashing: A fast access method for dynamic files*. ACM Trans. Database Syst. 4, 3, 1979, pp. 315–344.
Fieldtree	[Frank, 1983] A. Frank. *Problems of Realizing LIS: Storage Methods for Space Related Data: The Field Tree*. Technical Report 71, Institut for Geodesy and Photogrammetry, Swiss Federal Institut of Technology, Zurich, Switzerland, 1983.
Filter Tree	[Sevcik, Koudas, 1996] K. Sevcik, N. Koudas. *Filter trees for managing spatial data over a range of size granularities*. In Proceedings of the 22th International Conference on Very Large Data Bases (Bombay), 1996, pp. 16–27.
Gauss-Tree	[Bohm et al, 2006] C. Bohm, A. Pryakhin, M. Schubert. *The Gauss-Tree: Efficient Object Identification in Databases of Probabilistic Feature Vectors*. 22nd Int. Conf. on Data Engineering (ICDE'06), Atlanta, GA, 2006
GBD-Tree	[Ohsawa, Sakauchi, 1990] Y. Ohsawa, M. Sakauchi. *A new tree type data structure with homogeneous node suitable for a very large spatial database*. In Proceedings of the Sixth IEEE International Conference on Data Engineering, 1990, pp. 296–303.
Generalized Grid File	[Blanken et al, 1990] H. Blanken, A. Ijbema, P. Meek, B. Van den Akker. *The generalized grid file: Description and performance aspects*. In Proceedings of the Sixth IEEE International Conference on Data Engineering, 1990, pp. 380–388.
GiST (Generalizad Search Tree)	[Hellerstein et al, 1995] J. M. Hellerstein, J. F. Naughton, A. Pfeffer. *Generalized Search Trees for Database Systems*. Proc. 21st Int. Conf. on Very Large Databases, September 1995, pp. 562-573.
GNAT (Geometric Near-Neighbor Access Tree)	[Brin, 1995] S. Brin. *Near neighbor search in large metric spaces*. In VLDB '95: Proceedings of the 21th International Conference on Very Large Data Bases, pages 574–584, San Francisco, CA, USA, 1995. Morgan Kaufmann Publishers Inc.
Grid File	[Nievergelt et al, 1981] J. Nievergelt, H. Hinterberger, K. Sevcik. *The grid file: An adaptable, symmetric multikey file structure*. In Proceedings of the Third ECI Conference, A. Duijvestijn and P. Lockemann, Eds., LNCS 123, Springer-Verlag, Berlin/Heidelberg/New York, 1981, pp. 236–251.
G-Tree	[Kumar, 1994] A. Kumar. *G-tree: A new data structure for organizing multidimensional data*. IEEE Trans. Knowl. Data Eng. 6, 2, 1994, pp. 341-347.
Hana Tree	[Kwon, Jeong, 2000] Y. Kwon, C. Jeong. *Hana Tree: A Dynamic and Robust Access Method for Spatial Data Handling*. Lecture Notes in Computer Science, Springer Berlin / Heidelberg, Volume 1846/2000. Proceedings of Web-Age Information Management: First International Conference, WAIM 2000, Shanghai, China, June 21-23, 2000.
Hashing Technique	[Song, Roussopoulos, 2001] Z. Song, N. Roussopoulos. *Hashing Moving Objects*. In Mobile Data Management, pages 161–172, Jan. 2001.

hBP-Tree	[Evangelidis et al, 1995] G. Evangelidis, D. Lomet, B. Salzberg. *The hBP-tree: A modified hB-tree supporting concurrency, recovery and node consolidation*. In Proceedings of the 21st International Conference on Very Large Data Bases, 1995, pp. 551–561.
hB-Tree	[Lomet, Salzberg, 1989] D.B. Lomet, B. Salzberg. *The hBtree: A robust multiattribute search structure*. In Proceedings of the Fifth IEEE International Conference on Data Engineering, 1989, pp. 296–304.
Hilbert R-Tree	[Kamel, Faloutsos, 1994] I. Kamel, C. Faloutsos. *Hilbert R-tree: An improved R-tree using fractals*. In Proceedings of the Twentieth International Conference on Very Large Data Bases, 1994, pp. 500–509.
HR+-Tree	[Tao, Papadias, 2001a] Y. Tao, D. Papadias. *Efficient Historical R-trees*. In Proc. of the Intl. Conf. on Scientific and Statistical Database Management, SSDBM, pages 223–232, July 2001.
HR-Tree (Historical R-Tree)	[Nascimento, Silva, 1998] M. A. Nascimento, J.R.O. Silva. *Towards historical R-trees*. In Proc. of the ACM Symp. on Applied Computing, SAC, pages 235–240, Feb. 1998.
Interpolation-Based Grid File	[Ouksel, 1985] M. Ouksel. *The interpolation based grid file*. In Proceedings of the Fourth ACM SIGACT –SIGMOD Symposium on Principles of Database Systems, 1985, pp. 20–27.
Interval Tree	[Edelsbrunner 1980] H. Edelsbrunner. *Dynamic Rectangle Intersection Searching*. Institute for Information. Processing Report 47, Technical University of Graz, Austria, 1980.
IQ-Tree (Independent Quantization Tree)	[Berchtold et al, 2000] S. Berchtold, C. Bohm, H. V. Jagadish, H.-P. Kriegel, J. Sander. *Independent quantization: An index compression technique for high-dimensional data spaces*. In ICDE '00: Proceedings of the 16th International Conference on Data Engineering, page 577, Washington, DC, USA, 2000. IEEE Computer Society.
KD2B-Tree	[Oosterom, 1990] P. Oosterom. *Reactive data structures for geographic information systems*. Ph.D. Thesis, University of Leiden, The Netherlands. 1990.
KDB_{FD}-Tree KDB_{HD}-Tree	[Orlandic, Yu, 2002] R. Orlandic, B. Yu. *A retrieval technique for high-dimensional data and partially specified queries*. Data & Knowledge Engineering 2002; 42(2):1-21.
KDB_{KD}-Tree	[Yu et al, 2003] B. Yu, R. Orlandic, T. Bailey, J. Somavaram. KDBKD-Tree: A Compact KDB-Tree Structure for Indexing Multidimensional Data. International Conference on Information Technology: Computers and Communications, 2003.
K-D-B-Tree	[Robinson, 1981] J.T. Robinson. *The K-D-B-tree: A search structure for large multidimensional dynamic indexes*. In Proceedings of the ACM SIGMOD International Conference on Management of Data, 1981, pp. 10-18.
K-D-Tree	[Bentley, 1975] J. L. Bentley. *Multidimensional binary search trees used for associative searching*. Commun. ACM 18, 9, 1975, pp. 509–517.
kNR-Tree	[Mondal et al, 2005] A. Mondal, A. K. H. Tung, M. Kitsuregawa. *kNR-tree: A novel R-tree-based index for facilitating Spatial Window Queries on any k relations among N spatial relations in Mobile environments.* MDM 2005 05 Ayia Napa Cyprus, 2005
Level Balaced B-Tree	[Agarwal et al, 1999] P.K. Agarwal, L. Arge, G.S. Brodal, J.S. Vitter. *I/O-efficient.dynamic point location in monotone planar subdivisions*. In Proc. ACM-SIAM Symp. on Discrete Algorithms, 1999, pp.1116-1127
Linear Hashing	[Litwin, 1980] W. Litwin. *Linear hashing: A new tool for file and table addressing*. In Proceedings of the Sixth International Conference on Very Large Data Bases, 1980, pp. 212–223. [Larson, 1980] P. A. Larson. *Linear hashing with partial expansions*. In Proceedings of the Sixth International Conference on Very Large Data Bases, 1980, pp. 224–232.
Linear Quadtree	[Samet, 1990] H. Samet. *Applications of Spatial Data Structures*. Addison-Wesley, Reading, MA. 1990.

LPC-File (Local Polar Coordinate File)	[Cha et al, 2002] G.-H. Cha, X. Zhu, D. Petkovic, C.-W. Chung. *An efficient indexing method for nearest neighbor searches inhigh-dirnensional image databases*. IEEE Transactions on Multimedia, 4(1):76–87, 2002.
LSD-Tree	[Henrich et al, 1989] A. Henrich, H.-W. Six, P. Widmayer. *The LSD tree: Spatial access to multidimensional point and non-point objects*. In Proceedings of the Fifteenth International Conference on Very Large Data Bases, 1989, pp. 45–53.
LSH (Locality Sensitive Hashing)	[Gionis et al, 1999] A. Gionis, P. Indyk, R. Motwani. *Similarity search in high dimensions via hashing*. In VLDB '99: Proceedings of the 25th International Conference on Very Large Data Bases, pages 518–529, San Francisco, CA, USA, 1999. Morgan Kaufmann Publishers Inc. [Indyk, Motwani, 1998] P. Indyk, R. Motwani. *Approximate nearest neighbors: towards removing the curse of dimensionality*. In STOC '98: Proceedings of the thirtieth annual ACM symposium on Theory of computing, pages 604–613, New York, NY, USA, 1998. ACM Press.
LUR-Tree (Lazy Update R-Tree)	[Kwon et al, 2002] D. Kwon, S. Lee, S. Lee. *Indexing the Current Positions of Moving Objects Using the Lazy Update R-tree*. In Mobile Data Management, MDM, pages 113–120, Jan. 2002.
lz-Hashing	[Hutflesz et al, 1991] A. Hutflesz, P. Widmayer, C. Zimmermann. *Global order makes spatial access faster*. In Geographic Database Management Systems, G. Gambosi, M. Scholl, and H.-W. Six, Eds., Springer-Verlag, Berlin/Heidelberg/ New York, 1991, pp. 161–176.
MB+ Tree	[Yang et al, 1995] Q. Yang, A. Vellaikal, S. Dao. *MB+-Tree: A New Index Structure for Multimedia Databases*. Proceedings of the International Workshop on Multimedia Database Management Systems, August, 1995, pp. 151-158.
MDAM, ArM	[Markov, 1984] Kr. Markov. *A Multi-domain Access Method.* Proceedings of the International Conference on Computer Based Scientific Research. Plovdiv, 1984, pp.558-563. [Markov, 2004] Kr. Markov. *Multi-Domain Information Model*. International Journal "Information Theories and Applications", ISSN 1310-0513. Vol. 11, No: 4, 2004, pp. 303-308
MedRank (Median Rank)	[Fagin et al, 2003] R. Fagin, R. Kumar, D. Sivakumar. *Efficient similarity search and classification via rank aggregation*. In SIGMOD '03: Proceedings of the 2003 ACM SIGMOD international conference on Management of data, pages 301–312, New York, NY, USA, 2003. ACM Press.
MOLPHE	[Kriegel, Seeger, 1986] H.-P. Kriegel, B. Seeger. *Multidimensional order preserving linear hashing with partial expansions*. In Proceedings of the International Conference on Database Theory, LNCS 243, Springer-Verlag, Berlin/Heidelberg/New York. 1986.
mQp-Tree	[Salas, Polo, 2000] M. Salas, A. Polo. *The mQp-tree: a multidimensional access method based on a non-binary tree*. Proc of the VIIth Conference on Extending Database Technology, March 2000, Konstanz - Germany
MR-tree	[Xu et al, 1990] X. Xu, J. Han, W. Lu. *RT-Tree: An Improved R-Tree Indexing Structure for Temporal Spatial Databases*. In Proc. of the Intl. Symp. on Spatial Data Handling, SDH, pages 1040–1049, July 1990.
M-Tree (Metric Tree)	[Ciaccia et al, 1997] P. Ciaccia, M. Patella, P. Zezula. *M-tree: An efficient access method for similarity search in metric spaces*. In VLDB '97: Proceedings of the 23rd International Conference on Very Large Data Bases, pages 426–435, San Francisco, CA, USA, 1997. Morgan Kaufmann Publishers Inc.
Multi-Layer Grid File	[Six, Widmayer, 1988] H. Six, P. Widmayer. *Spatial searching in geometric databases*. In Proceedings of the Fourth IEEE International Conference on Data Engineering, 1988, pp. 496–503.
Multi-Level Grid File	[Whang, Krishnamurthy, 1985] K.-Y. Whang, R. Krishnamurthy. *Multilevel grid files*. IBM Research Laboratory, Yorktown Heights, NY. 1985.

MV3R-Tree	[Tao, Papadias, 2001b] Y. Tao, D. Papadias. *MV3R-Tree: A Spatio-Temporal Access Method for Timestamp and Interval Queries*. In Proc. of the Intl. Conf. on Very Large Data Bases, VLDB, pages 431–440, Sept. 2001.
MVB-Tree (Multi Version B-Tree)	[Becker et al, 1996] B. Becker, S. Gschwind, T. Ohler, B. Seeger, P. Widmayer. An *Asymptotically Optimal Multiversion B-Tree*. VLDB Journal, 5(4):264–275, 1996.
MVP Tree	[Bozkaya, Ozsoyoglu, 1997] T. Bozkaya, M. Ozsoyoglu. *Distance-Based Indexing for High-Dimensional Metric Spaces*. Proceedings of the 1997 ACM SIGMOD International Conference on Management of Data, May 1997, pp. 357-368.
MVSB-Tree (Multiversion SB-tree)	[Markowetz et al, 2001] D. Zhang, A. Markowetz, V. Tsotras, D. Gunopulos, B. Seeger. *Efficient computation of temporal aggregates with range predicates*. In Proc. Principles Of Database Systems, pages 237–245, 2001.
NA-Tree (Nine Areas Tree)	[Chang et al, 2003] Y.-I. Chang, C.–H. Liao, H.-L. Chen. *NA-Trees: A Dynamic Index for Spatial Data*. Journal of Information Science and Engineering, 19, 103-139, 2003.
Nested Interpolation Based Grid File	[Ouksel, Mayer, 1992] M.A. Ouksel, O. Mayer. *A robust and efficient spatial data structure*. Acta Inf. 29, 1992, pp. 335–373.
NSI-Tree	[Porkaew et al, 2001] K. Porkaew, I. Lazaridis, S. Mehrotra. *Querying Mobile Objects in Spatio-Temporal Databases*. In Proc. of the Intl. Symp. on Advances in Spatial and Temporal Databases, SSTD, pages 59–78, Redondo Beach, CA, July 2001.
Overlapping Linear Quadtree	[Tzouramanis et al, 1998] T. Tzouramanis, M. Vassilakopoulos, Y. Manolopoulos. *Overlapping Linear Quadtrees: A Spatio-Temporal Access Method*. In Proc. of the ACM workshop on Adv. in Geographic Info. Sys., ACM GIS, pages 1–7, Nov. 1998.
P+-Tree	[Zhang et al, 2004] R. Zhang, B. C. Ooi, K.-L. Tan. *Making the pyramid technique robust to query types and workloads*. In ICDE'04 : Proceedings of the Twenth International Conference on Data Engineering, pages 313–324, Washington, DC, USA, 2004. IEEE Computer Society.
Packed R-Tree	[Roussopoulos, Leifker, 1985] N. Roussopoulos, D. Leifker. *Direct spatial search on pictorial databases using packed R-trees*. In Proceedings of the ACM SIGMOD International Conference on Management of Data, 1985, pp. 17–31.
Parallel R-Tree	[Kamel, Faloutsos, 1992] I. Kamel, C. Faloutsos. *Parallel R-trees*. In Proceedings of the ACM SIGMOD International Conference on Management of Data, 1992, pp. 195–204.
Partitioned B-Tree	[Graefe, 2003] G. Graefe. *Sorting and Indexing with Partitioned B-Trees*. Conference on Innovative Data Systems Research, 2003.
PCURE (Paralel Implementation of Clustering Using Representatives)	[Berrani et al, 2003] S.-A. Berrani, L. Amsaleg, P. Gros. *Approximate searches: k-neighbors + precision*. In CIKM '03: Proceedings of the twelfth international conference on Information and knowledge management, pages 24–31, New York, NY, USA, 2003. ACM Press.
Persistent B-tree	[Sarnak, Tarjan, 1986] N. Sarnak, R.E. Tarjan. *Planar point location using persistent search trees*. Communication of the ACM, 1986, 29:669-679.
PLOP-Hashing	[Kriegel, Seeger, 1988] H.-P. Kriegel, B. Seeger. *PLOP-hashing: A grid file without directory*. In Proceedings of the Fourth IEEE International Conference on Data Engineering, 1988, pp. 369–376.
PM Quadtree	[Samet, Webber, 1985] H. Samet, R. E. Webber. *Storing a collection of polygons using quadtrees*. ACM Trans. Graph. 4, 3, 1985, 182–222.
PMR-Quadtree	[Nelson, Samet, 1986] R.C. Nelson, H. Samet. *A Consistent Hierarchical Representation for Vector Data*. In Proc. of the ACM SIGGRAPH, pages 197–206, Aug. 1986.
PMR-Quadtree for moving objects	[Tayeb et al, 1998] J. Tayeb, O . Ulusoy, O.Wolfson. *A Quadtree-Based Dynamic Attribute Indexing Method*. The Computer Journal, 41(3):185–200, 1998.

Point Quadtree	[Klinger, 1971] A. Klinger. *Pattern and search statistics*. In Optimizing Methods in Statistics, S.Rustagi, Ed., 1971, pp. 303–337.
PPR-Tree	[Kumar et al, 1998] A. Kumar, V. J. Tsotras, and C. Faloutsos. *Designing Access Methods for Bitemporal Databases*. IEEE Trans. on Knowledge and Data Engineering, TKDE, 10(1):1–20, 1998.
PPR-Tree with Linear Greedy	[Kollios et al, 2001] G. Kollios, V. J. Tsotras, D. Gunopulos, A. Delis, M. Hadjieleftheriou. *Indexing Animated Objects Using Spatiotemporal Access Methods*. IEEE Trans. on Knowledge and Data Engineering, TKDE, 13(5):758–777, 2001.
PPR-Tree with Polynomial Greedy	[Hadjieleftheriou et al, 2002] M. Hadjieleftheriou, G. Kollios, V. J. Tsotras, D. Gunopulos. *Efficient Indexing of Spatiotemporal Objects*. In Proc. of the Intl. Conf. on Extending Database Technology, EDBT, pages 251–268, Czech Republic, Mar. 2002.
Prefix Hash Tree	[Ramabhadran et al, 2004] S. Ramabhadran, S. Ratnasamy, J.M. Hellerstein, S. Shenker. *Prefix Hash Tree: An Indexing Data Structure over Distributed Hash Tables*. Submitted to PODC 2004
PR-Tree	[Cai, Revesz, 2000] M. Cai, P. Revesz. *Parametric R-Tree: An Index Structure for Moving Objects*. In Proc. of the Intl. Conf. on Management of Data, COMAD, Dec. 2000.
PSI (Parametric Space Indexing)	[Porkaew et al, 2001] K. Porkaew, I. Lazaridis, S. Mehrotra. *Querying Mobile Objects in Spatio-Temporal Databases*. In Proc. of the Intl. Symp. on Advances in Spatial and Temporal Databases, SSTD, pages 59–78, Redondo Beach, CA, July 2001.
P-Tree (J)	[Jagadish, 1990] H.V. Jagadish. *Spatial search with polyhedra*. In Proceedings of the Sixth IEEE International Conference on Data Engineering, 1990, pp. 311–319.
P-Tree (S)	[Schiwietz, 1993] M. Schiwietz. *Speicherung und anfragebearbeitung komplexer geo-objekte*. Ph.D. Thesis, Ludwig-Maximilians-Universitat Munchen, Germany (in German). 1993.
PvS Index	[Lejsek et al, 2005] H. Lejsek, F. H. Asmundsson, B. P. Jonsson, L. Amsaleg. *Efficient and effective image copyright enforcement*, Technical Report. Reykjavík University, 2005.
Pyramid Technique	[Berchtold et al, 1998] S. Berchtold, C. Bohm, H.-P. Kriegel. *The pyramid-technique: Towards breaking the curse of dimensionality*. In L. M. Haas and A. Tiwary, editors, SIGMOD 1998, Proceedings ACM SIGMOD International Conference on Management of Data, June 2-4, 1998, Seattle, Washington, USA, pages 142–153. ACM Press, 1998.
Q+R-Tree	[Yuni, Prabhakar, 2003] X. Yuni, S. Prabhakar. *Q+Rtree: efficient indexing for moving object databases*. Database Systems for Advanced Applications, 2003. (DASFAA 2003). Proceedings. Eighth International Conference on Publication Date: 26-28 March 2003. Pp: 175- 182
Quantile Hashing	[Kriegel, Seeger, 1987] H.-P. Kriegel, B. Seeger. *Multidimensional quantile hashing is very efficient for non-uniform record distributions*. In Proceedings of the Third IEEE International Conference on Data Engineering, 1987, pp. 10–17.
R*-Tree	[Beckmann et al, 1990] N. Beckmann, H.-P. Kriegel, R. Schneider, B. Seeger. *The R*-tree: An efficient and robust access method for points and rectangles*. In Proceedings of ACM SIGMOD International Conference on Management of Data, 1990, pp. 322–331.
R+-Tree	[Sellis et al, 1987] T. Sellis, N. Roussopoulos, C. Faloutsos. *The R+-tree: A dynamic index for multi-dimensional objects*. In Proceedings of the Thirteenth International Conference on Very Large Data Bases, 1987, pp. 507–518.
Reactive Tree	[Oosterom, 1993] P. van Oosterom. *Reactive Data Structures for Geographic Information Systems*. Oxford University Press, Oxford. 1993.
Region Quadtree	[Finkel, Bentley, 1974] R. Finkel, J. L. Bentley. *Quadtrees: A data structure for retrieval of composite keys*. Acta Inf. 4, 1, 1974, pp. 1–9.
Relational Interval Tree	[Kriegel et al, 2000] H.-P. Kriegel, M. Pötke, T. Seidl. *Managing Intervals Efficiently in Object-Relational Databases*. Proc. 26th Int. Conf. on Very Large Databases (VLDB): 407-418, 2000.

Relational R-Tree	[Ravi et al, 1999] K.V. Ravi Kanth, S. Ravada, J. Sharma, J. Banerjee. *Indexing Medium-dimensionality Data in Oracle*. Proc. ACM SIGMOD Int. Conf. on Management of Data: 521-522, 1999.
Relational X-Tree	[Berchtold et al, 1999] S. Berchtold, C. Böhm, H.-P. Kriegel, U. Michel. *Implementation of Multidimensional Index Structures for Knowledge Discovery in Relational Databases*. Proc. 1st Int. Conf. on Data Warehousing and Knowledge Discovery (DaWaK), LNCS 1676: 261-270, 1999.
R^{EXP}-Tree	[Saltenis, Jensen, 2002] S. Saltenis, C. S. Jensen. *Indexing of Moving Objects for Location-Based Services*. In Proc. of the Intl. Conf. on Data Engineering, ICDE, Feb. 2002.
R-File	[Hutflesz et al, 1990] A. Hutflesz, H.-W. Six, P. Widmayer. *The R-file: An efficient access structure for proximity queries*. In Proceedings of the Sixth IEEE International Conference on Data Engineering, 1990, pp. 372–379.
R-link Tree	[Ng, Kameda, 1994] V. Ng, T. Kameda. *The R-link tree: A recoverable index structure for spatial data*. In Proceedings of the Fifth Conference on Database and Expert Systems Applications (DEXA'94), D. Karagiannis, Ed., LNCS 856, Springer-Verlag, Berlin/Heidelberg/New York, 1994, 163–172.
R-Tree	[Guttman, 1984] A. Guttman. *R-trees: A dynamic index structure for spatial searching*. In Proceedings of the ACM SIGMOD International Conference on Management of Data, 1984, 47–54.
RT-Tree	[Xu et al, 1990] X. Xu, J. Han, W. Lu. *RT-Tree: An Improved R-Tree Indexing Structure for Temporal Spatial Databases*. In Proc. of the Intl. Symp. on Spatial Data Handling, SDH, pages 1040–1049, July 1990.
SB-Tree (Segment B-Tree)	[Yang, Widom, 2001] J. Yang, J. Widom. *Incremental Computation and Maintenance of Temporal Aggregates*. Proceedings of the 17th International Conference on Data Engineering, 2001, Pages: 51 - 60
SEB-Tree (Start/End Timestamp B-Tree)	[Song, Roussopoulos, 2003] Z. Song, N. Roussopoulos. *SEB-tree: An Approach to Index Continuously Moving Objects*. In Mobile Data Management, MDM, pages 340–344, Jan. 2003.
Segment Indexes	[Kolovson, Stonebraker, 1991] C. Kolovson, M. Stonebraker. *Segment indexes: Dynamic indexing techniques for multi-dimensional interval data*. In Proceedings of the ACM SIGMOD International Conference on Management of Data, 1991, pp. 138–147.
SETI (Scalable and Efficient Trajectory Index)	[Chakka et al, 2003] V. P. Chakka, A. Everspaugh, J. M. Patel. *Indexing Large Trajectory Data Sets with SETI*. In Proc. of the Conf. on Innovative Data Systems Research, CIDR, Asilomar, CA, Jan. 2003.
SH-Tree (Super Hybrid Tree)	[Dang, 2006] T.K. Dang. *The SH-Tree: A Novel and Flexible Super Hybrid Index Structure for Similarity Search on Multidimensional Data*. International Journal of Computer Science & Applications, 2006 Technomathematics Research Foundation Vol. III, No. I, pp. 1 - 25
SKD-Tree	[Ooi et al, 1987] B.C. Ooi, K.J. McDonell, R. Sacks-Davis. *Spatial kd-tree: An indexing mechanism for spatial databases*. In Proceedings of the IEEE Computer Software and Applications Conference, 1987, pp. 433–438.
Slim Tree	[Traina Jr. et al, 2000] C. Traina Jr., A. Traina, B. Seeger, C. Faloutsos. *Slim-trees: High Performance Metric Trees Minimizing Overlap Between Nodes*. International Conference on Extending Database Technology (EDBT) 2000, Konstanz, Germany, March 27-31, 2000.
Space Filling Curves	[Morton, 1966] G. Morton. *A computer oriented geodetic data base and a new technique in file sequencing*. IBM Ltd. 1966.
Sphere Tree	[Oosterom, 1990] P. Oosterom. *Reactive data structures for geographic information systems*. Ph.D. Thesis, University of Leiden, The Netherlands. 1990.

sQSF-Tree (Simple QSF-Tree)	[Yu et al, 1999] B. Yu, R. Orlandic, M. Evens. *Simple QSF-trees: an efficient and scalable spatial access method*. Proceedings of the Eighth International Conference on Information and Knowledge Management, p.5-14, November 02-06, 1999, Kansas City, Missouri, United States
SR Tree (Sphere-Rectangle Tree)	[Katayama, S. Satoh, 1997] N. Katayama, S. Satoh. *The sr-tree: an index structure for high-dimensional nearest neighbor queries*. In SIGMOD '97: Proceedings of the 1997 ACM SIGMOD international conference on Management of data, pages 369–380, New York, NY, USA, 1997. ACM Press.
SS Tree (Similarity Search Tree)	[White, Jain, 1996] D. A. White, R. Jain. *Similarity indexing with the ss-tree*. In ICDE '96: Proceedings of the Twelfth International Conference on Data Engineering, pages 516–523, Washington, DC, USA, 1996. IEEE Computer Society.
STAR-Tree (Spatio-temporal Self Adjusting R-Tree)	[Procopiuc et al, 2002] C. M. Procopiuc, P. K. Agarwal, S. Har-Peled. *STAR-Tree: An Efficient Self-Adjusting Index for Moving Objects*. In Proc. of the Workshop on Alg. Eng. and Experimentation, ALENEX, pages 178–193, Jan. 2002.
String B-tree	[Ferragina, Grossi, 1995] P. Ferragina, R. Grossi. *A fully-dynamic data structure for external substring search*. In Proc. ACM Symp. on Theory of Computation, 1995, pp.693-702.
STR-tree	[Pfoser et al, 2000] D. Pfoser, C. S. Jensen, and Y. Theodoridis. *Novel Approaches in Query Processing for Moving Object Trajectories*. In Proc. of the Intl. Conf. on Very Large Data Bases, VLDB, pages 395–406, Sept. 2000.
SV-Model	[Chon et al, 2001] H. D. Chon, D. Agrawal, A. E. Abbadi. *Storage and Retrieval of Moving Objects*. In Mobile Data Management, pages 173–184, Jan. 2001.
TB-Tree (Trajectory-Bundle Tree)	[Pfoser et al, 2000] D. Pfoser, C. S. Jensen, and Y. Theodoridis. *Novel Approaches in Query Processing for Moving Object Trajectories*. In Proc. of the Intl. Conf. on Very Large Data Bases, VLDB, pages 395–406, Sept. 2000.
TPR*-Tree	[Tao et al, 2003] Y. Tao, D. Papadias, J. Sun. *The TPR*-Tree: An Optimized Spatio-Temporal Access Method for Predictive Queries*. Proceedings of the 29th VLDB Conference, Berlin, Germany, 2003
TPR-tree (Time Parameterized R-Tree)	[Saltenis et al, 2000] S. Saltenis, C. S. Jensen, S. T. Leutenegger, M. A. Lopez. *Indexing the Positions of Continuously Moving Objects*. In Proc. of the ACM Intl. Conf. on Management of Data, SIGMOD, pages 331–342, May 2000.
TR*-Tree	[Schneider, Kriegel, 1992] R. Schneider, H.-P. Kriegel. *The TR*-tree: A new representation of polygonal objects supporting spatial queries and operations*. In Proceedings of the Seventh Workshop on Computational Geometry, LNCS 553, Springer-Verlag, Berlin/Heidelberg/New York, 1992, pp. 249–264.
TR-Tree (Temporal R-Tree)	[Almeida et al, 1999] V. T. Almeida, J. M. Souza, G. Zimbrao. *The Temporal R-Tree*. Technical Report No. ES-492/99, COPPE Sistemas/UFRJ, 1999.
TSB-Tree	[Lomet, Salzberg, 1989] D.B. Lomet, B. Salzberg. *Access Methods for Multiversion Data*. In Proc. of the ACM Intl. Conf. on Management of Data, SIGMOD, pages 315–324, May 1989.
TV-Tree (Telescoping Vector Tree)	[Lin et al, 1994] K. I. Lin, H. V. Jagadish, C. Faloutsos. *The tv-tree: an index structure for high-dimensional data*. The VLDB Journal, 3(4):517–542, 1994.
Twin Grid File	[Hutflesz et al, 1988b] A. Hutflesz, H.-W. Six, P. Widmayer. *Twin grid files: Space optimizing access schemes*. In Proceedings of the ACM SIGMOD International Conference on Management of Data, 1988, pp. 183–190.
Two-Level Grid File	[Hinrichs, 1985] K. Hinrichs. *Implementation of the grid file: Design concepts and experience*. BIT 25, 1985, pp. 569–592.
UB-Tree	[Bayer, 1996] R. Bayer. *The universal B-tree for multidimensional indexing*. Tech. Rep. I9639, Technische Universitat Munchen, Munich, Germany. 1996.

VA+-File	[Ferhatosmanoglu et al, 2000] H. Ferhatosmanoglu, E. Tuncel, D. Agrawal, A. E. Abbadi. *Vector approximation based indexing for non-uniform high dimensional data sets*. In CIKM '00: Proceedings of the ninth international conference on Information and knowledge management, pages 202–209, New York, NY, USA, 2000. ACM Press.
VA-File (Vector Approximation File)	[Weber et al, 2000] R. Weber, K. Bohm, H.-J. Schek. Interactive-time similarity search for large image collections using parallel va-files. In ICDE, page 197, 2000.
VCI R-Tree (Velocity Constrained Indexing R-Tree)	[Prabhakar et al, 2002] S. Prabhakar, Y. Xia, D. V. Kalashnikov, W. G. Aref, S. E. Hambrusch. *Query Indexing and Velocity Constrained Indexing: Scalable Techniques for Continuous Queries on Moving Objects*. IEEE Transactions on Computers, 51(10):1124–1140, 2002.
VP Tree (Vantage Point Tree)	[Yianilos, 1993] P. N. Yianilos. *Data structures and algorithms for nearest neighbor search in general metric spaces*. In SODA '93: Proceedings of the fourth annual ACM-SIAM Symposium on Discrete algorithms, pages 311–321, Philadelphia, PA, USA, 1993. Society for Industrial and Applied Mathematics.
V-Reactive Tree	[Li et al, 2001] J. Li, N. Jing, M. Sun. *Spatial Database Techniques Oriented to Visualization in 3D GIS*. In Proceedings of the 2nd International Symposium on Digital Earth. 2001.
Weight-balanced B-tree	[Arge, Vitter, 1996] L. Arge, J. S.Vitter. *Optimal Dynamic Interval Management in External Memory*. Proc. 37th Annual Symp. on Foundations of Computer Science: 560-569, 1996.
X-Tree	[Berchtold et al, 1996] S. Berchtold, D. Keim, H.-P. Kriegel. *The X-tree: An index structure for high-dimensional data*. In Proceedings of the 22nd International Conference on Very Large Data Bases, (Bombay) 1996, pp. 28–39.
Z-Hashing	[Hutflesz et al, 1988a] A. Hutflesz, H.-W. Six, P. Widmayer. *Globally order preserving multidimensional linear hashing*. In Proceedings of the Fourth IEEE International Conference on Data Engineering, 1988, pp. 572–579.
zkdB+tree	[Orenstein, 1986] J. A. Orenstein. *Spatial query processing in an object-oriented database system*. In Proceedings of the ACM SIGMOD International Conference on Management of Data, 1986, 326–333.
Z-Ordering	[Orenstein, Merrett, 1984] J. Orenstein, T.H. Merrett. *A class of data structures for associative searching*. In Proceedings of the Third ACM SIGACT–SIGMOD Symposium on Principles of Database Systems, 1984, pp. 181–190.

Additional bibliography

[Arge, 2002] L. Arge. *External Memory Data Structures*. Part 4, chapter 9 in Handbook of Massive Datasets. J. Abello, P.M. Pardalos, M.G.C. Resende (eds), Kluwer Academic Publishers, 2002, pp. 313-357.

[Bayer, 1996] R. Bayer. *The Universal B-tree for Multidimensional Indexing*. Techical Report I9639, Technische Universitat Munchen, Munich, Germany. 1996.

[Bellman 1961] R. Bellman. *Adaptive Control Processes: A Guided Tour*. Princeton University Press., 1961

[Bouteldja et al, 2006] N. Bouteldja, V. Gouet-Brunet, M. Scholl. *Back to the Curse of Dimensionality with Local Image Descriptors.* CEDRIC Research Report no 1049. July 20, 2006.

[Chakrabarti, 2001] K. Chakrabarti. *Managing Large Multidimensional Datasets Inside a Database System*. Phd Thesis, University of Illinois at Urbana-Champaign. Urbana, Illinois, 2001

[Chavez et al, 2001] E. Chavez, G. Navarro, R. Baeza-Yates, J. Marroquin. *Searching in Metric Spaces*. ACM Computing Surveys, 33(3):273–321, Sept. 2001.

[CODASYL, 1971] Codasyl Systems Committee. *Feature Analysis of Generalized Data Base Management Systems*. Technical Report, May, 1971 / Информационные системы общего предназначения (Аналитический обзор систем управления базами данных). Москва, Статистика, 1975.

[Connolly, Begg, 2002] T.M. Connolly, C.E.Begg. *Database Systems. A Practical Approach to Design, Implementation, and Managemen*t. Third Edition. Addison-Wesley Longman, Inc. – Pearson Education Ltd., 1995, 2002 / Т.Коннолли, К.Бегг. *Базы данных. Проектирование, реализация и сопровождение*. Теория и практика. Москва-Санкт Петербург-Киев, Издателский дом "Вильямс", 2003. 1440 с.

[Date, 1977] C.J. Date. *An Introduction to Database Systems*. Addison-Wesley Inc. 1975. / К.Дейт. *Введение в системы баз данных*. Москва, Наука, 1980.

[Gaede, Günther, 1998] V. Gaede, O. Günther. *Multidimensional Access Methods*. ACM Computing Surveys, Vol. 30, No. 2, June 1998.

[IBM, 1965-68] *IBM System/360, Disk Operating System, Data Management Concepts*. IBM System Reference Library, IBM Corp. 1965, Major Revision, February 1968.

[Markov, 2006] Kr. Markov. *Multidimensional Context-Free Access Method*. PhD Thesis. Institute of Mathematic and Informatics, Bulgarian Academy of Sciences. Sofia, 2006, 130p.

[Markov, 2004] Kr. Markov. *Multi-Domain Information Model*. International Journal "Information Theories and Applications", ISSN 1310-0513. Vol. 11, No: 4, 2004, pp. 303-308

[Martin, 1975] J.Martin. *Computer Data-Base Organization*. Prentice-Hall, Inc., Englewood Cliffs, New Jersey / Дж. Мартин. *Организация баз данных в вычислительных системах*. Москва, Мир, 1978.

[Moënne-Loccoz, 2005] N. Moënne-Loccoz. *High-Dimensional Access Methods for Efficient Similarity Queries*. Technical Report N:0505, University of Geneva, Computer Vision and Multimedia Laboratory, May 2005.

[Mokbel et al, 2003] M. F. Mokbel, T. M. Ghanem, W. G. Aref. *Spatio-temporal Access Methods*. IEEE Data Engineering Bulletin, 26(2), 40-49, June, 2003.

[Ooi et al, 1993] B.C. Ooi, R. Sacks-Davis, J. Han. *Indexing in Spatial Databases*. Technical Report. 1993.

[Schlosser et al, 2005] S.W. Schlosser, J. Schindler, S. Papadomanolakis, M. Shao, A. Ailamaki, C. Faloutsos, G.R. Ganger. *On Multidimensional Data and Modern Disks*. Proceedings of the 4th USENIX Conference on File and Storage Technology (FAST '05). San Francisco, CA. December 13-16, 2005.

[Stably, 1970] D. Stably. *Logical Programming with System/360*. New York, 1970 / Д.Стэбли. *Логическое программирование в системе/360*. Москва, Мир, 1974.

Authors' Information

Krassimir Markov *- Institute of Mathematics and Informatics, BAS, Acad.G.Bonthev St., bl.8, Sofia-1113, Bulgaria; Institute of Information Theories and Applications FOI ITHEA, P.O.Box: 775, Sofia-1090, Bulgaria; e-mail: markov@foibg.com*

Krassimira Ivanova *- Institute of Mathematics and Informatics, BAS, Acad.G.Bonthev St., bl.8, Sofia-1113, Bulgaria; e-mail: kivanova@math.bas.bg*

Ilia Mitov *- Institute of Information Theories and Applications FOI ITHEA, P.O.Box: 775, Sofia-1090, Bulgaria; e-mail: mitov@foibg.com*

Stefan Karastanev *- Institute of Mechanics and Biomechanics, BAS, Acad.G.Bonthev St., bl.4, Sofia-1113, Bulgaria; e-mail: stefan@info.imbm.bas.bg*

PRINCIPLES OF INTEGRATION OF RUSSIAN AND JAPANESE DATABASES ON INORGANIC MATERIALS

Nadezhda Kiselyova, Shuichi Iwata, Victor Dudarev, Ilya Prokoshev, Valentin Khorbenko, Victor Zemskov

***Abstract**: The methods and software for integration of databases (DBs) on inorganic material and substance properties were developed. The integration of information systems is based on combination of known approaches: EII (Enterprise Information Integration) and EAI (Enterprise Application Integration). The metabase - special database that stores data on contents of integrated DBs is a kernel of integrated system. Proposed methods were applied for integrated system of DBs creation in the field of inorganic chemistry and materials science. Important feature of developed integrated system is ability to include DBs that were created by means of different DBMS using essentially various computer platforms: Sun (DB "Diagram") and Intel (other DBs) and diverse operating systems: Sun Solaris (DB "Diagram") and Microsoft Windows Server (other DBs).*

***Keywords**: Integration of databases, metabase, distributed information system, inorganic substances and materials, EII, EAI.*

***ACM Classification Keywords**: H.2.4 Distributed databases, H.2.8 Scientific databases, J.2 Chemistry.*

Introduction

At present rich variety of databases on properties of inorganic substances and materials were developed and maintained in the world [Bale and Eriksson, 1990; Dudarev et al., 2006; Eriguchi and Shimura, 1990; Khristoforov et al., 2001; Kiselyova, 2005; Kiselyova et al., 2004, 2005, 2006; Villars et al., 2004; Xu et al., 2006; Zemskov et al., 1998]. Traditional areas, that DBs cover, are thermodynamic, thermo-chemical, crystallographic and crystal chemical properties. The majority of large industrial corporations support development of DBs that contain the information on physical, technical and technological parameters of materials and substances. The development tendencies of modern DBs on properties of inorganic substances and materials are following:

1. Internet-access to the information.

2. The use of powerful DBMS: Microsoft SQL Server, Oracle, IBM DB2, etc.

3. Great attention has been concentrated on the quality (reliability) of stored information. Highly skilled specialists are engaged in development process of the most "advanced" commercial information systems for data capture and expert estimation of data reliability. So users receive not simply "row" information but recommended values passed filtration for elimination of misprints.

4. Often DBs are supplemented with information analysis tools: from traditional thermodynamic calculations and statistical procedures up to modern means for regularities search in the data allowing predicting behavior of objects and making decisions. In the last case usual DBs, oriented to transaction processing, are often supplemented, for example, with special integrated information systems, that are named in the English literature as *Data Warehouse* [Kimball and Caserta, 2004]. They are intended for data coordination and integration from various information sources and its preparation for the subsequent computer analysis.

5. Integration of DBs on substances and materials. In this case the user can find the most complete cumulative information on properties of a certain substance.

The last problem information integration of resources on inorganic substance and material properties is the most important today. The data about various properties of a certain substance or material are distributed among different heterogeneous DBs. The chemist or material scientist has to look through a great number of DBs in

order to find the necessary information. Therefore some superstructure above DBs, that will allow to output some cumulative – the integrated information on all set of properties of a substance stored in different information systems, is required. That is, an integration of DBs is necessary. The solution of this problem is concerned with several difficulties. Databases on inorganic substance and material properties were developed in various organizations and countries and thus they use different database management and operating systems. Taking into consideration differences in data quality, data expertise procedures, data formats, languages and many other troubles it should be stated that full and smooth integration of information resources is practically impossible problem. We developed the approach to DBs integration taking into consideration the peculiarities of DBs on inorganic substance and material properties. The approach can be used for integration of Russian and Japanese DBs in this knowledge domain.

Known Approaches to Database Integration

Principally there are three approaches to database integration [Imhoff, 2005]:

1) Data Warehouse based on ETL (Extract, Transform, Load) paradigm – technology [Kimball and Caserta, 2004].

2) EII (Enterprise Information Integration) – technology [Morgenthal, 2005].

3) EAI (Enterprise Application Integration) – technology [Morgenthal, 2000].

These approaches can be used for solution of wide set of problems: from real-time integration to batch integration and from data integration to applications integration. Fig. 1 illustrates these approaches area of application in relation to the different task types [Imhoff, 2005]. The EII technology is the best approach for real-time data integration. The ETL technology allows the best batch data integration. The EAI technology gives the best results at applications integration in real-time or batch modes.

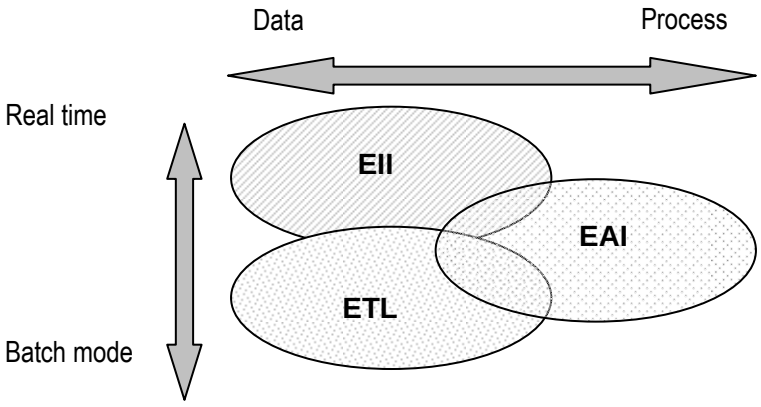


Fig. 1. Modern approaches for information systems integration.

The ETL-technology implies full merging of existing resources (fig. 2). That is the case when database complex is a single information system (*megabase*) for end users, operators and administrators. This approach is also known as Data Warehouse [Imhoff, 2005; Kimball and Caserta, 2004]. So at first information is extracted from DBs to be integrated. Then these data are somehow processed for clearing (that is, check for discrepancies and elimination of obviously false data) and transformations – series of special procedures that allow to get a common unified format and dimension. Only after these stages cleared and unified data are input into data warehouse or megabase. Database exploitation costs reduction and information duplication reduction can be mentioned among advantages of this integration approach.

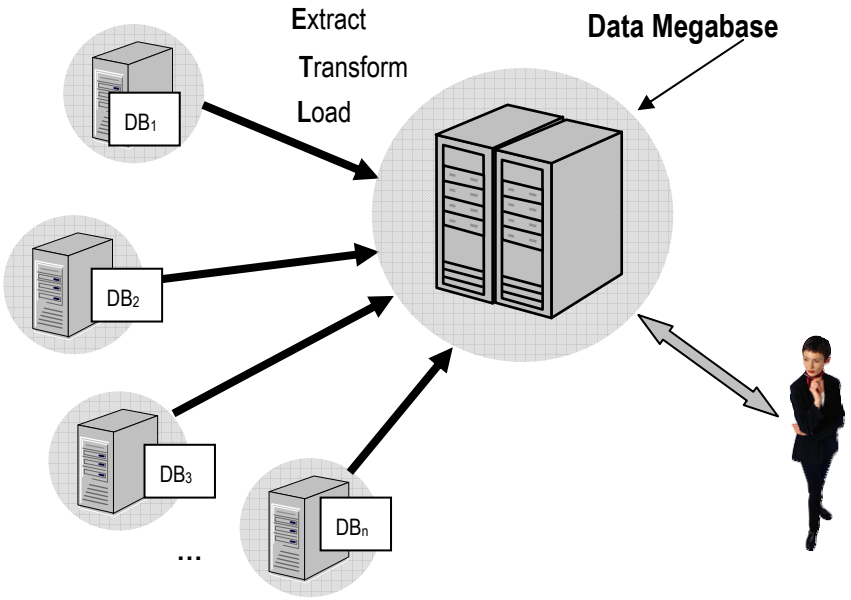


Fig. 2. ETL-approach – full merging of existing DBs.

The second integration approach is based on EII-technology (fig. 3). It is not going to integrate databases themselves [Imhoff, 2005; Morgenthal, 2005]. Integrated data are not transferred into a central megabase but remain in the same information systems, as before. Instead the program interface for data access is developed that allows retrieving required data. EII is data integration means from multiple systems into a unified, consistent and accurate representation format geared toward the data browsing. So the data are aggregated, restructured and relabeled (if it is necessary) and presented to a user. Usually the result of this approach is a virtually integrated heterogeneous distributed information system.

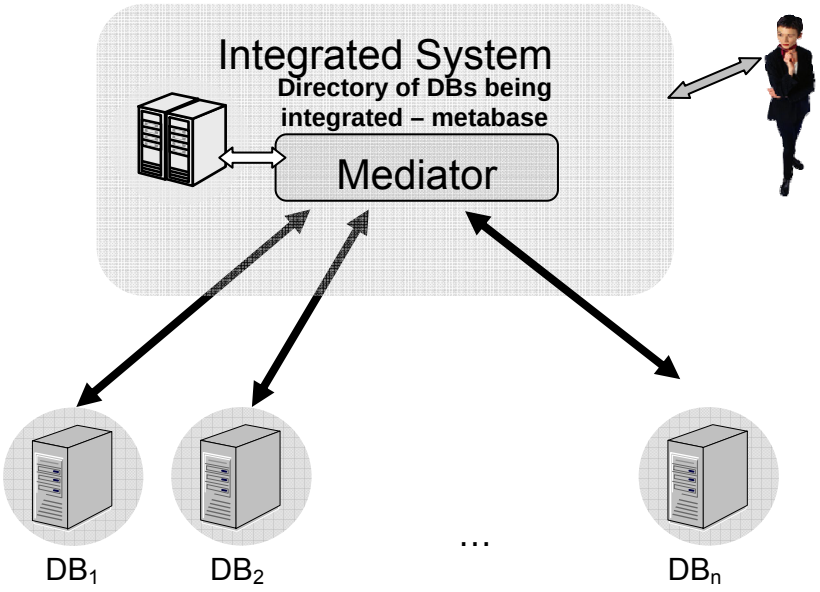


Fig. 3. EII-approach – data integration in real time.

The third approach – EAI – (fig. 4) is aimed for applications integration [Imhoff, 2005; Morgenthal, 2000]. Integration can be carried out in a batch mode or in real-time mode. Combined work of two and more applications can be achieved using this approach. This approach is based on message exchange between several applications. Frequently such information exchange is carried out through some common message exchange infrastructure known as message bus. Applications are connected to this common message bus by means of special adapters.

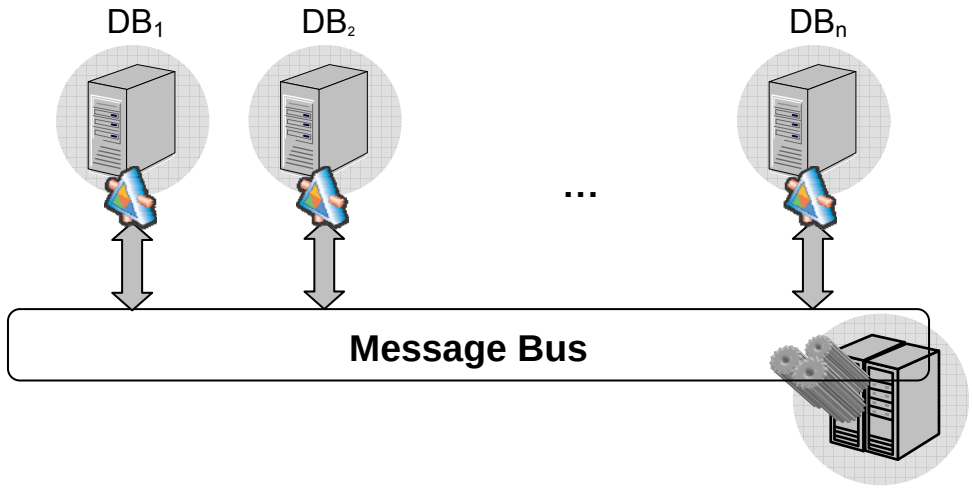


Fig.4. EAI-approach – integration of applications.

The EII and EAI technologies allow not to change every integrated database structure dramatically (and thus established database administration technology). So called "virtual" database integration and heterogeneous distributed information system creation implies an independence in evolution of separate subsystems and at the same time allows to end user to get access to the whole "live" data array on a certain chemical substance or material that are stored in databases of virtually united system.

So EAI technology integrates transactions of two or more applications, ETL technology merges the data of several information sources into a single one, and EII technology carries out virtual data integration from various information sources. It should be mentioned that no one approach can solve all tasks arising when integrating information systems on material and substance properties.

It is necessary to take into consideration that every data center on materials properties is a point of information concentration and data analytical processing bases on different software and hardware. The technology of information accumulating and data processing has been settled down in each organization. So great investments that were made in hardware and software do not allow mechanically transporting all the data into some centralized database. Moreover many DBs on material and substance properties are equipped with ancillary programs for substance parameters calculation. Therefore taking into consideration current development conditions of databases on inorganic substance and material properties the integrated system based on both EAI- and EII-technologies was developed in Baikov Institute [Dudarev et al., 2006; Kornuyshko and Dudarev, 2006] (fig. 5). It allows dynamically integrating a plenty of heterogeneous databases that are supplied with any computational subsystems.

Integration of Russian Databases on Inorganic Material and Substance Properties

From the beginning the proposed approach was used for integration of Russian DBs on inorganic material and substance properties. The successful task decision of integration needs is some coordinating center, which "knows" what information is stored in every integrated DB. Such function can be carried out by *metabase* – a special metadata database that stores information on integrated DBs contents, namely, about chemical systems,

substances and its modifications. Chemical systems are identified by a set of, which are included into its composition. Chemical substances are determined by a set of chemical elements (as systems) and their numeric contents in substance. Chemical modifications are defined as chemical substances having special crystal structure of phases. Metabase contains also information on properties, which data are stored in different DBs, and other data. This information is enough to make search for relevant chemical systems and data on properties of substances and materials.

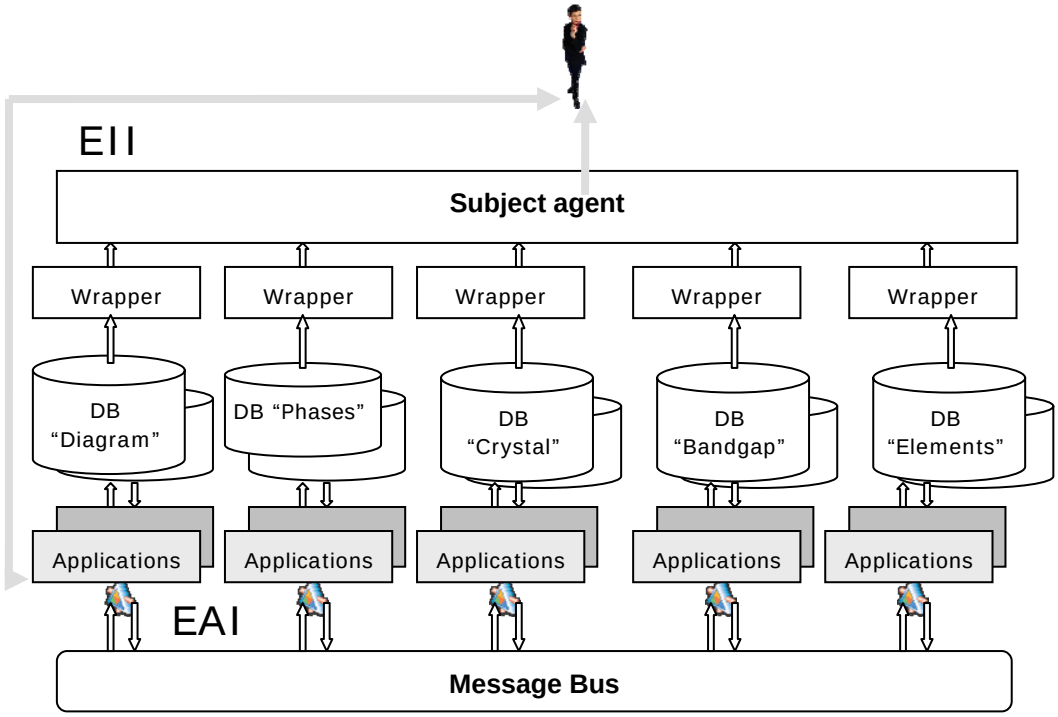


Fig. 5. Structure of integrated DBs system of Baikov Institute.

Now the integrated information system includes five DBs that were developed by Baikov Institute: DB on the properties of inorganic compounds "Phases" [Kiselyova et al., 2006], DB on phase diagrams of semiconducting systems "Diagram" [Khristoforov et al., 2001], DB on substances with significant acousto-optical, electro-optical and nonlinear-optical properties "Crystal" [Kiselyova et al., 2004], DB on width of the forbidden zone of inorganic substances "Bandgap" [Dudarev et al., 2006] and DB on properties of chemical elements "Elements" (fig. 5). One of the most important features of the developed integrated system is that DBs which have been included into integrated system were created with various DBMS using essentially different computer platforms: Sun (DB "Diagram") and Intel (other DBs) and different operational systems: Sun Solaris (DB "Diagram") and Microsoft Windows 2003 Server (other DBs). However the way, offered by us, has appeared successful even in such a difficult case for program realization.

Integration of Russian and Japanese Databases on Inorganic Material and Substance Properties

Next stage is an expansion of integrated system. Information system of Baikov Institute will be integrated with other Russian [Zemskov et al., 1998] and foreign DBs [Villars et al., 2004; Xu et al., 2006] on inorganic materials and substances. Principles of integration are based on the application of metabase and combined approach that was developed in Baikov Institute [Dudarev et al., 2006; Kornuyshko and Dudarev, 2006]. Sometimes small additional tables, that contain information about sets of elements and their contents in substance and crystal structure, will be included into these DBs.

The following structure of metabase can be used for integration of the Web-applications of DBs on properties of inorganic substances and materials (fig. 6).

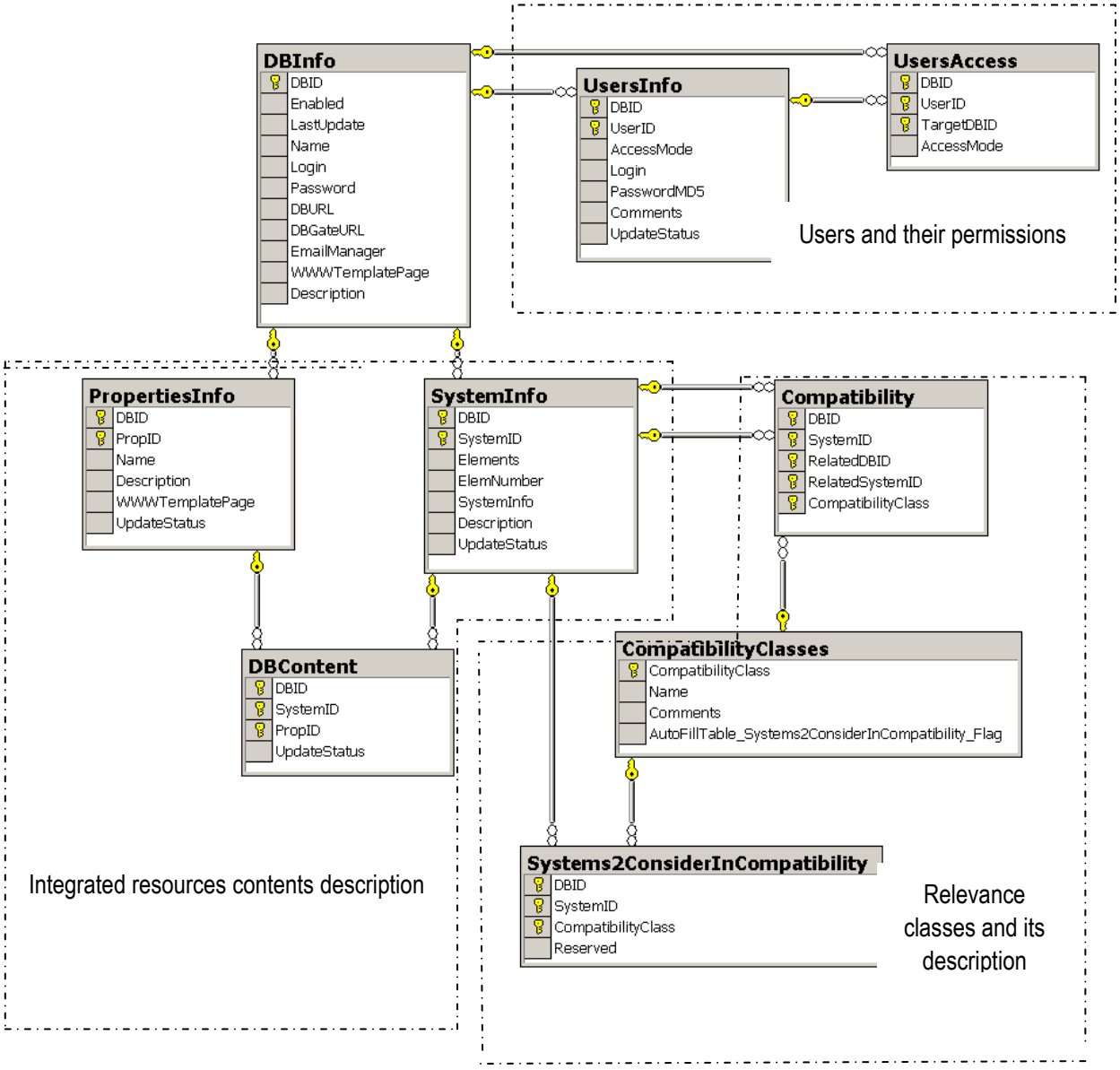


Fig. 6. Metabase structure for DBs Web-applications integration.

Purpose of the tables (fig. 6): **DBInfo** – main table containing the information about DBs Web-applications to be integrated; **UsersInfo**, **UsersAccess** - tables containing the information on the users of integrated system and their access permissions to information; **SystemInfo**, **PropertiesInfo**, **DBContent** – tables that describe contents of resources to be integrated (what information on chemical systems and their properties is stored in what DB); **CompatibilityClasses**, **Compatibility**, **Systems2ConsiderInCompatibility** – tables that contain information on relevance classes and determe relevant chemical systems.

Conclusion

The complex approach to information integration combining the integration at a level of data and user interfaces (EII+EAI) is offered. Within proposed approach access was implemented to all current user interfaces of virtually

united information system. Moreover the system allows to users to move freely between different applications (EAI). According to the common developed information schema subject mediator was implemented. It provides rich opportunities on information extraction and aggregation from diverse distributed data sources on material and substance properties (EII).

The tasks of search for the relevant data in integrated information systems and implementation of transparent user transition between DBs Web-applications (taking into account the security issues) were solved at DBs Web-applications integration. For search mechanisms implementation for the relevant information was used metadata database (metabase) – special reference database containing metadata only – the information on information systems to be integrated. Diverse data sources integration is based on conceptual structure of the knowledge domain (inorganic chemistry) and development of the heterogeneity conflicts resolution ways.

The system of databases on inorganic material and substance properties is accessible for registered users of the Internet: http://www.imet-db.ru.

The work is supported by RFBR, grants №06-07-89120 and 05-03-39009.

Bibliography

[Bale and Eriksson, 1990] C.W.Bale and G.Eriksson. Metallurgical thermochemical databases a review. Can.Met.Quart. 1990, v.29.

[Dudarev et al., 2006] V.A.Dudarev, N.N.Kiselyova, V.S.Zemskov. Integrated system of databases on properties of materials for electronics. Perspektivnye Materialy, 2006, N.5 (Russ.).

[Eriguchi and Shimura, 1990] K.Eriguchi and K.Shimura. Factual databases for materials design and manufacturing. ISIJ Int., 1990, v.30.

[Imhoff, 2005] C.Imhoff. Intelligent Solutions: Understanding the Three E's of Integration EAI, EII and ETL. DM Review Magazine, 2005, apr. (http://www.dmreview.com/article_sub.cfm?articleId=1023893).

[Khristoforov et al., 2001] Yu.I.Khristoforov, V.V.Khorbenko, N.N.Kiselyova, et al. Internet-accessible database on phase diagrams of semiconductor systems. Izvestiya VUZov. Materialy elektron.tekhniki, 2001, №4 (Russ.).

[Kimball and Caserta, 2004] R.Kimball and J.Caserta. The Data Warehouse ETL Toolkit: Practical Techniques for Extracting, Cleaning, Conforming, and Delivering Data. John Wiley & Sons, 2004.

[Kiselyova, 2005] N.N.Kiselyova. Computer Design of Inorganic Compounds. Application of Databases and Artificial Intelligence. Nauka, Moscow, 2005 (Russ.).

[Kiselyova et al., 2005] N.N.Kiselyova, V.A.Dudarev, I.V.Prokoshev, et al. The distributed system of databases on properties of inorganic substances and materials. Int.J."Information Theories & Applications", 2005, v.12.

[Kiselyova et al., 2006] N.Kiselyova, D.Murat, A.Stolyarenko, et al. Database on ternary inorganic compound properties "Phases" in Internet. Informazionnye resursy Rossii, 2006, N.4 (Russ.).

[Kiselyova et al., 2004] N.N.Kiselyova, I.V.Prokoshev, V.A.Dudarev, et al. Internet-accessible electronic materials database system. Inorganic materials, 2004, v.42, №3.

[Kornuyshko and Dudarev, 2006] V.Kornuyshko and V.Dudarev. Software Development for Distributed System of Russian Databases on Electronics Materials. Int. J. "Information Theories & Applications", 2006, v.13.

[Morgenthal, 2000] J.P.Morgenthal. Enterprise Applications Integration with XML and Java. Prentice Hall PTR; Bk&CD Rom edition, 2000.

[Morgenthal, 2005] J.P.Morgenthal. Enterprise Information Integration: A Pragmatic Approach. Lulu.com, 2005.

[Villars et al., 2004] P.Villars, M.Berndt, K.Brandenburg, et al. The Pauling File, binaries edition. J.Alloys and Compounds, 2004, v.367.

[Xu et al., 2006] Y.Xu, M.Yamazaki, H.Wang, K.Yagi. Development of an Internet system for composite design and thermophysical property prediction. Mater.Trans., 2006, v.47.

[Zemskov et al., 1998] V.S.Zemskov, F.A.Kuznetsov, V.B.Ufimtsev. Databanks on semiconducting and other materials for electronics and technology of their productions. Izvestiya VUZov. Materialy elektronnoi tekhniki, 1998, N.3 (Russ.).

Authors' Information

Shuichi Iwata *– Graduate School of Frontier Sciences, The University of Tokyo, P.O.Box: 113-8656, Room 507A, Building E12 QUEST, 7-3-1 Hongo, Bunkyo-ku, Tokyo, Japan, e-mail: iwata@k.u-tokyo.ac.jp*

Nadezhda Kiselyova *– e-mail: kis@ultra.imet.ac.ru*

Victor Dudarev *– e-mail: vic@osg.ru*

Ilya Prokoshev *– e-mail: eldream@e-music.ru*

Valentin Khorbenko *– e-mail: Khorbenko_v@mail.ru*

Victor Zemskov *– e-mail: zemskov@ultra.imet.ac.ru*

A.A.Baikov Institute of Metallurgy and Materials Science of Russian Academy of Sciences, P.O.Box: 119991 GSP-1, 49, Leninskii Prospect, Moscow, Russia

BENEFITS OF TSPI IN A SOFTWARE PROJECT UNDER A SMALL SETTINGS ENVIRONMENT

Jose Calvo-Manzano, Gonzalo Cuevas, Tomás San Feliu, Edgar Caballero

Abstract*: This article introduces a small setting case study about the benefits of using TSPi in a software project. An adapted process from the current process based on the TSPi was defined. The pilot project had schedule and budget constraints. The process began by gathering historical data from previous projects in order to get a measurement repository. The project was launched with the following goals: increase the productivity, reduce the test time and improve the product quality. Finally, the results were analysed and the goals were verified.*

Keywords*: TSPi, Small Settings, Process improvement.*

ACM Classification Keywords*: D.2.9 Software Engineering – Management, software process models*

Introduction

Large and small and medium enterprises have common problems related to the management and quality [IPRC, 2006] of software projects. This generates costs overruns, low quality and cancelled projects [Standish Group, 2004].

Some processes models like the CMMI are successful enough, but they are not affordable for the small organizations [IPRC, 2006].

Organizations have recognized that the control of their software processes affects the success of their projects, *"they know what to do but not how to apply it"* [Noopur, 2003].

A new research line based on the process improvement in small settings is arising in order to facilitate competitive capabilities for this environment in a global market [Glazer, 2006]. Small Settings include small and medium organizations, and small software projects [Garcia 2006].

Garcia [Garcia, 2005] and Serrano [Serrano, 2006] show how to get CMMI maturity levels using TSP in Small Settings. Some CMMI level 5 organizations have improved their quality levels using TSP [Noopur, 2003].

Team Software Process (TSP) is a framework that provides a customizable process based in an excellent experience in planning and managing software projects [Humphrey, 2006]. It guides teams in managing cost, schedule and quality [Noopur, 2003].

This article shows through a case study the results of using an adapted process based on the introduction of the Team Software Process (TSPi) in a small organization.

The following goals have been established for the adapted process:

1. To finish the project within the established schedule, cost and effort.
2. To reduce the test time.
3. To increase the productivity and improve the product quality.

The organization decided to use TSPi in order to accomplish the previous goals assuming the risk of modifying its previous processes. Besides, there was not enough time or resources to elaborate a complete training in TSPi and PSP.

Therefore, the organization decided to apply the basic TSPi principles, getting a customized process as a result of combining TSPi with the previous organizational process.

A basic training was provided for the new process, and historical data were collected in order to facilitate the estimation of the pilot project and the comparative study.

In the following section, the article shows the organization, the development context and the pilot project attributes. Later, the historical data collection will be described, and the new process will be showed and the advantages pointed out. The project goals will be verified using the project measures, and finally, the conclusions will be showed.

The schema showed in Figure 1 resumes the factors considered in the project.

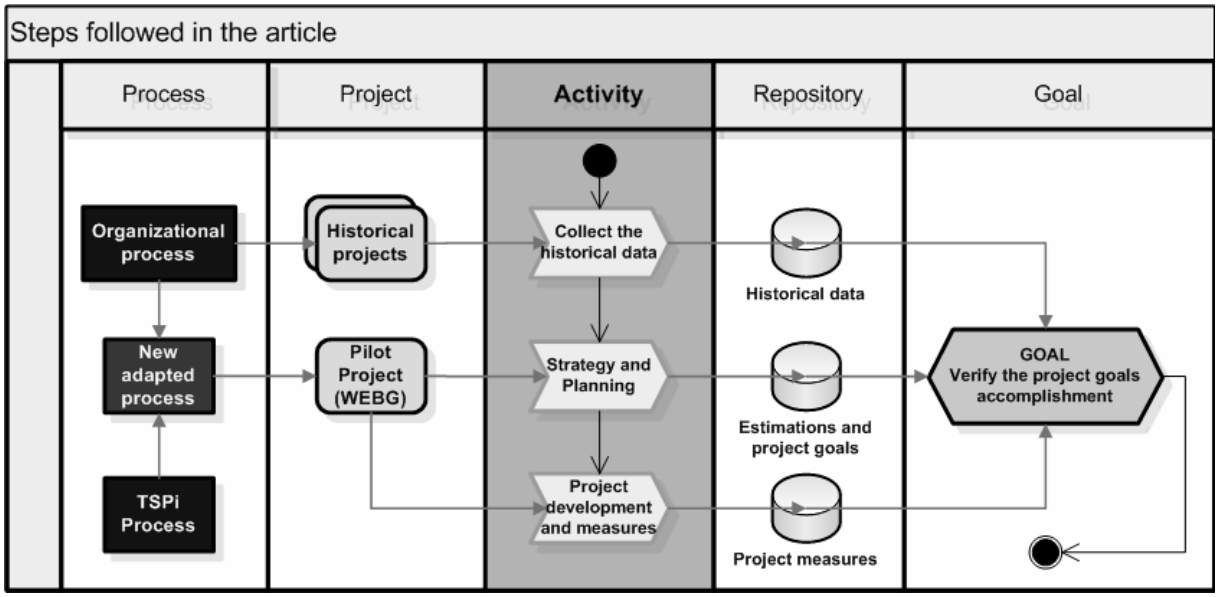


Figure 1 – Factors considered in the project

Context: The organization and the pilot project

UPTA is a Spanish intersectorial organization who takes care of all the scopes of economic activity which they are exerted by self – employment. UPTA leads a lot of projects which generally need a specific software development and in the last year, the number of software projects has increased.

The working scenario has changed to a new environment where many projects were developed simultaneously, and with a greater number of involved people. As a result of an internal assessment, senior management detected that projects were delayed, dedicating additional efforts to accomplish the objectives. Moreover, products quality had decreased.

UPTA was interested in introducing a process model such as CMMI, but it could not afford it.

Besides this handicap, UPTA had a project (called PRO) with schedule and budget constraints, and was delayed. The organization selected this project as the pilot. The purpose of "PRO" was to develop a tool that allows creation of a web site based on templates and a basic content system where the end users does not need technical knowledge.

According to TSPi criteria, during the strategy phase, the team agreed to reduce the initial functionality by 20%. The team established the following project goals (see Table 1):

Table 1 – Project goals

Measure	Goal
Schedule deviation	< 8% (1 week)
Effort deviation	< 15%
Budget deviation	< 15%
Test productivity	< 33.4 hours/KLOC (historical average)
Project productivity	> 7.3 LOC/hour (historical average)
% Release defects	< 5%

Collecting the historical data

Data on previous UPTA projects were not enough. There were only schedule and budget data, but in order to verify the project goals, defects and phases efforts data were needed.

Estimations values and measures related to schedule, size, effort and defects were collected. In addition, some derived metrics were calculated in order to analyze the project results.

Lines of code (LOC) were chosen because it could be done automatically. Based on the LOC and the effort of previous projects, the historical average productivity was calculated. This information was used to estimate the pilot project size.

In order to support the analysis, historical projects were divided into three phases:

- Development phase (process): From the launch until before the test.
- Test phase: It includes integration and system tests.
- Release phase: From the product release to the customer until the end of the third month of use.

Phase effort and defects data of these phases are approximate values because there was no previous data repository.

The selected historical projects were: HIS-1 (23 KLOC), HIS-2 (7 KLOC), HIS-3 (33 KLOC), HIS-4 (11 KLOC) and HIS-5 (104 KLOC).

The process

The process is a customized process as a result of blending the basic TSPi principles and the previous organizational process.

Once the new process was defined, the project started with training on the new process and the launching meeting.

The TSPi phases were used in the new process in order to get benefits from its procedures and metrics, but the intermediate products, such as requirements or design specifications, were based on the previous organizational process in order to reduce the change impact.

The focus on quality is the main difference with the previous organizational process. Examples of this approach are the quality plan relative to the phases and processes performance, inspections and reviews.

With respect to project management, weekly meetings and the earned value method were introduced. These gave to the project a real visibility and an effective tracking. The schedule, goals, risks, and change requests were evaluated in the weekly meetings.

The team was empowered to estimate and plan the project balancing the workload, and so, were more committed. Also, a good role definition was adopted.

Table 2 shows the basic TSPi principles applied in the process and the difference with the previous process.

Table 2 –TSPi principles applied in the new process

New process	Previous process
Process well defined. It makes easier the estimation and tracking project	Process with inconsistencies. The phases are not well defined
Team motivated, participative and collaborative	Only a project leader elaborates the project plan and the task distribution
Quality focus based in an early defect detection and reduction	Since the schedules are restricted, the quality was not considered
Introduction of inspection activities in the process	Only personal reviews without a quality control
Detailed plan in order to avoid schedule, and effort deviation	Projects begin with cost and schedules pre-established and restricted
Tracking and project visibility with the earned value method	There is no mechanism to track the project status
Weekly meeting to analyse the project and resolve process issues	There are no formal meetings and they are preformed only when there are problems

Verifying the project goals

In order to verify the project goals, measures were evaluated based on the initial plan (see Table 1).

5.1. Goal 1: To finish the project within the established schedule, cost and effort.

The results obtained in the project related to schedule, size and effort are (see Table 3):

Table 3 – Estimation vs. Actual

Measure	Estimation	Actual	Deviation
Schedule [SEM]	13.0	14.0	7.7%
Effort [HRA]	950.0	1121.0	18.0%
Size [KLOC]	6.9	8.5	22.5%

Table 4 shows that only there was one week of delay in the schedule. The effort can be considered acceptable because the actual value is close to the estimated value.

Table 4 – Goal 1 results

Measure	Goal	Actual	Deviation
Schedule deviation	< 8% (1 week)	7.7%	-3.8%
Effort deviation	< 15%	18.0%	20.2%
Cost deviation	< 15%	18.0%	20.2%

Cost data are derivated from effort and the results are similar.

As an example of the earned value tracking visibility, to deliver the product on time, the team decided to work with a little more intensity at seventh week because they observed a possible delay. Also, as can be seen in Figure 2, in last week (13) there was no earned value because the team was dedicated to fix a defect detected in the system test phase.

The weekly meetings and the earned value method allowed the improvement of the project management [Humphrey, 1995] (See Figure 2).

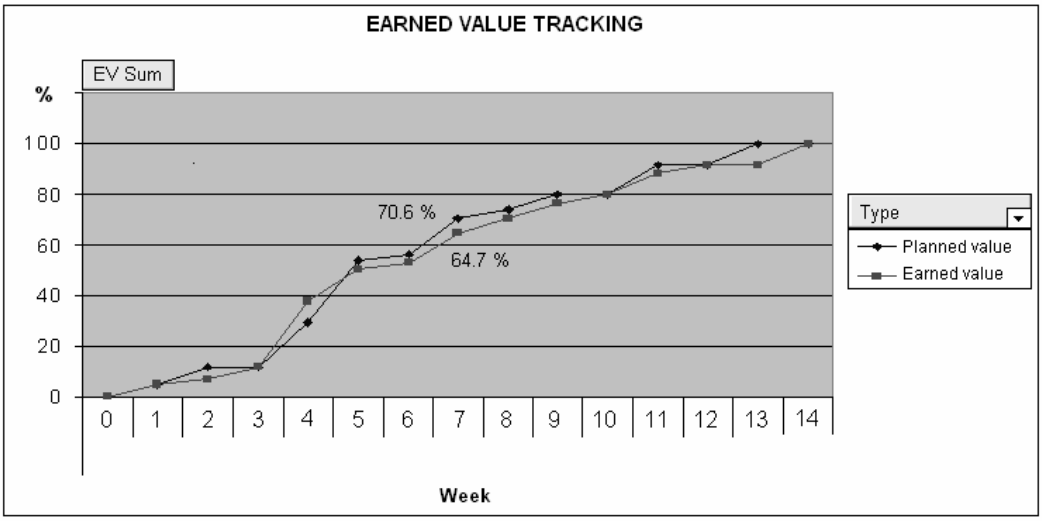


Figure 2 – Earned value tracking

5.2. Goal 2: To reduce the test time

Table 5 shows the reduction on the test time and test productivity. Note that the goal values were established using the average of the historical data

Table 5 – Goal 2 results

Measure	Goal	Actual	Deviation
Test time reduction	< 24.4 %	10.0 %	-59.1%
Test productivity [Hours/KLOC]	< 33.4	13.2	-60.5%

Figure 3 shows that the test productivity has improved 20.2 hours/KLOC, which means 60.5 % better than the historical average.

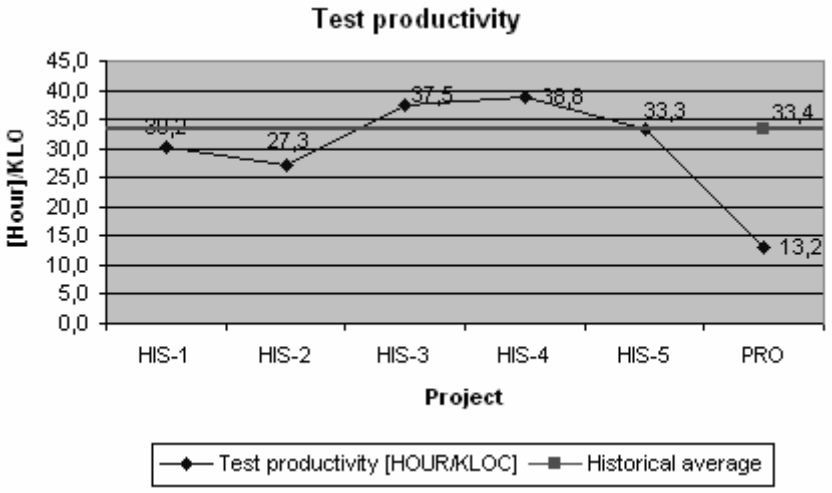


Figure 3 – Test productivity

Figure 4 shows the test time reduction. Only 10.0% of the project time was needed, which means 59.1% lower than the historical average.

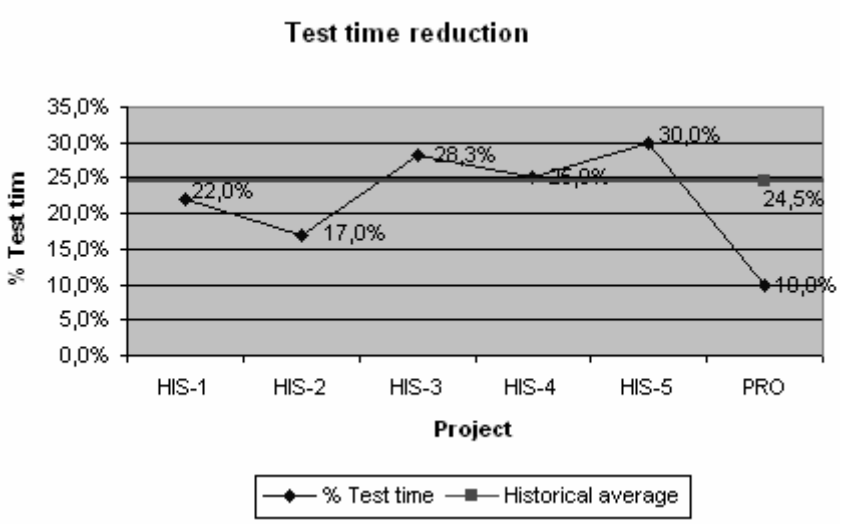


Figure 4 – Test time reduction

5.3. Goal 3: Increase the productivity and improve the product quality

Table 6 shows the project productivity and released defects goals. The project productivity had no important improvement.

Table 6 – Goal 3 results

Measure	Goal	Actual	Deviation
Project productivity [LOC/Hour]	> 7.3 %	7.6 %	3.9%
% Released defects	< 5.0	3.8	-24.8%

Figure 5 shows the project productivity improvement compared to the historical average.

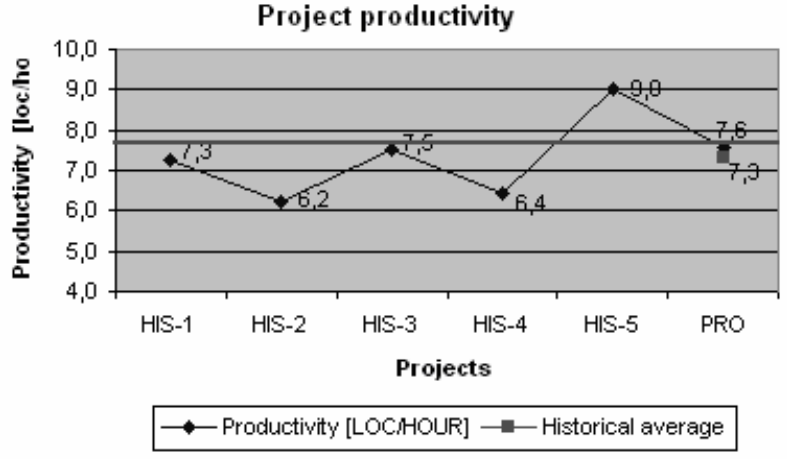


Figure 5 – Project productivity

One of the best results of this project was the reduction of the released defects. A released defect is a defect found during the first three months of operation. This was possible because the quality TSPi principles were applied, introducing reviews and inspections to get an early defect detection.

Figure 6 compares the released product defects using the new process with the historical average.

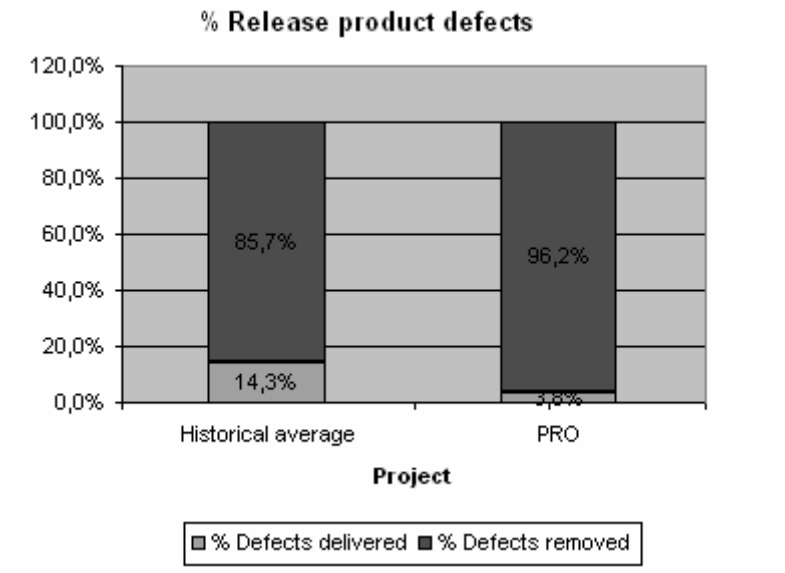


Figure 6 – Release product defects

Conclusion

The use of TSPi principles in the new process allowed the accomplishment of project goals based on the following considerations.

1. The team integration, the detailed plan, the change management, the weekly meetings and the earned value method allowed the accomplishment of these goals.
2. Along the project, the responsibility of the team members increased the test productivity by reducing the rework.
3. The reviews, inspections and quality plans allowed the reduction of test effort. The team members understood the test phase as a quality evaluation and not as a defect detection activity.

With an affordable investment in process definition, it has been demonstrated that using TSPi with adaptations has permitted a good solution for process improvement in Small Settings.

This article will be the foundation for future actions such as establishing the adapted process performance or comparing the adapted process quality versus the TSPi quality indicators.

Bibliography

[Standish Group, 2004] Standish group. CHAOS Report, 2004.

[Serrano, 2006] Serrano, M., Montes, C., Cedillo, K. An Experience on Implementing the CMMI in a Small Organization Using the TSP, 81-92. http://www.sei.cmu.edu/pub/documents/06.reports/pdf/06sr001.pdf, 2006.

[IPRC, 2006] International Process Research Consortium. IPSS White Paper. Improving Process in Small Settings, 2006 http://www.sei.cmu.edu/iprc/ipss-white-paper-v1-1.pdf

[Noopur, 2003] Noopur, D. The Team Software Process in Practice: A Summary of Recent Results. SEI Technical Report CMU/SEI-2003-TR-014, 2003.

[Humphrey, 2006] Humphrey, W. TSP: Coaching Development Teams. Ed. Addison-Wesley Publishing Company, 2006

[Humphrey, 1999] Humphrey, W. Introduction to the Team Software Process. Ed. Addison-Wesley Publishing Company, 1999.

[Humphrey, 1995] Humphrey, W. A Discipline for Software Engineering. Ed. Addison-Wesley Publishing Company, 1995

[Garcia, 2006] Garcia, S. Graettinger, C., Kost K. Proceedings of the First International Research Workshop for Process Improvement in Small Settings. SEI Special Report CMU/SEI-2006-SR-001, 2005.

[Garcia, 2005] Garcia, S. Thoughts on Applying CMMI in Small Settings.
http://www.sei.cmu.edu/cmmi/adoption/pdf/garcia-thoughts.pdf, 2005.

[Glazer, 2006] Glazer, H. Time to Market vs. Process Discipline.http://www.sei.cmu.edu/iprc/sepg2006/glazer.pdf, 2006.

Authors' Information

Calvo-Manzano Jose A.- *Universidad Politécnica de Madrid, Facultad de Informática, Campus Montegancedo, Boadilla del Monte-28660, Madrid-España; e-mail: jacalvo@fi.upm.es*

Cuevas Gonzalo - *Universidad Politécnica de Madrid, Facultad de Informática, Campus Montegancedo, Boadilla del Monte-28660, Madrid-España; e-mail: gcuevas@fi.upm.es*

San Feliu Tomás - *Universidad Politécnica de Madrid, Facultad de Informática, Campus Montegancedo, Boadilla del Monte-28660, Madrid-España; e-mail: tsanfe@fi.upm.es*

Caballero Edgar Henry – *Universidad Politécnica de Madrid, Facultad de Informática, Campus Montegancedo, Boadilla del Monte-28660, Madrid-España; e-mail: ecaballero@zipi.fi.upm.es*

АВТОМАТИЗАЦИЯ ТЕСТИРОВАНИЯ И ДОКУМЕНТИРОВАНИЯ ИНФОРМАЦИОННЫХ СИСТЕМ

Антон Цыбин, Людмила Лядова

Abstract*: Статья посвящена описанию подходов к автоматизации тестирования и создания пользовательской документации в программных системах с графическим пользовательским интерфейсом. Для автоматизации тестирования применён комбинированный подход на основе машин состояний и data mining. Для автоматического составления документации применён подход, основанный на использовании описания системы с помощью метаданных, представленных связанными списками. Предложенные методы позволяют ускорить и повысить качество процессов тестирования и документирования приложений без требования задания сложных входных данных.*

Keywords*: тестирование программ, автоматическое создание документации, data mining, фиксированные грамматики, изменяемые грамматики.*

ACM Classification Keywords*: D.2 Software Engineering: D.2.2 Design Tools and Techniques; D.2.4 Software - Program Verification; I.2 Artificial Intelligence: I.2.2 Automatic Programming.*

Введение

На сегодняшний день одним из базовых понятий методологии проектирования программных систем является понятие жизненного цикла. Жизненный цикл – это непрерывный процесс, который начинается с момента принятия решения о необходимости создания информационной системы (ИС) и заканчивается в момент окончания её эксплуатации.

Структура жизненного цикла включает три группы процессов: основные, вспомогательные и организационные. Основные процессы включают следующие этапы:

- создание,
- внедрение и эксплуатация,
- сопровождение.

Этап создания программы также является сложным и включает в себя:

- оценку жизненного цикла системы,
- анализ требований,
- проектирование структуры компонентов,
- реализацию проекта.

На этапе внедрения и эксплуатации производятся контрольное тестирование и приёмо-сдаточные испытания. Данный этап направлен на обеспечение качества программного продукта.

Все рассмотренные этапы создания программной системы сопровождаются вспомогательным процессом – документированием. В рамках данного процесса производятся работы по созданию различных инструкций и руководств, документации пользователя и программиста.

Из всех описанных этапов жизненного цикла наиболее формализованными и обеспеченными автоматизированными средствами являются этапы анализа требований и проектирования модели предметной области. Меньше внимания в жизненном цикле уделяют этапам тестирования и создания документации. Такая закономерность наблюдается, прежде всего, в небольших компаниях по разработке программного обеспечения (ПО). Причина состоит в сложности автоматизации тестирования и составления документации. Несмотря на обилие средств автоматизации, среди них не существует универсального продукта, подходящего для любой предметной области, для которой создается ПО. На выполнение данных этапов вручную требуется большой объём времени и средств, зачастую эти затраты не окупаются для небольших компаний. Однако тестирование – очень важный этап жизненного цикла, так как недостаточное внимание к нему не позволяет получить качественный, надежно функционирующий программный продукт. Для измерения степени автоматизации тестирования в жизненном цикле Институтом Иллинойса была предложена модель зрелости тестирования (TMM – Test Maturity Model) [1], содержащая набор приёмов повышения степени автоматизации.

Многие организации игнорируют этап создания документации, особенно на поздних стадиях создания продукта, что резко снижает эффективность внедрения и сопровождения, так как при отсутствии описания требований, описания функций системы затруднено изменение программного кода системы в случае обнаружения ошибок или переноса системы на новую программно-аппаратную платформу. При отсутствии или ненадлежащем состоянии пользовательской документации осложняются этапы внедрения и сопровождения системы, так как в этом случае снижается эффективность обучения пользователей, их неверные действия при работе с ПО приводят к необходимости исправления ошибок в данных и увеличению числа вопросов к службе технической поддержки.

Подходы к тестированию

На сегодняшний день рынок программного обеспечения предлагает огромный выбор продуктов, позволяющих автоматизировать тестирование программ на различных стадиях разработки. Существуют средства, проверяющие полноту и непротиворечивость требований к программе, тестирующие сетевое взаимодействие, основную функциональность приложений методами «чёрного», «белого» и «серого» ящиков, надёжность работы системы при больших и запредельных нагрузках. Существуют даже инструменты, позволяющие автоматизировать тестирование удобства использования графического интерфейса приложения.

Исторически первым методом тестирования является так называемый *метод «чёрного ящика»* [2]. Данный метод подходит для тестирования небольших программных модулей и сложен в использовании при тестировании больших программных систем, так как требует выполнения большого количества тестов для достижения приемлемого уровня уверенности в правильности программы.

Качественно новый подход к тестированию был предложен Энтони Хоаром – *метод доказательства корректности* программ [3]. Единственным серьёзным недостатком данного подхода является

чрезвычайная сложность построения математических моделей для больших программ. После некоторых попыток применения подхода, от него решено было отказаться.

Позднее были предложены иные методы тестирования, основанные на *анализе исходного кода программ для выявления ошибок*. Наиболее известным является метод «белого ящика», предложенный Филлисом Франклом и Элайном Веюкером в 1988 году [2]. Недостаток данных методов состоит в том, что тестирование программы производится без наличия знаний о правильном результате работы программы для каждого теста. При использовании данных методов предполагается, что результаты работы программы на тестах проверит человек. Но инструментальные средства тестирования с помощью покрытия кода генерируют огромное количество тестовых наборов, так что проверить результат работы каждого теста вручную затруднительно. В связи с этим был предложен *критерий выборочной проверки* (так называемая *N-выборка*), основанный на статистических методах.

Среди программ автоматизированного тестирования существуют *средства записи-воспроизведения действий человека*, осуществляющего тестирование. В Rational Robot имеется возможность определения правильности результата по свойствам визуальных элементов управления.

Существуют также научные разработки, использующие средства записи-воспроизведения (такие, как Rational Robot) для комбинирования достоинств нескольких подходов к организации тестирования. Например, в работе [4] описан оригинальный способ тестирования правильности работы графического интерфейса пользователя в программах. Недостаток такого подхода состоит в сложности построения и сопровождения диаграмм, описывающих состояния интерфейса.

Подходы к автоматическому созданию документации

На сегодняшний день на рынке ПО наибольшее распространение получили программы автоматического документирования исходного кода проектов. Рассмотрим несколько подходов.

Для создания программной документации существует множество программ, использующих одинаковый подход на основе грамматик. Грамматикой будем называть любой формализованный набор правил построения какой-либо системы.

Использование фиксированных грамматик

Большинство современных систем документирования исходного кода программных проектов используют фиксированные грамматики. Идея данного метода состоит в *использовании информации, содержащейся в правилах построения программ*. Правила построения программ иначе называют *грамматикой языка программирования*. Данные правила описаны в спецификациях языка программирования и, как правило, не могут быть изменены, т.е. являются фиксированными.

Например, существует программа NDoc для документирования исходного кода проектов на платформе dotNET. В данном примере в качестве грамматики можно рассматривать правила построения объектно-ориентированной программы для платформы dotNET, а также правила обработки XML-тегов комментариев.

Полностью аналогичный подход используется системой JavaDoc при построении программной документации проектов на языке Java.

Существуют также системы, создающие документацию для специфических языков программирования. Система LPdoc создает документацию для языков Lisp и Prolog. В качестве грамматики здесь можно рассматривать правила синтаксиса и семантики языков Lisp и Prolog.

Использование изменяемых грамматик

Реализация системы документирования программных проектов, полностью построенная на идее *конечных автоматов*, описана в работе [5].

Реализуемая в этом подходе архитектура позволяет создать общее описание проекта, а затем в любое время преобразовать данное описание в документацию. Аналогичные принципы построения компонента документирования используются и в представленном здесь исследовании.

Создание документации происходит на основе *шаблонов*, заранее заданных пользователем. Шаблоны содержат информацию о том, какую структуру должен иметь полученный документ.

В данном случае грамматика – это набор правил извлечения информации из определённым образом структурированного файла.

Отличие от подхода, основанного на фиксированных грамматиках, состоит в том, что наборы правил по выделению информации из различных файлов вынесены в XML-файл и хранятся вне системы документирования.

Предлагаемые подходы к автоматизации тестирования и создания документации

Данная работа посвящена исследованию возможности автоматизированной проверки правильности выходных данных программы с графическим пользовательским интерфейсом (GUI), а также разработке компонента автоматической генерации документации пользователя для информационных систем.

Приложения с графическим интерфейсом были выбраны в качестве объекта исследования, так как большинство современных прикладных систем реализуют такой интерфейс. Документация пользователя содержит описание функций системы, выполнение типовых операций через графический интерфейс. Поэтому необходимо реализовать автоматизированную проверку приложения через GUI с последующим описанием основной функциональности приложения через формы, с которыми работает пользователь.

Основной проблемой тестирования была и остаётся сложность описания правил, в соответствии с которыми должна работать проверяемая программа. Если для программ, осуществляющих научные вычисления, количество правил составляет 10-20, то для сложных программных комплексов их число составляет сотни тысяч. Несмотря на достоинства автоматического создания тестов метода «белого ящика», данный метод не в состоянии автоматически проверить правильность выполнения теста. Для этого либо необходим набор правил спецификации, либо требуется помощь человека.

Идея предлагаемого подхода состоит в том, чтобы при тестировании программы через графический интерфейс отслеживать значения ряда параметров, представляющих состояние системы. В качестве параметров предлагается выбрать внутренние переменные или значения свойств визуальных элементов управления. Благодаря возможности проверки внутренних переменных данный подход шире, чем тот, что используется в Rational Robot[1]. Входные данные для интерфейса можно генерировать на основе исходного кода. Например, используя критерий покрытия структур или потоков данных, можно создать множество тестовых наборов для элементов графического пользовательского интерфейса таким образом, чтобы максимально проверить существующий код тестируемой программы. От пользователя требуется лишь указать набор внутренних переменных для определения результата работы.

Генерацию данных для пользовательского интерфейса можно производить случайным образом. Такой метод тестирования менее эффективен, чем метод «белого ящика», поскольку не предусматривает целенаправленного тестирования узких мест в коде.

В процессе тестирования программы происходит сбор данных о значениях указанных параметров (состояния программы) и событиях, вызывающих переходы из одного состояния в другое. Поскольку отсутствует информация о спецификации на программу, необходимо каким-либо образом проверить, правильно или нет выполняется каждый тест. Для решения данной проблемы предлагается использовать методы Data mining.

[1] Система Rational Robot не позволяет отслеживать значения внутренних переменных тестируемой программы

Причина выбора нечёткого метода для идентификации правильности выполнения теста связана со сложностью задания спецификации на тестируемую программу. Поскольку создание набора правил, в соответствии с которыми должна работать программа, – сложная задача, можно привлечь набор нечётких методов для извлечения знаний из результатов тестирования. Иными словами, без вмешательства человека постараться выявить в результатах тестирования определённые закономерности, ассоциации, зависимости, последовательности и т.п. На основе полученных знаний можно делать выводы об исключительных ситуациях, аномалиях в результатах тестирования, а также в соответствии с определёнными критериями выбирать определённые тесты для проверки пользователем. Таким образом, после выполнения сотен тестов программы через графический интерфейс, система тестирования выдаст программисту несколько тестов для ручной проверки правильности. Получив информацию о правильности тестов от пользователя, система учитывает её при построении закономерностей и впоследствии ищет ошибки с учётом полученной информации.

В качестве методов Data mining подходящими для применения в данном подходе являются:

- *Алгоритм ограниченного перебора*. Проверяет простые логические события в различных подгруппах данных. Выбор логических событий производится эвристически.
- *Алгоритм выявления ассоциаций*. Перевычисляет меры доверия к комбинациям фактов на основе байесовской условной вероятности и понятия шанса.
- *Статистические методы*. Использование усреднённых величин для проверки статистических параметров распределений, доказательство гипотез о распределении случайных величин, проверка линейной корреляции.
- *Нейронная сеть*. Аппроксимация зависимостей с помощью нелинейных функций.

Применив перечисленные методы к данным о состояниях системы и переходам между состояниями, можно получить знания о взаимосвязях между указанными параметрами программы, а также об их связи с воздействиями через графический интерфейс. На основе полученных знаний выделяются значения параметров, отклоняющиеся от найденных закономерностей.

В рамках данной исследовательской работы разрабатывается компонент, позволяющий автоматизировать процесс создания пользовательской документации для информационных систем [6].

На данном этапе разработки созданный компонент поддерживает создание документации для CASE-системы METAS.

METAS представляет собой программную систему, позволяющую проектировать и создавать её средствами различные информационные системы (ИС), динамически настраиваемые на условия эксплуатации и потребности пользователей [7]. Система METAS использует метаданные, описывающие предметную область, для которой создается система, интерфейс пользователя и создаваемые в системе документы, бизнес-процессы и т.д. в режиме интерпретации, что позволяет создавать информационные системы без изменения программного кода приложений, а также осуществлять динамическую настройку системы в ходе ее эксплуатации. Алгоритмы компонента документирования основаны на использовании метаданных METAS, представленных в виде многоуровневых связанных списков.

В процессе исследовательской работы была разработана и реализована двухслойная структура компонента документирования. Разработанный компонент имеет структуру, показанную на рис. 1.

Реализованная схема компонента документирования позволяет разработчику задать структуру документа. Параметры представления документа подаются на вход генератора XML-файла шаблона. Использование XML-файла документации позволяет хранить данные описания ИС независимо от представления документа, то есть в дальнейшем возможен анализ данного файла и преобразование информации, находящейся в нём в конкретный вид. Например, файл Microsoft Word, HTML или Help-файл.

Разработчик документации должен иметь возможность определения структуры и содержания документации, а именно: задавать различные сочетания вложенности элементов документации, возможность добавлять или удалять элементы описания.

Для решения поставленной задачи был разработан специальный интерфейс администратора.

Структура документа представляется в виде дерева. Дерево содержания можно задавать как визуально, так и с помощью текстового описания.

Текстовое описание синхронизируется с деревом. В интерфейсе реализована функция проверки синтаксических и семантических ошибок описания структуры документа (недопустимая вложенность элементов содержания документа). Задача проверки семантических ошибок возникла по причине ограниченного числа комбинаций вложенности элементов описания.

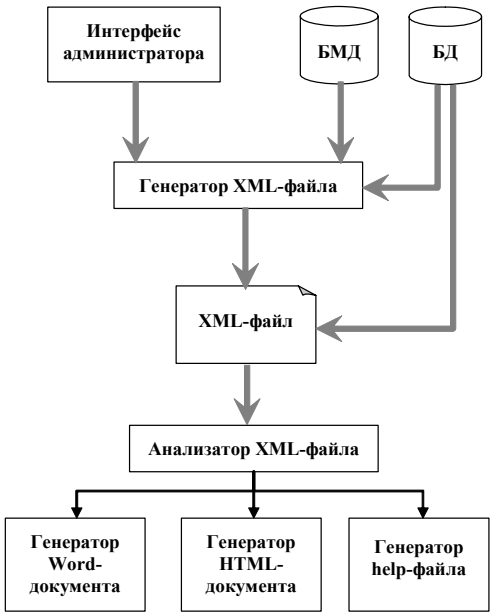


Рис. 1. Общая схема работы компонента документирования

Интерфейс передаёт информацию о структуре документа в виде XML-файла структуры генератору XML-файла.

Основная задача генератора XML-файла – создать XML-файл, содержащий описание ИС в соответствии со структурой, заданной в интерфейсе разработчика документа. Генератор обходит списки метаданных в порядке, заданном структурой документа. Каждая вершина XML-файла описывает определённый объект.

В генераторе XML-файла также реализован алгоритм поиска путей к сущности в рекурсивном дереве. Задача возникла в связи с тем, что система METAS позволяет на главной форме настроить дерево объектов (сущностей) ИС для удобного доступа пользователей к сущностям. Поэтому может потребоваться описать путь к вершине, представляющей сущность в дереве, показать, как пользователь может до неё добраться.

Для описания форм в генератор была добавлена функция «фотографирования» формы. Изображение формы сохраняется в отдельном файле, а затем в XML-файл добавляется ссылка на данное изображение.

Анализатор XML-файла выполняет разбор созданных ранее XML-файлов описаний ИС и передаёт данные описания в специализированные генераторы конечных документов, поддерживающие определённый интерфейс. Анализатор использует технологию событийного разбора документов SAX.

Каждая вершина XML-файла описывает определённый объект. Кроме атрибутов, содержащих данные описания, каждая вершина имеет специальный атрибут, значением которого является индекс соответствующей записи в таблице (рис. 2).

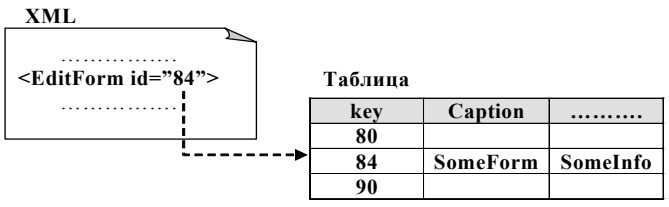


Рис. 2. Связь атрибута в XML и ключа в БД

Такой подход позволяет впоследствии при анализе XML-файла быстро находить нужную запись в БД, а также проверять на основе имеющихся описаний, тот ли это объект и существует ли он вообще. Несоответствие данных может быть следствием двух действий:

- произведена переиндексация таблиц,
- вручную изменены данные в XML-файле.

Во всех таких случаях анализатор спрашивает пользователя, откуда брать описание объекта – из XML или из ИС.

Для иллюстрации работы созданной схемы реализован генератор документа в представлении Microsoft Word. Взаимодействие с Word происходит через механизм OLE. Данные об ИС поступают непосредственно от анализатора XML-файла.

Разработанный компонент позволяет:

- представить данные ИС в универсальном виде (XML);
- разделить процессы выборки информации и создания документации;
- преобразовать данные ИС к любому представлению;
- по-разному структурировать документацию пользователя;
- поддерживать связь документации с базой данных ИС.

Пример работы компонента создания документации в соответствии с описанной ниже схемой показан на рис. 3.

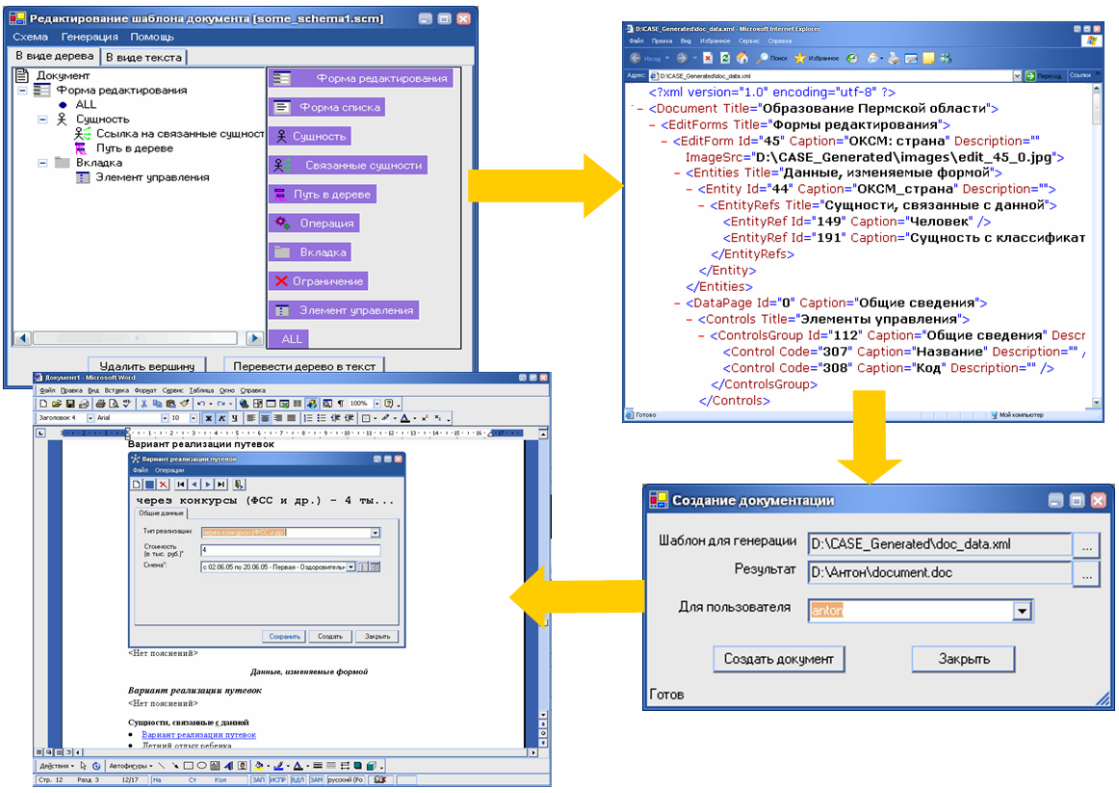


Рис. 3. Пример работы компонента создания документации

На рисунке изображён процесс создания документации, состоящий из шагов:

- создание структуры будущего документа,
- создание XML-шаблона, содержащего всю необходимую для документации информацию,
- указание параметров генерации конечного документа,
- создание конечного документа в формате Microsoft Word.

Реализованный компонент создания документации позволяет создавать XML-файлы шаблонов документов требуемой структуры, содержащие данные ИС. Также реализована возможность обработки XML-шаблонов и создания документации пользователя в формате Microsoft Word. Таким образом, имеется возможность в кратчайшие сроки создавать для информационной системы документацию различной структуры и представления.

В настоящее время исследуется возможность применения в генераторе шаблонов изменяемой грамматики. Алгоритм обработки элементов метаданных при построении документа меняется в зависимости от смысла отношений элементов метаданных. В ходе выполнения работы была поставлена задача представления на декларативном уровне правил обработки сочетаний элементов метаданных. Требуется предоставить разработчику возможность изменять данные правила. При этом информация о связях элементов метаданных представляет собой метаметаданные, или метаданные второго уровня. Такой подход позволяет реализовать компонент в виде отдельного модуля с возможностью подключения к любой информационной системе, основанной на списках. Подход также предоставляет более широкие возможности для приведения структуры документации к требуемому виду.

Разработана математическая модель применения изменяемых грамматик в алгоритме документирования, формально доказана обоснованность принятых проектных решений.

Заключение

Предложенный метод тестирования позволяет в автоматизированном режиме осуществить тестирование программы с графическим интерфейсом на любой стадии разработки благодаря возможности выбора контролируемых параметров программы. Кроме того, существует возможность проведения регрессионного тестирования, если при тестировании система будет сохранять выполняемые тесты.

Структура компонента документирования обеспечивает дальнейшее расширение возможностей. В ходе исследований появились также новые идеи построения логики работы компонента документирования.

Реализованный компонент планируется использовать для создания документации пользователей для различных информационных систем, а также для создания «заготовок» документов, которые впоследствии могут быть доработаны специалистом по созданию документации.

Библиографический список

[1] Дастин Э., Рэшка Д., Пол Д. Автоматизированное тестирование программного обеспечения. М.: ЛОРИ, 2003. С. 15-20.

[2] Gregory M. Kapfhammer. Software testing: [Электронный документ] (http://cs.allegheny.edu/~gkapfham/research/publish/software_testing_chapter.pdf).

[3] Дейкстра Э. Программирование как дисциплина математической природы: [Электронный документ] (http://khpi-iip.mipk.kharkiv.edu/library/extent/dijkstra/pp/ewd361.html).

[4] Калинов А.Я., Косачёв А.С. Автоматическая генерация тестов для графического пользовательского интерфейса по UML диаграммам действий: [Электронный документ] (http://www.citforum.ru/SE/testing/generation_uml).

[5] Суясов Д.И., Шалыто А.А. Автоматическое документирование программных проектов на основе автоматного подхода: [Электронный документ] (http://is.ifmo.ru).

[6] Цыбин А.В. Автоматическая генерация документации пользователя в информационных системах, управляемых метаданными // Сб. тезисов конференции-конкурса «Технологии Microsoft в теории и практике программирования» / Новосибирск: НГУ, 2007. С. 78-80.

[7] Лядова Л.Н., Рыжков С.А. CASE-технология METAS // Математика программных систем: Сб. науч. тр. / Пермь: Перм. ун-т, 2003. С. 4-18.

Сведения об авторах

Антон Цыбин *– Пермский государственный университет, студент магистратуры кафедры математического обеспечения вычислительных систем; Россия, г. Пермь, 614990, ул. Букирева, 15; e-mail: magicdr@mail.ru*

Людмила Лядова *– Пермский государственный университет, заведующий кафедрой математического обеспечения вычислительных систем; Россия, г. Пермь, 614990, ул. Букирева, 15; e-mail: LNLyadova@mail.ru*

АРХИТЕКТУРА И РЕАЛИЗАЦИЯ СРЕДСТВ РЕПОРТИНГА В ДИНАМИЧЕСКИ НАСТРАИВАЕМЫХ ИНФОРМАЦИОННЫХ СИСТЕМАХ

Вячеслав Ланин

Аннотация*: Статья посвящена описанию архитектуры и реализации подсистемы работы с запросами и отчетами в адаптируемых динамически расширяемых информационных системах. Разработанный метод характеризуется универсальностью применения, ориентацией на пользователя и возможностью интеграции с внешними информационными системами. Реализация программных средств основана на использовании многоуровневых метаданных.*

Keywords*: CASE-технология, адаптируемая информационная система, электронный документ, построитель запросов, генерация отчетов.*

ACM Classification Keywords*: D.2 Software Engineering: D.2.2 Design Tools and Techniques – Computer-aided software engineering (CASE); H.2: Database Management:: H.2.3 Languages – Report writers; H.3.3 Information Search and Retrieval – Query formulation.*

Введение

В настоящее время под термином «business intelligence» понимаются инструментальные средства для анализа данных, построения отчетов и запросов, помогающие бизнес-пользователям обрабатывать большие объемы данных для того, чтобы извлечь из них и синтезировать значимую информацию. Получение результатов обработки данных, хранящихся в базе данных (БД), хранилище данных – одно из основных назначений любой программной системы. Данные вводятся пользователями и накапливаются для последующей их обработки (системы OLTP); выполнения анализа, прогнозирования и поддержки принятия обоснованных управленческих решений (OLAP и DSS) и т.п. Результаты, получаемые на основе данных, размещенных в БД, должны быть не только определенным образом обработаны, но и визуализированы, представлены в виде документов различных форматов в зависимости от текущих информационных потребностей пользователей.

Таким образом, при построении любой информационной системы возникают задачи реализации средств создания запросов, подготовки и генерации отчетов, по возможности, не требующих навыков программирования, доступных пользователям, умеющим работать в среде современных офисных продуктов. Средства «business intelligence» являются компонентом практически каждой информационной системы, эффективность их реализации во многом определяет эффективность всей системы.

Требования к подсистеме управления документами в CASE-системе METAS

Разработка информационных систем (ИС), допускающих возможность динамической настройки на меняющиеся условия и потребности пользователей, расширения функциональности в ходе эксплуатации системы предусматривает необходимость использования специальных инструментальных средств разработки.

CASE-технология METAS (METAdata System) – это основа для создания адаптируемых информационных систем, управляемых метаданными [1]. Данная технология предназначена для снижения трудоемкости разработки корпоративных информационных систем и повышения их гибкости, масштабируемости и адаптируемости непосредственно в процессе эксплуатации. Основное отличие рассматриваемой системы от многих существующих CASE-систем, генерирующих по некоторым спецификациям, описывающим предметную область, код на каком-либо языке программирования, состоит в том, что METAS использует это описание *во время* своей работы, выполняя функции и отображая данные, определенные метаданными. Это дает возможность гибкой настройки приложения, реструктуризации описываемых объектов в процессе эксплуатации системы и создает хорошие предпосылки для создания интеллектуальной системы, которая может настраиваться на потребности пользователя. Технология METAS ориентированна на создание открытых распределенных приложений. Подсистема репортинга должна учитывать эти особенности технологии [2].

В подсистеме репортинга необходима реализация следующих функций:

- создание пользователем запросов к БД в терминах предметной области;
- генерация отчетных документов на основе разработанных пользователем шаблонов;
- передача шаблонов запросов и отчетов между узлами распределенной системы;
- интеграция с внешними системами.

Рассмотрим перечисленные выше требования более подробно с учетом специфики технологии METAS и предполагаемого характера применения ИС, построенных на основе данной технологии.

Основным требованием к системе репортинга является обеспечение интерфейса пользователя, облегчающего пользователям-непрограммистам *подготовку отчетов и запросов в привычных для пользователя терминах предметной области*. Отчеты могут быть достаточно сложными, включать элементы дополнительной обработки, анализа и визуализации сложно организованных данных. Система должна предусматривать *возможность расширения*, то есть создания новых отчетных форм и запросов. При этом от пользователя не должно требоваться владение специальными навыками программирования или знание языка SQL. Пользователь при построении отчета *должен работать в знакомой ему среде*. Из перечисленных выше требований следует что, необходимым условием реализации данного компонента является использование *развитой системы метаданных для описания предметной области*. Эти метаданные в системах, построенных на технологии METAS, уже имеются: с помощью них описывается модель предметной области ИС. Таким образом, и предметная область, и создаваемые пользователями запросы и отчеты могут описываться одними и теми же терминами.

Система должна обеспечивать хранение построенных запросов и шаблонов отчетов с возможностью их *повторного применения* и использования в качестве источника данных результатов выполнения ранее созданных запросов. Данное требование приводит к необходимости включения в структуру метаданных METAS *модели репортинга*, описывающей запросы и отчеты. В распределенной ИС необходимы развитые средства обмена данными между различными подсистемами, то есть построенные запросы и шаблоны отчетов, метаданные, описывающие их, должны быть легко *переносимы*.

Построение запросов требует *формирования сложных условий и выражений*. Как правило, для этой цели применяется так называемый «редактор выражений» – компонент, позволяющий пользователю строить выражения из некоторых базовых конструктов, обеспечивающий проверку правильности созданного

выражения. Для увеличения мощности редактора выражений необходима реализация *встроенного языка* для создания формул.

Пользователю должны быть предоставлены *инструменты позиционирования элементов отчета* при разработке шаблона. Требования к точности позиционирования элементов отчета бывают достаточно высокими, особенно когда данный отчет служит целям внешней отчетности, допускает возможность автоматизированной обработки.

Для просмотра и редактирования документов непосредственно из системы должны быть доступны средства *интеграции с внешними программными продуктами*. Особенно необходима тесная интеграция с офисными программами, системами электронной почты, являющимися в организациях одним из основных средств передачи документов.

Еще одно требование – возможность настройки на *использование различных СУБД*.

Неотъемлемой частью предлагаемого решения является наличие *единого механизма обработки документов из различных источников*, *автоматизация процессов обмена данными* с различными внешними информационными системами, обеспечение *возможности импорта текстов и документов из файлов и баз данных разнообразных форматов*. Необходимо предусмотреть возможность применения средств OLAP и Data Mining для анализа данных.

Подходы к созданию отчетов в информационных системах

Все средства создания (генераторы) отчетов можно условно разделить на три категории:

1. *Встроенные средства* – генераторы отчетов, встроенные в средства разработки, электронные таблицы и настольные СУБД.
2. *Специализированные средства* – генераторы отчетов, выпущенные в виде отдельных приложений (как правило, они достаточно универсальны, предназначены для работы с различными БД и категориями пользователей).
3. *«Нетрадиционные» средства* – приложения, которые, с одной стороны, имеют высококачественные средства управления печатью документов или конвертирования их в различные форматы, а с другой стороны, являются серверами автоматизации, предоставляющими доступ к этим возможностям с помощью своих объектных моделей.

Рассмотрим каждую из категорий более подробно и приведем примеры программных продуктов, являющихся наиболее типичными представителями данных категорий.

Инструменты класса *встроенных средств* (Rave Reports, SQL Server 2005 Reporting Services, ReportBuilder) обычно обладают меньшими возможностями, нежели средства создания отчетов других классов. Встроенные средства создания отчетов предоставляют достаточно богатые возможности по представлению информации, но от пользователя в большинстве случаев требуется знание языка SQL для построения запросов, пользователь также должен разбираться в структуре таблиц БД, к которой строится запрос. Все это накладывает достаточно жесткие требования на уровень квалификации пользователя и ограничивает сферу применимости данных средств. Безусловно, мы можем заранее заложить в систему шаблоны стандартных отчетов, созданные с помощью встроенных средств, можно также предусмотреть определенные возможности по настройке данных шаблонов. Однако вряд ли таким образом можно предусмотреть все потребности, которые могут возникать у пользователей в ходе эксплуатации ИС.

Кроме встроенных средств на рынке программного обеспечения представлено несколько продуктов, относящихся к *специализированным средствам* создания отчетов для разных категорий пользователей (специалистов, занимающихся подготовкой отчетов, рядовых пользователей, разработчиков приложений, разработчиков Web-сайтов и др.). Нередко один и тот же продукт существует в нескольких редакциях, ориентированных на различные категории пользователей. Специализированные средства создания отчетов обычно характеризуются поддержкой различных механизмов доступа к данным и наличием

мастеров и визуальных инструментов, ориентированных в первую очередь на пользователей-непрограммистов. Также следует отметить наличие встроенных средств деловой графики, интеграции с различными офисными приложениями и поддержки публикации данных в Internet. Для опытных пользователей и программистов предоставляются средства интеграции с наиболее популярными средствами разработки, встроенные языки для создания формул, возможности создания сложных аналитических отчетов. Безусловными лидерами на рынке специализированных средств создания отчетов являются продукт Seagate Crystal Reports фирмы Seagate Software (Crystal Decision) и комплекс продуктов фирмы BusinessObjects.

Нередко в качестве средств создания отчетов применяются приложения, которые имеют высококачественные средства управления печатью документов или конвертирования их в различные форматы. Наиболее часто в качестве таких генераторов отчетов используются приложения Microsoft Office или другие подобные продукты. В этом случае, как правило, при создании отчета возможность спроектировать его макет с помощью визуальных средств отсутствует, поэтому чаще всего такой способ генерации печатных документов используется не рядовыми пользователями, а программистами, создающими решения на базе Microsoft Office. Приложения, являющиеся частью таких решений и «заставляющие» сервер автоматизации выполнять те или иные действия (в данном случае создавать или печатать документ требуемого формата, содержащий запрашиваемые данные), называются контроллерами автоматизации. К преимуществам создания отчетов с помощью Microsoft Office относится определенная гибкость при выборе механизма доступа к данным: можно применять как механизмы доступа к данным самого Office, так и механизмы доступа к данным, которые поддерживаются средством разработки, применяемым при создании такого приложения. Немаловажным преимуществом использования средств MS Office для генерации отчетов является также его распространенность и популярность в широких кругах пользователей.

Применение встроенных средств создания отчетов оправдано, в случае разработки системы с ограниченным количеством встроенных фиксированных отчетов. Однако эти средства требуют навыков программирования при необходимости создания новых отчетов или изменения шаблонов существующих. Это делает неприемлемым их применение в разрабатываемой системе.

Использование внешнего генератора отчетов, например BusinessObjects (BO), позволяет выполнить большинство требований. BusinessObjects имеет возможность настройки на предметную область за счет средств семантической прослойки (Юниверса), обеспечивая за счет этого работу пользователя в терминах предметной области. Основным недостатком данного продукта является его высокая стоимость, делающая невозможным его массовое применение. Кроме того, интерфейс BO достаточно сложен для понимания неподготовленного пользователя. При применении данного средства необходимо ведение двух систем метаданных: метаданных самой ИС и метаданных семантической прослойки BO, как следствие, необходимо также отслеживать непротиворечивость метаданных двух систем. Вышеперечисленное делает невозможным применение BusinessObjects в качестве средства репортинга разрабатываемой системы.

Другой известный продукт – Crystal Reports – не удовлетворяет требованиям, предъявляемым к подсистеме репортинга ИС, т.к. не обеспечивает возможность работы пользователя в терминах предметной области. В этом продукте имеется так называемый «Словарь» (Dictionary), однако он лишь позволяет задать псевдонимы для полей и таблиц БД, что недостаточно в нашем случае.

Таким образом, для реализации поставленной задачи не подходят решения, представленные в первых двух категориях. Необходимо решение из третьей категории, включающие в себя преимущества первых двух. Разработка собственного генератора отчетов – наиболее гибкий вариант в отношении удовлетворения требований. При таком подходе имеется возможность тесной интеграции с приложениями пакета Microsoft Office и остальными компонентами системы.

Данное решение обладает следующими преимуществами:

- Поскольку разрабатывается собственное средство репортинга, есть возможность *учесть все требования, предъявляемые к подсистеме репортинга*. В частности, мы можем использовать метаданные системы для обеспечения возможности работы в терминах предметной области.
- Использование Microsoft Word и Microsoft Excel в качестве графической оболочки для разработки макета отчета *позволяет исключить затраты на разработку собственного средства представления документов*. Офисные продукты Microsoft имеют развитые средства оформления документов и точного позиционирования элементов, возможность применения средств анализа, деловой графики и т.п.
- Документы данного типа являются стандартом де-факто электронного документа в России. Большинство пользователей знакомы с данной программой, что позволит *снизить затраты на обучение пользователей и обеспечит их комфортную работу*. Microsoft Office инсталлирован на большинстве пользовательских компьютеров, следовательно не будет необходимости закупки данного продукта.

Следует сделать одно замечание относительно создания сложных *аналитических отчетов*. Можно выделить два противоположных подхода к их реализации. Первый подход, применяющийся в большинстве существующих систем, состоит в том, что для добавления нового отчета требуется написание программного кода и создание запросов на языке SQL. Очевидно, что в этом случае ИС требует сопровождения профессиональным программистом, хорошо разбирающимся в предметной области данной системы, обработка данных таком подходе вынесена в программный код. Второй подход заключается в том, что при необходимости создания сложных отчетов сначала формируются с помощью реализованных средств простые отчеты, а затем используются программные продукты сторонних производителей, такие как Microsoft Excel или Word и т.д., которые позволяют реализовать дополнительную обработку полученных результатов (анализ данных, визуализация в виде диаграмм и пр.). Конечно, первым способом можно построить отчет любой степени сложности, но в силу того, что одна из основных концепций системы METAS – исключение сопровождения ИС программистами на всех стадиях существования системы, данный подход неприменим. Таким образом, необходима реализация подхода, с одной стороны, достаточно простого в использовании, не требующего в процессе эксплуатации навыков программирования, а с другой стороны, содержащего потенциальные механизмы создания сложных аналитических отчетов.

Реализация подсистемы репортинга CASE-системы METAS

Как уже было сказано выше, одно из основных требований к подсистемам создания запросов и отчетов – это возможность их разработки пользователями-непрограммистами. Такое требование может быть выполнено только за счет введения дополнительного семантического слоя, основой которого могут быть метаданные, уже присутствующие в системе. Это даtт пользователю возможность работы с данными в соответствии с терминологией, принятой в конкретной предметной области, позволит абстрагироваться от физической нормализованной структуры таблиц. В предлагаемом подходе система репортинга состоит из двух компонентов: *построителя запросов* и *генератора отчетов*.

За основу инструмента создания запроса был взят аналогичный инструмент Microsoft Access, знакомый большинству опытных пользователей. Согласно предложенной концепции пользователь выбирает сущности, участвующие в запросе, и необходимые связи между ними. Затем включает в запрос интересующие его атрибуты сущностей и другие параметры, влияющие на сортировку и группировку данных. В результате учета требований пользователя и интерпретации метаданных построитель запросов автоматически генерирует SQL-запрос к БД ИС.

Для создания отчетов разработан специальный инструмент «Менеджер отчетов». В качестве шаблонов для отчетов могут быть использованы документы Word и рабочие книги Excel. Для обеспечения обмена с

другими узлами распределенной ИС построитель имеет функции экспорта и импорта шаблонов отчетов и документов.

В соответствии с предлагаемым подходом последовательность действий пользователя для *создания нового отчета* выглядит следующим образом:

1) подготовка необходимых запросов с помощью «Менеджера запросов»;
2) подготовка шаблона офисного документа (включение в шаблон статической информации: элементов оформления, формул и диаграмм и т.д.) и его разметка (включении в шаблон информации о диапазонах, куда будут помещаться данные при генерации документа на базе данного шаблона);
3) связывание запросов и соответствующих диапазонов документа, в которые должны быть помещены результаты их выполнения;
4) сохранение полученного шаблона в базе метаданных системы для последующего использования.

Предложенный подход обладает рядом преимуществ. Во-первых, для создания нового отчета не требуется программирование и написание запросов на языке SQL. При необходимости аналитическая обработка информации может быть произведении в Microsoft Excel. Во-вторых, хранение шаблона отчета в базе метаданных (БМД) делает этот отчет частью метаданных. При таком подходе отчет совместно с запросами, на базе которых он получен, может тиражироваться между узлами распределенной информационной системы. При поступлении нового типа отчета в узлах информационной системы не потребуется обновления программного обеспечения, в отличие от традиционных систем.

Последовательность *генерации отчетов* является следующей:

1) из БМД извлекаются шаблоны отчета, на их основе создаются документы;
2) выполняются запросы к БД ИС, связанные с каждым из шаблонов;
3) происходит вставка результатов выполнения запросов в диапазоны отчета, при необходимости в нем происходят дополнительные вычисления;
4) созданный отчет сохраняется в БД как документ, он может быть распечатан или передан по сети.

Благодаря реализованной возможности хранения электронных документов в БД, сгенерированные отчеты становятся частью данных ИС. Анализ первичных отчетов за различные временные промежутки позволяет создавать новые – сводные – отчеты.

Описанные выше подходы к разработке средств репортинга могут быть реализованы в системе, имеющей архитектуру, показанную на рис. 1.

Была разработана *модель репортинга*, которая включает в себя две подмодели: подмодель запроса и подмодель отчета. Модели тесно интегрированы в общую концепцию метаданных системы, они опираются на функциональность физической и логической моделей, а также на модель безопасности. Программный интерфейс модели репортинга служит для:

- выполнения, добавления, удаления и редактирования запросов к данным ИС, формулируемым в терминах логической модели ядра метаданных, т.е. в терминах сущностей и их атрибутов;
- управления шаблонами отчетов: генерации отчета по заданному шаблону, добавления, удаления и изменения шаблона отчета.

Основное предназначение модели запроса – хранение информации о созданных пользователем запросах к данным ИС, редактирования ранее созданных запросов, возможность экспорта и импорта запросов.

Модель отчета хранит информацию об отчетах, доступных в системе, позволяет добавлять новые, изменять и удалять уже существующие отчеты. Особенностью модели является возможность непосредственного хранения шаблона отчета в метаданных.

Менеджер запросов используется для управления пользовательскими запросами. С помощью данного компонента возможны создание, редактирование, удаление и выполнение пользовательских запросов к базе данных информационной системы. «Менеджер запросов» представляет собой интерфейс для

работы с моделью отчета. Компонент «Мастер связей» предназначен для связи диапазонов шаблонов и запросов в визуальном режиме. При генерации отчета в диапазоны шаблонов отчета помещается результат выполнения связанных с ними запросов.

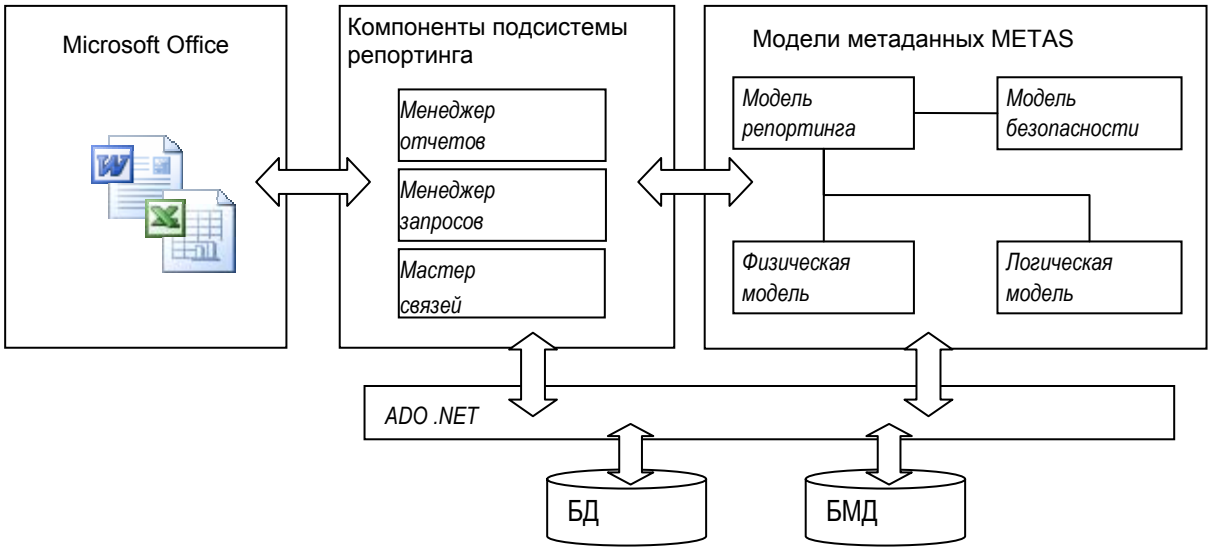


Рис.1. Архитектура подсистемы репортинга

Алгоритм генерации запроса

Как было описано выше, составленный пользователем запрос хранится в модели репортинга БМД. Данные ИС хранятся в ее реляционной БД, поэтому возникает задача трансляции запроса в терминах модели на язык запросов SQL.

Пользователь формирует свой запрос в терминах сущностей, ограничения и выражения для полей запроса он задает с помощью выражений, в которых используются термины логической модели (сущности, атрибуты и т.д.). Так как конечным результатом процедуры генерации запроса должен быть SQL-запрос к БД ИС, очевидно, что применение терминов сущностей и атрибутов недопустимо. Таким образом, необходима трансляция терминов логической модели в термины таблиц БД. Но решение такой задачи является очень сложным: требуется мощный анализатор выражений либо необходимо ограничить пользователя в сложности выражений.

Предлагается другое решение, не ограничивающее пользователя в сложности условий и не требующее сложного программирования. Источником данных будут не таблицы БД, а предварительно приведенные к первой нормальной форме представления сущностей. Логическая модель обладает средствами генерации запросов для таких представлений. Благодаря такому подходу, мы получим одинаковое представление источников данных и налагаемых на них ограничений.

С учетом модели запроса и особенностей функционирования ядра исполнительной среды CASE-системы METAS, необходимо привести запрос к следующему виду:

SELECT поля_запроса
FROM ((запрос_для_Сущности1) Псевдоним_для_Сущности1
INNER*|*LEFT*|*RIGHT JOIN (запрос_для_Сущности2) Псевдоним _для_Сущности2
    ON условие_соединения_Сущности1_и_ Сущности2, …
INNER*|*LEFT*|*RIGHT JOIN (запрос для СущностиN) Псевдоним _для_СущностиN
    ON условие_соединения_СущностиN_и_ СущностиM),
        Таблицы_Соединения_Сущностей_M:M, (…)Псевдоним_для_Сущности(N+1),…,
        (…)Псевдоним _для_Сущности(N+n),
WHERE Условия_Соединения_Сущностей_M:M, Условия_Пользователя

GROUP BY *group_by_expression*
ORDER BY *order_expression* [ASC | DESC]

Ключевыми этапами алгоритма являются следующие шаги:

1. Формирование списка полей, выводимых в качестве результата запроса.
2. Разбиение множества используемых в запросе сущностей на три подмножества: связанные с помощью внутреннего, левого или правого объединения (*JoinLinked*); связанные отношением «многие ко многим» и не входящие в первое множество (*MultiLinked*); к последнему множеству относятся несвязанные сущности (*NotLinked*).
3. Обращение к логической модели для получения текста SQL-запроса для каждой из используемых сущностей. Логическая модель сгенерирует запрос, возвращающий таблицу экземпляров сущности в первой нормальной форме.
4. Формирование раздела запроса «FROM»:
   - Объединение запросов, являющихся результатом выполнения шага 3 алгоритма, для сущностей из множества *JoinLinked* с помощью указанного типа объединения (LEFT JOIN, RIGHT JOIN, INNER JOIN), присваивание псевдонима результату подзапроса.
   - Перечисление запросов для сущностей из множества *MultiLinked*, в ходе которого присваиваются псевдонимы результатам подзапросов и добавляется имя вспомогательной таблицы, организующей связь «многие со многими».
   - Перечисление запросов для сущностей из множества *NotLinked*, с присваиванием псевдонимов результатам подзапросов.
5. Формирование раздела запроса «WHERE»:
   - Для каждой сущности из множества *MultiLinked*, добавляются условие связи двух сущностей.
   - Добавляются ограничения, наложенные пользователем на атрибуты сущностей.
6. Формирование раздела запроса «ORDER BY» путем объединения элементов сортировки.
7. Формирование раздела запроса «GROUP BY» путем объединения элементов группировки.

Описанная последовательность действий приведет к генерации SQL-запроса в терминах таблиц и полей БД ИС. Данный запрос будет произведен непосредственно к базе данных без обращения к моделям ядра метаданных системы.

Очевидно, что сформированный запрос не будет являться оптимальным по своей структуре, в силу противоречивости предъявленных к нему требований. Но, учитывая, что все современные СУБД имеют встроенные внутренние средства оптимизации запросов, и то, что скорость отклика системы некритична, предложенное решение удовлетворяет основным требованиям.

Следует заметить, что источники оптимизации запроса касаются логической модели. Сейчас логическая модель в качестве запроса для сущности возвращает ее первую нормальную форму, в которую входят все ее атрибуты, совершенно необязательные для выполнения запроса. Таким образом, объем текста запроса и скорость его выполнения можно уменьшить за счет добавления дополнительной функциональности в логическую модель.

Заключение

Предложенный подход к построению системы репортинга CASE-системы METAS позволил создать гибкую, ориентированную на пользователя систему работы с отчетами и запросами. Применение метаданных позволяет уйти от использования SQL при формировании запроса к реляционной базе данных. Реализованная подсистема репортинга стала частью комплексного подхода к управлению электронными документами в CASE-системе METAS [3].

Библиографический список

[1] Лядова Л.Н., Рыжков С.А. CASE-технология METAS // Математика программных систем: Межвуз. сб. науч. трудов / Перм. ун-т. Пермь, 2003. С. 4-18.

[2] Бакланов Д.М., Варламов А.А., Ланин В.В., Лядова Л.Н. Подсистема репортинга программного комплекса MDK METAS // Математика программных систем: Межвуз. сб. научных трудов / Перм. ун-т. Пермь, 2003. С. 19-34.

[3] Ланин В.В. Подсистема управления документами CASE-системы METAS // Математика программных систем: Межвуз. сб. науч. трудов / Перм. ун-т. Пермь, 2006. С. 135-146.

Сведения об авторе

Вячеслав Ланин *– Пермский государственный университет, аспирант кафедры математического обеспечения вычислительных систем; Россия, г. Пермь, 614990, ул. Букирева, д. 15; e-mail: lanin@psu.ru*

СРЕДА РАЗРАБОТЧИКА ИСКУССТВЕННЫХ НЕЙРОННЫХ СЕТЕЙ

Кирилл Юрков

Аннотация: *В данной статье мы рассмотрим проблему создания инструментальной среды разработчика искусственных нейронных сетей. Будут рассмотрены различные подходы к созданию инструментальной среды, вопросы проектирования объектно-ориентированной модели нейросетей и предложен программный продукт EasyNet – инструментальная среда разработчика искусственных нейронных сетей.*

Keywords: *искусственные нейронные сети.*

ACM Classification Keywords: *I.2.6 Learning – Connectionism and neural nets*

Введение

Прошло то время, когда необходимо было доказывать полезность искусственных нейронных сетей (далее ИНС), и на данный момент сформировалось ряд областей обработки информации, где применимость ИНС доказана практикой. Промышленные приложения, решающие задачи распознавания образов, прогнозирования, кластеризации и анализа данных, активно и с успехом применят достижения теории ИНС. И хотя очень важно отметить, что прикладная теория ИНС бурно развивается и ширится круг задач решаемых ею, уровень, достигнутый данной теорией, позволят выделять общие концепции, которые могут быть положены в основу программного продукта, позволяющего применять всю мощь ИНС. Но каким требованиям должен удовлетворять такой продукт? Благодаря популярности ИНС среди очень широкого круга исследователей и разработчиков, в последнее десятилетие были разработаны десятки типов ИНС и сотни алгоритмов обучения, более того новые модели продолжают появляться и нет причин считать, что в ближайшее время данная тенденция изменится. Прикладная значимость теории ИНС делает необходимым создание инструмента, который бы облегчил труд разработчиков, применяющих ИНС.

Среда EasyNet

Разработчик, планирующий использовать ИНС в рамках собственного программного продукта, как правило, требует от среды, во-первых, возможности применения всех достижений теории ИНС, во-вторых, максимального облегчения процесса экспериментирования, так как процесс подбора ИНС под задачу, остается сложным и нетривиальным, и, в-третьих, возможность работать с системой в терминах

теории ИНС. Отдельным требованием является также визуальный интуитивно понятный интерфейс для создания и редактирования ИНС. На данный момент разработчик вынужден:

- программировать ИНС, пользуясь универсальными языками программирования, что делает процесс поиска оптимальной сети непозволительно длительным;
- моделировать ИНС универсальными средствами моделирования такими, как Stratum или MatLab, которые не позволяют работать с системой в терминах ИНС и не обладают необходимых набором встроенных примитивов для работы с ИНС;
- пользоваться профессиональными нейропакетами, которые часто не обладают достаточной гибкостью и в связи с высокой стоимость малоэффективны при разовом использовании.

Отметим, что даже среди профессиональных нейропакетов не существует программного продукта, который в полной мере удовлетворял бы нуждам разработчика-экспериментатора.

Проанализировав требования к современной инструментальной среде разработчика ИНС, а также недостатки и преимущества существующих средств, нами была спроектирована и реализована среда EasyNet, позволяющая

1. создавать ИНС как полностью под управлением разработчика, так и с помощью встроенных мастеров,
2. манипулировать сетью на нейронном уровне с помощью удобного визуального интерфейса;
3. применять не только алгоритмы обучения сети, но и алгоритмы оптимизации топологии;
4. динамически расширять набор поддерживаемых ИНС, алгоритмов обучения и алгоритмов оптимизации ИНС;
5. заносить в журнал данные о проводимых экспериментах;
6. использовать сохраненные в журнале данные для воспроизведения ранее проведенных экспериментов.

Для реализации было применено объектно-ориентированное моделирование ИНС на уровне нейронов, что позволило создать наиболее гибкую систему, предоставляющую разработчику максимум возможностей при проектировании ИНС под конкретную задачу. Подобный подход позволяет в рамках удобной визуальной среды добавлять, редактировать и удалять не только целые слои ИНС, но и отдельные нейроны, что позволяет, например, изменять передаточные функции у выбранных нейронов, а не у всего слоя сразу. Реализованы мастера сетей и слоев, дающие возможность создавать готовую к обучению ИНС за несколько секунд. Поддержка алгоритмов оптимизации позволяет автоматизировать подбор оптимальной ИНС под задачу в рамках, определенного типа сети.

Моделирование ИНС на уровне нейронов оставляет открытым вопрос эффективности по времени алгоритмов обучения ИНС. Однако благодаря тому, что алгоритм обучения сети, являясь отдельным компонентом, применим только, для определенного набора сетей, в его рамках возможна реализация перехода от объектно-ориентированного подхода к матричному, что позволяет сделать процесс обучения менее трудоемким.

Разработчику ИНС предоставляются библиотеки базовых классов, а также описание их интерфейса. Создавая собственные классы, наследующие от базовых, разработчик имеет возможность вносить в систему новые типы сетей, алгоритмов обучения и оптимизации, слоев, нейронов и даже передаточных функций и сумматоров. Благодаря применению платформы .Net, разработчик имеет возможность выбрать из целого ряда современных языков программирования тот язык, который лучше подходит для реализации нового компонента. Для того чтобы встроить разработанный компонент в систему, достаточно занести информацию о нем в базу метаданных системы, что может быть сделано с помощью самой среды. Информация о компоненте содержит путь к DLL библиотеке, содержащий код для данного компонента, имя конструктора и его параметры, тип параметра и значение по умолчанию, вопрос, который должен быть задан пользователю, для того чтобы он ввел значение данного параметра. Для поддержки мастеров сетей и слоев, соответствующие компоненты должны иметь по две записи конструкторов в базе

метаданных: о базовом конструкторе и о конструкторе для мастера. Система, работая с базой метаданных, определяет набор доступных компонент и по запросу пользователя в режиме диалога создает объект запрашиваемого типа.

На данном этапе были реализованы сети типа многослойный персептрон и Кохонена, а также алгоритмы обучения для данных сетей, алгоритм генетической оптимизации для многослойных персептронов, целый ряд различных типов слоев, нейронов, передаточных функций для каждой из сетей. В дальнейшем планируется постепенно расширять число компонент.

Важной частью системы является база данных экспериментов, куда сохраняется информация о применяемой сети, алгоритмах обучения, обрабатываемых данных и погрешности сети на этих данных. Предоставляется возможность просмотра журнала и импортирования решений предыдущих экспериментов. В силу того, что разработчик, как правило, работает со схожими задачами, со временем, извлекая знания из базы данных экспериментов, возможно создание экспертной системы по подбору ИНС под задачу из данной проблемной области, что позволит полностью автоматизировать труд разработчика-экспериментатора.

На рисунке 1 представлена схема разработанной среды EasyNet. Вполне очевидно, что база данных экспериментов может переноситься отдельно от среды. Таким образом, с течением времени разработчики из разных областей применения ИНС, смогут заполнить свои базы данных экспериментов. Объединяя их, и извлекая знания из полученной обобщенной базы данных, возможно, создать экспертную систему для подбора ИНС под задачу.

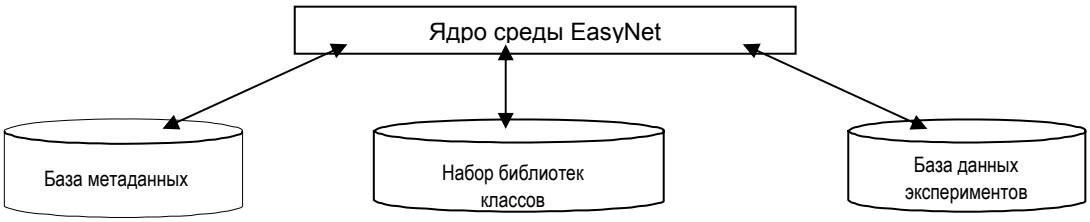


Рис. 1. Инструментальная среда EasyNet

Важной особенностью EasyNer является ее удобный визуальный интерфейс, позволяющий в полной мере воспользоваться плюсами нейронного подхода. Графический интерфейс, представленный на рисунке 2, поддерживает как стандартные операции тапа масштабирования, перетаскивания объектов, работы с выделенной группой объектов, так и специфичные, например, добавить нейрон, протянуть связь и т.д.

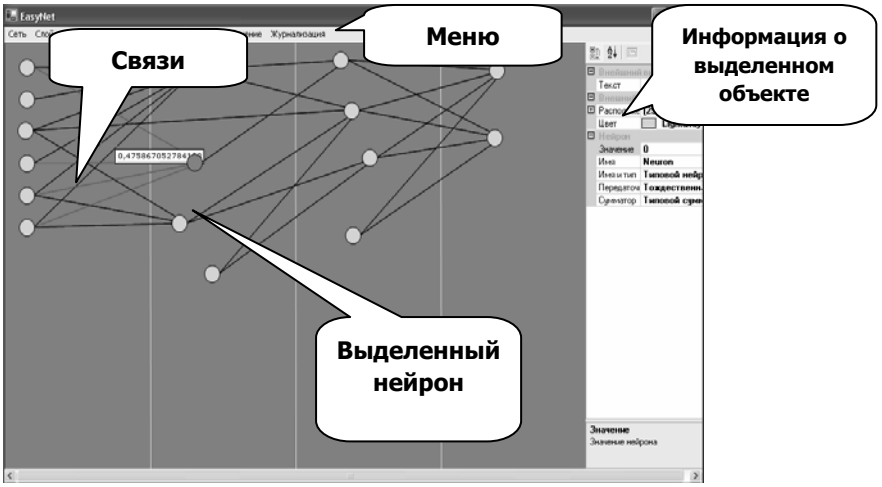


Рис. 2. Интерфейс инструментальной среды EasyNet

Применение EasyNet в учебном процессе

Преподавание дисциплин ИИ связано с целом рядом проблем. В частности при обучения студентов работе с ИНС важно дать возможность студентам применять типичные ИНС и создавать собственные. Для облегчения понимания необходимо предоставить средство позволяющее сделать работу с ИНС как можно более наглядной. EasyNet как нельзя лучше подходит на роль инструмента используемого как на этапе ознакамления с азами теории ИНС, так и на этапе создания собственной ИНС. В частности Easynet позволяет:

- Облегчить преподавание основ ИНС и наглядно продемонстрировать процесс конструирования и обучения ИНС
- Максимально ускорить ознакомление с азами теории ИНС и перейти к практике решения конкретных задач
- В процессе обучения учесть уровень подготовленности пользователя
- Контролировать выполнение работ обучающимся (посредствам журнализации)
- Ознакомиться со спецификой создания новых типов сетей и алгоритмов обучения для уже существующих типов ИНС
- Облегчить использование результатов проведенных экспериментов для решения новых задач

Замечания по дальнейшему развитию системы

На текущий момент среда EasyNet представляет собой законченное приложение, однако в перспективе планируется расширить ее возможности. В частности, будет реализована возможность экспортировать полученные сети в виде DLL библиотек, будет расширяться число компонент. Также планируется расширить возможности среды путем внедрения модуля анализа база данных экспериментов и, в перспективе, генерации из нее экспертной системы, позволяющей помочь разработчику в выборе конкретной ИНС под задачу.

Для облегчения применения EasyNet в учебном процессе предполагается также добавить возможность создания макросов – наборов поименованных действий в системе. Таким образом, будет предоставлена возможность создания упражнений на основе среды, а также демонстраций методов их решения.

Заключение

В данной статье была рассмотрена современная инструментальная среда разработчика ИНС EasyNet. Была продемонстрировано, что внедрение данной среды позволит сделать труд разработчика ИНС эффективнее, за счет того, что

1. Среда предоставляет удобный визуальный интерфейс для создания и редактирования ИНС.
2. Среда отличается гибкостью, что делает возможным создание, практически любой сети на базе, существующих компонентов.
3. В случае если, существующих компонентов не достаточно, разработчик всегда может расширить среду, добавив собственные компоненты.
4. Существует возможность, переложить на «виртуальные плечи» среды наиболее рутинную работу по подбору оптимальной топологии (числа слоев, нейронов в слое и т.д.).
5. Разработан и реализован механизм журнализации, позволяющий сохранять и применять информацию о проводимых экспериментах и применяемых сетях.

Authors' Information

Кирилл Юрков *– Пермский Государственный Университет, студент; Россия, 614990, Пермь, ул. Букирев, д. 15; e-mail: forfin@mail.ru*

HOW TO USE A DESKTOP VERSION OF A DBMS FOR CLIENT-SERVER APPLICATIONS

Julian Vasilev

***Abstract**: DBMS (Data base management systems) still have a very high price for small and middle enterprises in Bulgaria. Desktop versions are free but they cannot function in multi-user environment. We will try to make an application server which will make a Desktop version of a DBMS open to many users. Thus, this approach will be appropriate for client-server applications. The author of the article gives a concise observation of the problem and a possible way of solution.*

***Keywords**: Database management systems (DBMS), Information technology, parallel processing, Cache, client-server applications, application server, sockets.*

***ACM Classification Keywords**: H.2.8 Database Applications, H.4 information systems applications.*

Introduction

Single user versions of some DBMS are also called Desktop versions. They are usually free for commercial, home and office use. We can give Intersystem's Cache as an example [1]. The license fee for the use of the multi-user version of this DBMS is 245 EUR per process excluding VAT (Value Added Tax). If we calculate it with VAT then price is 294 EUR. If a company buys accounting software for 200 EUR and wants to use it as a client-server application for 4 computers, it has to pay 200 EUR for the software and 1176 EUR for license fees just for the right to use the multi-user version of the DBMS. This fact obstructs many small and middle enterprises in buying software products. That is why we have a possible solution. We will build an application server which will receive queries from workstations and redirect them to a single-user database. After receiving the answer from the database it will be redirected to the appropriate workstation. In this way, end users cannot feel that they use a single-user database. Moreover there is no need in changing the existing software, for instance the accounting software.

Background of the problem

Form a technological point of view this idea can be realized in Delphi, C#, Visual Basic. Moreover, there is a version of Delphi for Linux, called "Kylix". In this way the future application server can be compiled for another operating system. The program implementation consists of two parts: a server application and a client application. According to Cantu [2, 749] the idea can be realized by using the communication interface DCOM.

"DCOM is directly available in Windows NT/2000 and 98/Me, and it requires no additional run-time applications on the server. You still have to install it on Windows 95 machines. DCOM is basically an extension of COM technology that allows a client application to use server objects that exist and execute on a separate computer. The DCOM infrastructure allows you to use stateless COM objects, available in the COM+ and in the older MTS (Microsoft Transaction Server) architectures. Both COM+ and MTS provide features such as security, component management, and database transactions, and are available in Windows NT/2000 and in Windows 98/Me. Due to the complexity of DCOM configuration and of its problems in passing through firewalls, even Microsoft is abandoning DCOM in favour of SOAP-based solutions."

We made several experiments and few basic problems occurred. First, computers with different version of operating systems (for instance Windows 98 and Windows XP) cannot communicate. Second, DCOM does not keep alive several connections. Third, there is a limit in the number of connections. That is why our research continues. We want to find a better solution. Let us examine the work of a multi-user database (fig. 1).

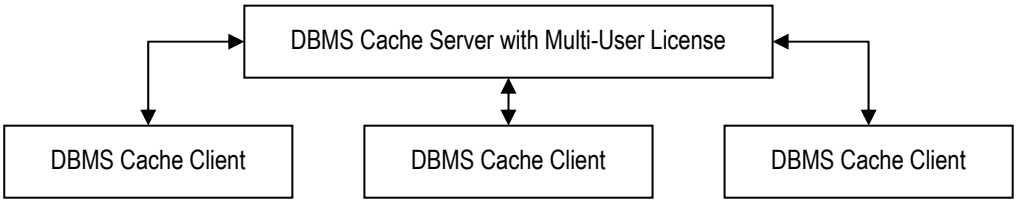


Figure 1: Technology of using a multi-user DBMS

The server part is usually installed on a computer, named "server" and the client part of the DBMS – on workstations. When we use a different DBMS the installation process is similar.

A possible solution

We have to use another information technology to solve this problem. The communication can be realized using Transmission Control Protocol/Internet Protocol (TCP/IP for short). In a local area network (LAN) each computer has a unique IP Address. Connections between computers can be implemented by TCP ports. Each TCP connection takes place though a port. Some TCP ports have a standard usage for specific high-level protocols and services. In other words, you should use those port numbers when implementing those services and stay away from them in any other case. Here is a short list (table 1).

Table 1 Ports for some protocols

Protocol	Port
HTTP (Hypertext Transfer Protocol)	80
FTP (File Transfer Protocol)	21
SMTP (Simple Mail Transfer Protocol)	25
POP3 (Post Office Protocol, version 3)	110
Telnet	23

HTTP, SMTP and FTP are standard protocols. If we want a custom communisation between server and workstations we have to define custom protocol. A set of communication rules is generally indicated as a protocol. Basically, the server can receive different requests and, depending on the type of request and whether it can be accomplished, replies to the client. The server will respond to many requests. Transfer protocols are at a higher level than transmission protocols. That is why protocols are independent not only from the operating system and the hardware but also from the physical network. Communication can be started only if we launch a server program which accepts client connections. The client requests a connection indicating the server it wishes to connect to. When the client sends the request, the server can accept the connection, starting a specific server side socket, which connects to the client-side socket.

Methodology of implementation

Delphi 5 ships with three sets of socket components. Newer versions of Delphi also support Socket components. They can be used to read and write information over a TCP/IP connection. The Internet page of the palette hosts the Client Socket and Server Socket components. Sending text to server can be done by issuing method "SendText".

*ClientSocket.Socket.SendText('Select * From Customers Where CustNo = 1394')*

In this way we can send to the server a SQL (structured query language) statement. The server will receive the sent message as simple text. The Server Sockets reads the text by calling the method "Client Read". The text is actually contained in the property "Receive Text".

```
SQL_to_execute := ServerSocket.Socket.ReceiveText;
```

The server can use blocking or non-blocking connections. When the server uses blocking connections requests are processed in sequence. Huge information systems cannot scale by blocking connections. One of the possible solutions is the use of non-blocking connections. If we build a large system a good idea is to use threads to communicate with the database.

For the starting of the server we have to do the following:

```
ServerSocket.Port := 1974;
ServerSocket.Active := True;
```

On the client side, to connect to the server we have to connect to the server:

```
ClientSocket.Port := 1974;
ClientSocket.Host := '192.168.23.117'; // This is the IP address of the server
ClientSocket.Active := True;
```

As we mentioned we send a SQL statement to the server.

```
ClientSocket.Socket.SendText( SQL_to_execute );
```

The server receives request. This event fires the method OnClientRead of the Server Socket.

```
procedure TForm1.ServerSocket1ClientRead(Sender: TObject;
  Socket: TCustomWinSocket);
var
  i:integer;
  st : string;
begin
  for i := 0 to ServerSocket.Socket.ActiveConnections-1 do
  begin
    with ServerSocket..Socket.Connections[i] do
    begin
      st := ReceiveText;
      if st <> '' then
      begin
        Memo1.Lines.Add(RemoteAddress +  sends:');
        Memo1.Lines.Add(str);
      end;
    end;
  end;
end;
```

In this way received text is added in a Memo.

We have to redirect it to a database. We can use ADOConnection and a Windows UDL file to access different types of databases (for instance MS Access, Oracle, MS SQL Server, Informix, Sybase, Cache, DB2 and others).

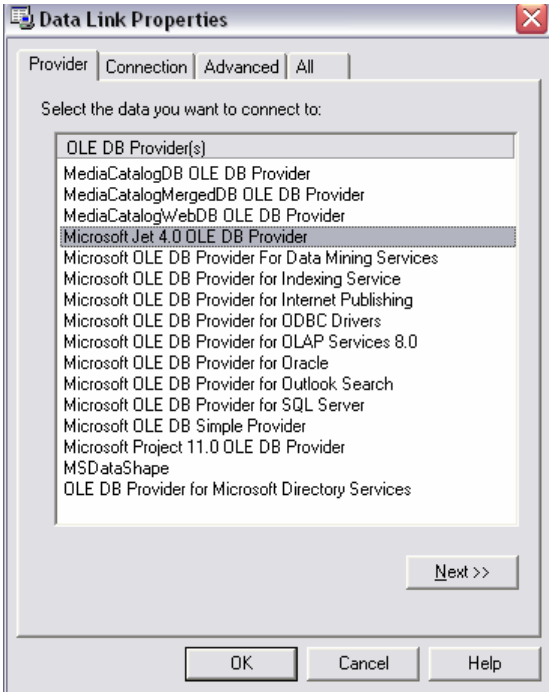


Figure 2 Provider for Data Link Properties

To make an UDL file we just make a simple text file and we change its extension from "txt" to "udl". We open the file. On the "Provider" tab we set the type of DBMS we want to use (fig. 2).

The marked choice is for MS Access. If we use Oracle, we have to choose Microsoft OLE DB Provider for Oracle. If we use MS SQL Server, we have to choose Microsoft OLE DB Provider for MS SQL Server. In the next step we choose the database, username and password for login and we can test the connection. In this way our server

application is DBMS independent. Moreover, the connection to the database is initialized through a text file which looks like an “ini” file. It is a stand-alone file and can be modified without the need of compilation.

As we highlighted the server receives requests. They are redirected to a DBMS using “ADOConnection” and “ADOQuery”. The result of the execution of a query is in the form of DataSet. This dataset is two-dimensional. It consists of rows and columns. To be send back to the server it has to be represented as a simple string. That is why we have to use 2 delimiters: one - for rows and another one – for columns. They can be “tab character” – ASCII code “9” for field delimiter and “line feed and carriage return” – ASCII codes “10”, followed by “13” – for record delimiter (table 2).

Table 2 Simple dataset returned as a result of execution a query

Order_number	Order_date	Cust_code
12345	03.04.2007	1395
12336	04.04.2007	1391

This tabular data will be transformed in one string as follows:

Result_string := ‘12345’+#9+’03.04.207’+#9+’1395’+#10#13+‘12346’+#9+’06.04.207’+#9+’1391’;

Actually, the result string is formed by using two “for” cycles. The sample code is too simple. That is why we skipped it. The next step is sending the result dataset to the client.

ServerSocket.Socket.Connections[nConnection].SendText(Result_string);

The variable “nConnection” indicates the unique number of connection. The client socket receives the result dataset. The Client Socket fires the event “OnRead”. To read the incoming message from the server we have to write the following:

Received_text := socket.ReceiveText;

The next step is to convert the string into a two-dimensional array in order to visualize the dataset in a tabular format. This operation is simple. That is why we go on. A corner-stone can be the size of the string send over socket connection. A possible solution is to send the result dataset in several parts. OLE fields are another corner-stone. If we have to send a file or multimedia fields or BLOBs (binary large objects) we can use streams over socket connections.

Software development in brief

To realize the idea of using a desktop DBMS in a local area network (multi-user environment) we need to do the following. Firstly, we have to use the technology of communication by using the built-in port in Windows. For instance Trojan horses and some DBMS (Cache uses port 1972) use them for communication. Likewise, web servers use this port to communicate with end users. The default port is 80. This fact determines the use of sockets. We can use Client Sockets and a Server Socket. Secondly, we need an application server which will stay one level above the DBMS and will dispatch queries. Its kernel is based upon a Server Socket. Thirdly, a small application is needed on workstations with built-in Client Socket. Fourthly, we need a standard for communication between end-users’ workstation and the application server. An example of such organization of work is described in fig. 3.

In another aspect this application server can be used as a dispatching server for cluster applications. This server can have several subordinate sub-servers which process users’ queries. In this way we get better parallel processing of queries. The result is similar to that in search engines such as google.com, yahoo.com, live.com, ask.co.uk.

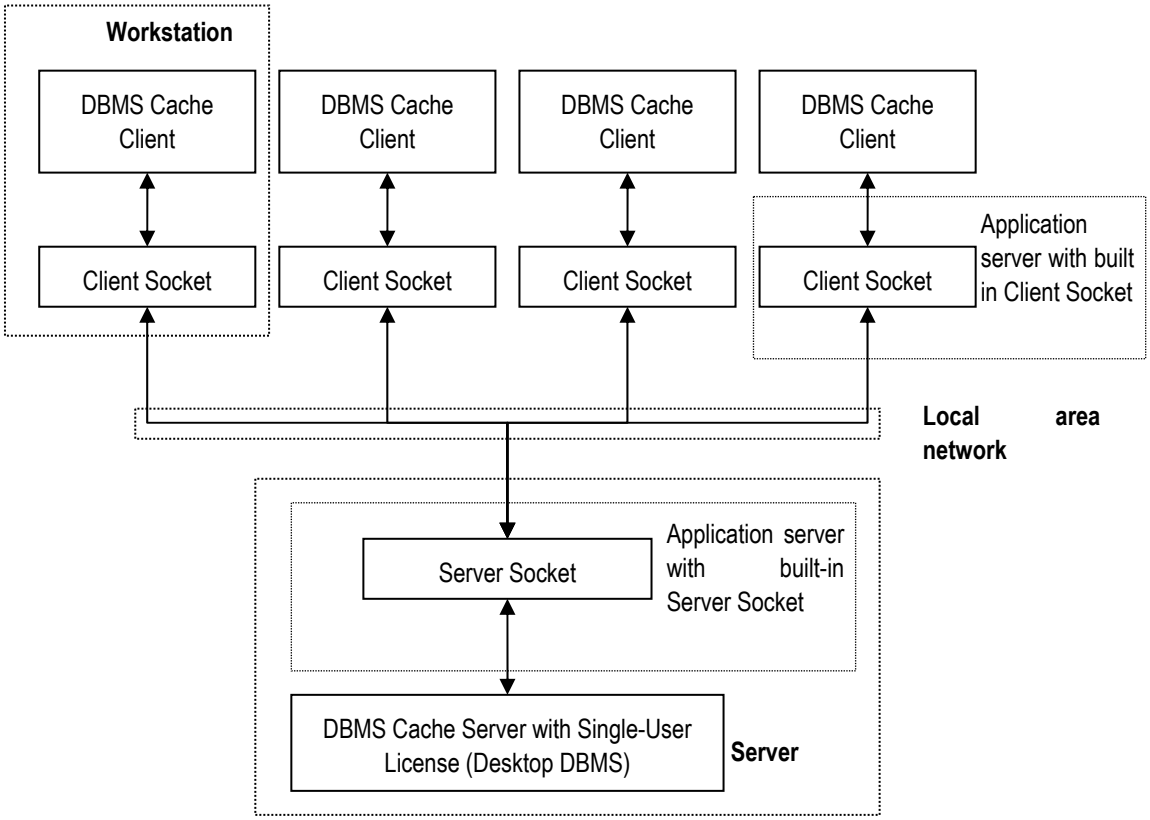


Figure 3: Technology of using a single-user DBMS in multi-user environment

Conclusions

As the reader sees, we succeeded in using a single user DBMS in multi-user environment. We gave a methodology of implementation of this idea. It was realized in a software product installed in 2005 at the Varna University of Economics – graphic user interface software for testing students' knowledge. The system is used in 7 disciplines. The achieved positive results are obvious evidence for the positive aspects of issued concepts. Moreover, software companies in Bulgaria can easily adapt this idea and make their products accessible to many small and middle enterprises.

Bibliography

[1] www.e-dbms.com.

[2] Cantu, M. Mastering Delphi 6, Sybex, Alameda, CA, 2001.

Authors' Information

Julian Vasilev – *Chief assistant professor, Department of Informatics, Varna University of Economics; 77, Kniaz Boris I str.; Varna; Bulgaria; e-mail: vasilev@ue-varna.bg*

Advanced Technologies

MATHEMATICAL MODEL AND SIMULATION OF A PNEUMATIC APPARATUS FOR IN-DRILLING ALIGNMENT OF AN INERTIAL NAVIGATION UNIT DURING HORIZONTAL WELL DRILLING

Alexander Djurkov, Justin Cloutier, Martin P. Mintchev

Abstract: *Conventional methods in horizontal drilling processes incorporate magnetic surveying techniques for determining the position and orientation of the bottom-hole assembly (BHA). Such means result in an increased weight of the drilling assembly, higher cost due to the use of non-magnetic collars necessary for the shielding of the magnetometers, and significant errors in the position of the drilling bit. A fiber-optic gyroscope (FOG) based inertial navigation system (INS) has been proposed as an alternative to magnetometer -based downhole surveying. The utilizing of a tactical-grade FOG based surveying system in the harsh downhole environment has been shown to be theoretically feasible, yielding a significant BHA position error reduction (less than 100m over a 2-h experiment). To limit the growing errors of the INS, an in-drilling alignment (IDA) method for the INS has been proposed. This article aims at describing a simple, pneumatics-based design of the IDA apparatus and its implementation downhole. A mathematical model of the setup is developed and tested with Bloodshed Dev-C++. The simulations demonstrate a simple, low cost and feasible IDA apparatus.*

Keywords*: Mathematical Modeling, Measurement-While-Drilling, In-Drilling Alignment*

ACM Keywords*: Mathematical Modeling*

List of Abbreviations

BHA	Bottom-hole assembly	INS	Inertial Navigation System
FOG	Fiber-optic gyroscope	MWD	Measuring-while-drilling
IDA	In-drilling alignment	ZUPT	Zero velocity update
IMU	Inertial Measurement Unit		

Nomenclature:

a	Orifice area (m^2)	Pa	Air Pressure in Chamber A
Aa	Piston area enclosing Chamber A (m^2)	P_b	Air Pressure in Chamber B
Ab	Piston area enclosing Chamber B (m^2)	R	Gas constant for air (287 J/kg/K)
c_p	Constant air pressure specific heat (1003.5 $Jkg^{-1}K^{-1}$)	$T_{a,\,b}$	Cylinder's chamber temperatures (K)
c_q	Orifice Discharge Coefficient	$T_{s,\,ex}$	Air tank temperatures (K)
c_v	Constant air volume specific heat (718.6$Jkg^{-1}K^{-1}$)	V_{da}	Chamber A dead volume (m^3)
m_a	Mass of air in Chamber A (kg)	V_{db}	Chamber B dead volume (m^3)
m_b	Mass of air in Chamber B (kg)	x	Displacement of piston
M	Combined mass of piston, piston rod and IMU (kg)	x_1	Cylinder's stroke

1. Introduction

1.1 Conventional Horizontal Drilling Techniques

Horizontal drilling features several advantages when it comes to oil exploration and production. First, it facilitates the accessibility of reservoirs in complex locations: under riverbeds, mountains and even cities [1]. Secondly, if a particular reservoir is characterized by a large surface area, but is distributed over a thin horizontal layer, a horizontal well will yield a larger contact area with the reservoir and thus lead to a higher productivity and longevity when compared to vertical ones [2]. Present applications of horizontal wells include intersecting of fractures; eliminating of coning problems in wells with gas and water coning problems; the improving of draining area per well in gas production, resulting in a reduction of the number of wells required to drain the reservoir; and providing larger reservoir contact area and enhancing injectivity of an injection well [3].

The drilling of a directional (horizontal) well begins by drilling vertically from the surface to a kick-off point at a predetermined depth. Then, the well bore is deviated intentionally from the vertical at a controlled rate. To implement this complex drilling trajectory, measurement-while-drilling (MWD) equipment, steerable setup and surveying sensors must be incorporated within the drilling assembly [4]. The drilling assembly utilizes a diamond bit and a mud turbo-drill motor installed in front of a trajectory control sub, nonmagnetic drill collars which include the magnetic surveying sensors, and a drill pipe [5], (Fig.1).

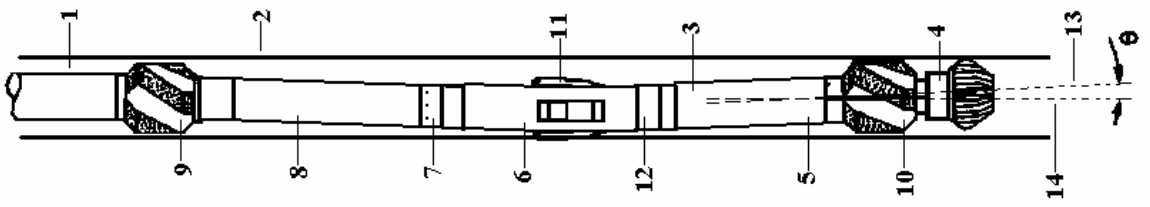


Fig.1: Drilling Assembly: 1 – drill string, 2 – borehole, 3 – bottom hole assembly (BHA), 4 – drill bit, 5 – drilling motor, 6 – trajectory control sub, 7 – bypass sub, 8 – MWD tool included in nonmagnetic collars, 9, 10 – upper and lower stabilizers for centering the drilling assembly in the borehole, 11 – stabilizer blades, 12 – induced bend to provide angular offset (θ) between the axis of the drill bit (13) and the center line (14).

1.2 Principles of Magnetic Surveying

The conventional measurement-while-drilling (MWD) surveying system presently utilizes three-axis accelerometers and three-axis magnetometers fixed in three mutually orthogonal directions [13]. At a certain predetermined surveying stations, the drilling assembly is brought to rest. At that point, the body frame of the MWD surveying system, formed by the axes of the accelerometers and magnetometers, is an angular transformation of the reference (North-East-Vertical) frame. Since the position of the bottom-hole assembly (BHA) is known, the direction and magnitude of Earth's acceleration are known as well. By comparing the acceleration vector formed from the measurements of the three accelerometers with the known vector of Earth's gravitational acceleration in the reference frame, the pitch (θ) and row (Φ) can be calculated (Fig.2) [7].

Then, the measurements from the magnetometers are combined with the calculated pitch and row to determine the azimuth angle (Ψ). The BHA trajectory is then computed by assuming a certain trajectory between the two successive stations.

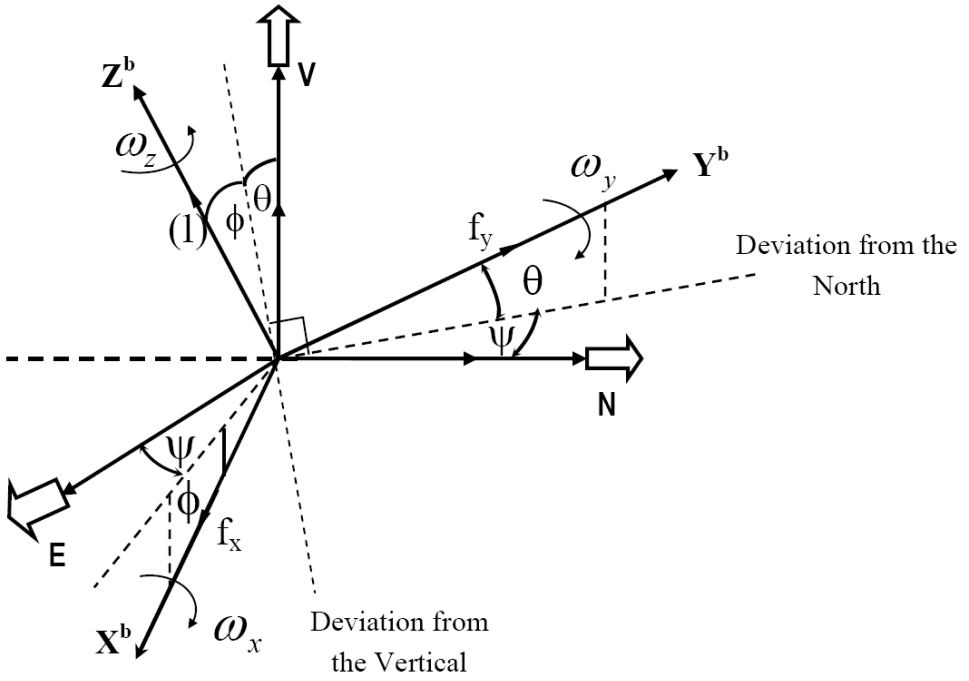


Fig.2: Orientation of the MWD magneto-surveying system with respect to North, East, and Vertical directions: the pitch (θ), the roll (Φ), and the azimuth (Ψ). In the drawing, X^b, Y^b and Z^b form the body frame, with its axes coinciding with the axes of the accelerometers and magnetometers. E, N, and V denote East, North, and Vertical and form the reference frame. The measured accelerations along the axes x, y and z of the body frame are respectively f_x, f_y, and f_z. The measured angular rates in the body frame about the x, y and z axes are respectively ω_x, ω_y, and ω_z.

1.3 Problems with MWD Magneto-Surveying System

Several external factors affect the performance of the magnetic surveying sensors. Such factors encompass the presence of randomly located ore deposits and geomagnetic influences. Moreover, the dynamic behavior of the magnetometers is negatively affected by magnetic interferences from the drill string. This requires the utilization of nonmagnetic collars for protecting the magnetic sensors. Although the accuracy of the magneto-surveying system increases with length of the nonmagnetic collars, this results in heavier and more costly MWD apparatus. Additionally, another source of error is introduced. Since the surveying sensors are located approximately 20 meters away from the drill bit, some rotations of the near-drill bit assembly may not be recognized [6].

1.4 Review of Current Inertial Navigation System (INS) -Based Navigation

In order to avoid the problems associated with magnetometers and non-magnetic collars, an INS based inertial measurement unit (IMU) incorporating a single fiber-optic gyroscope (FOG) and three-axis accelerometer has been proposed [7]. The INS determines the position, velocity and orientation of the drilling assembly in three-dimensional space by integrating the measured components of the acceleration (provided by the accelerometers) and the angular velocity (provided by the gyroscope). However, due to the small errors in the measurements of the accelerometers and the fiber-optic gyroscope, a continuous error growth in the position and the velocity of the BHA is observed [8]. Several approaches to limit this error growth have been proposed.

The first approach is based on continuous surveying with the aid of velocity and altitude updates through a Kalman filter. It has been reported that this method yields an inclination and azimuth angle errors of less than 0.4° and 1°, respectively, over a two-hour experiment. Moreover, the altitude errors have not exceeded ±0.5m over the entire experiment, while the errors along the East and North directions, dependant on the accelerometer bias, have been kept less than 50m and 20m respectively, over a two hour experiment [8].

The second approach was applied when velocity updates were not available. The approach involved the interrupting of the BHA motion at some predetermined station to apply the velocity zero update (ZUPT) for resetting the velocity errors and stopping the growth of position errors. The ZUPT approach was associated with position errors of less than 25m and 100m along the East and North directions respectively [8]. However, these results did not show substantial advantage over standard magnetic surveying.

A third method, called the In-Drilling Alignment Method (IDA), involves the induction of motion on the IMU in the horizontal North-East plane, while the entire bottom-hole assembly (BHA) is at rest. If the acceleration of the IMU at any time during the induced motion is known more precisely than the accuracy of the accelerometers on the IMU, the observations may be used as acceleration updates to align the accelerometers. Separately, an angular motion of the IMU about the axis of its gyroscope may be induced with accurately known angular rate and be used as an update for the gyroscope [9]. Such an IDA apparatus that will perform effectively in bore-hole drilling conditions has not been designed.

The aims of this paper are: (1) to design an In-Drilling Alignment apparatus for testing this newly-proposed concept; and (2) to mathematically model the expected results provided by such an apparatus.

2. Methods and Materials

2.1 Inducing Motion on the IMU in the North-East Horizontal Plane

A pneumatically-based solution is proposed for inducing a motion on the IMU in the North-East horizontal plane while the BHA is at rest. A compact, cylindrical capsule containing an IMU, RF transmitter and a small battery to power the IMU and the transmitter is attached to the end of a piston rod of a pneumatic cylinder via a bearing. The bearing allows the capsule to rotate freely around the cylinder's rod. By correctly regulating the pressure on each side of the piston, desired linear accelerations of the piston rod-IMU assembly can be obtained (Fig.3).

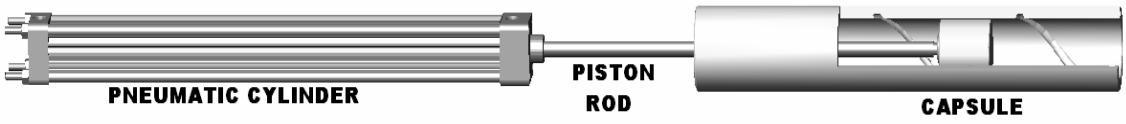


Fig.3: In-Drilling Alignment Apparatus

This linear motion can further be employed for inducing an angular motion on the IMU about the axis of one its gyroscopes. On the exterior surface of the cylindrical capsule, around its axis, ball bearings are positioned in a helical pattern. Similar helical thread is machined on the inner side of a pipe, to allow the bearings on the capsule to smoothly traverse along it. Thus, any linear motion induced on the capsule by the pneumatic cylinder will simultaneously cause an angular motion. If the linear acceleration of the IMU-containing capsule and the angular step of the helical thread are accurately known, then the angular acceleration of the capsule can be calculated easily. This in turn can be integrated to yield the angular rotation rate of the capsule (Fig.4).

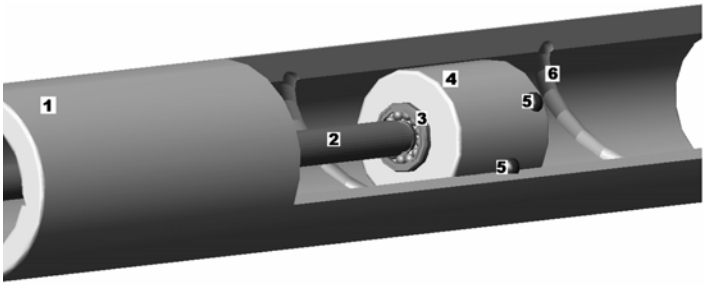


Fig.4: Schematic of the angular motion inducing mechanism: 1-pipe, 2-pneumatic cylinder rod, 3-bearing, 4-capsule enclosing IMU, battery and RF transmitter, 5-ball bearings aligned in a helical pattern over the surface of the capsule, 6-helical thread machined on the interior surface of the pipe.

2.2 Monitoring the Induced Motion of the IMU

The principle of the magnetostrictive effect is employed for monitoring the position of the piston in the pneumatic cylinder. For this purpose, the piston is equipped with tiny magnets, and a special piston position-sensing unit is installed along the cylinder (Fig.5) [11].

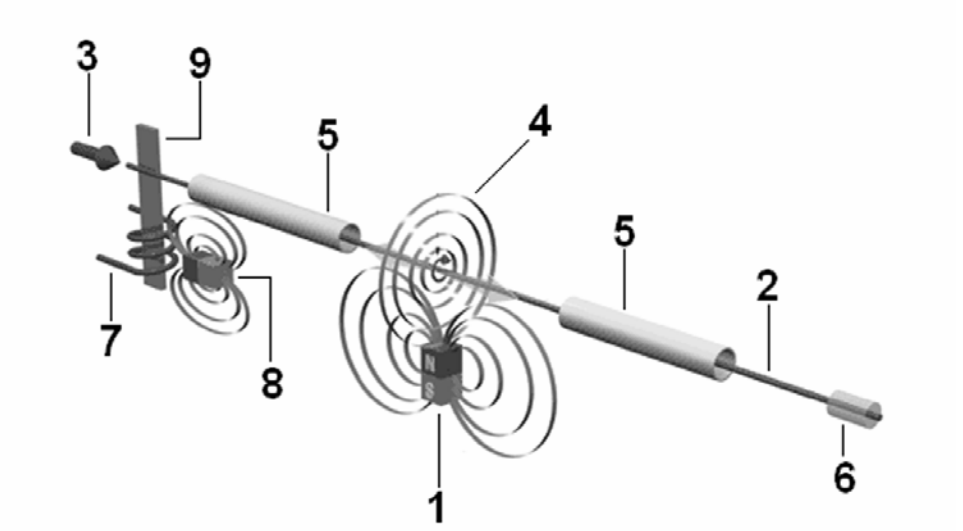


Fig. 5: Schematic of the operation of a magnetostrictive effect-based piston position sensing unit: 1-piston magnet, 2-waveguide, 3-short current pulse, 4-magnetic field around the waveguide due to the current pulse (3), 5-protective casing, 6-dampener, 7-mechanical wave detecting coil, 8-magnet providing a magnetic field in which the detecting coil is located (7), 9-strip along which the deformation wave is transmitted to the coil.

The unit consists of a “waveguide” made of a special nickel-alloy tube through which runs a copper wire. The initiation of a measurement is denoted by a short electric pulse through this wire, which sets up a circular magnetic field around it. At the point along the “waveguide” where the produced field intersects the perpendicular magnetic filed due to the magnets located in the piston of a pneumatic cylinder, an elastic deformation of the nickel-alloy tube is caused according to the magnetostrictive effect. The component of the deformation wave that traverses the “waveguide” toward its back end is dampened, while the component that arrives at the signal converter is transformed into an electric pulse. Since the travel time for the pulse is directly proportional to the position of the magnetic piston [11], by determining the elapsed time between the initiating pulse and received pulse, the piston’s position can be estimated with high accuracy in the order of 5μm [11].

Once the position of the piston is accurately known, a differentiation yields its velocity and acceleration. However, since the IMU capsule is affixed to the piston rod of the pneumatic cylinder, its linear component of motion is completely defined. Moreover, the angular rate of the IMU around the axis of the pneumatic cylinder can be calculated according to:

$$\omega = \mathbf{v} \cdot \lambda \tag{1}$$

where (ω) is the angular speed, (v) is the linear speed and (λ) is the angular step of the machined helical thread.

2.3 Pneumatic Setup of the IDA Apparatus

The following simplified pneumatic setup is proposed for inducing and controlling the motion of the IMU (Fig.6).

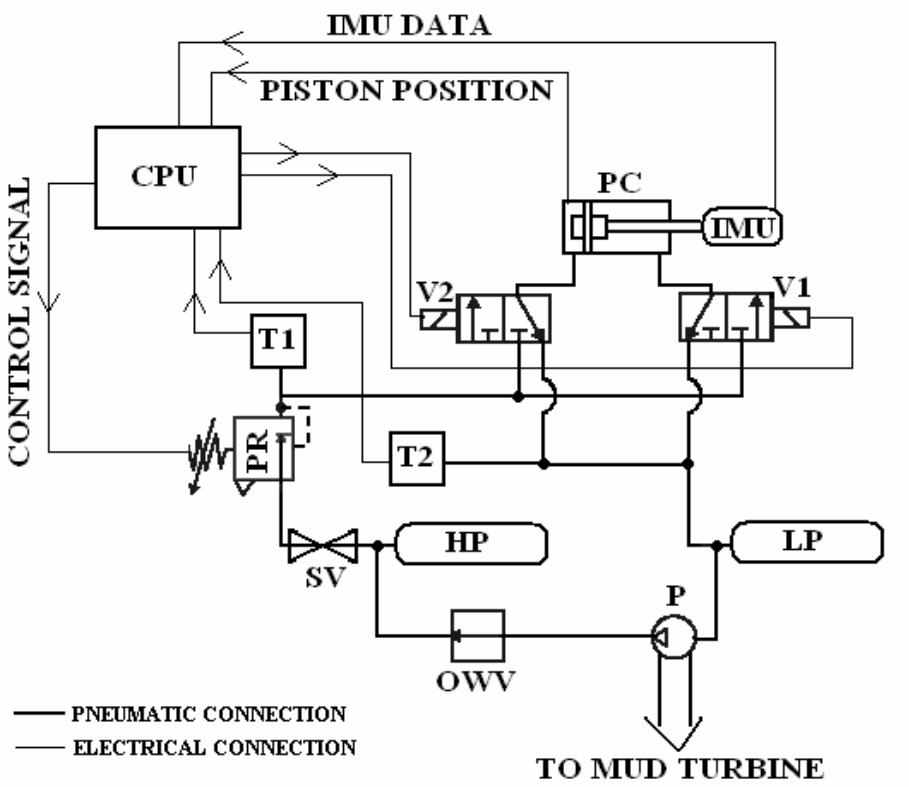


Fig.6: Pneumatic System Setup: HP-high pressure air tank; LP-low pressure air tank; PC-pneumatic cylinder, cushioned at both ends; V1, V2-two way solenoid valves; PR-proportional electric pressure regulator, T1, T2-electric pressure transducers, SV-shutoff valve; OWV-one-way air valve; P-air pump.

Initially, the system comprises a high (HP) and a low (LP) pressurized air tanks. The Central Processing Unit (CPU) can independently control the two solenoid valves (V1) and (V2) through which the pneumatic cylinder is connected to the rest of the pneumatic system. By feeding the appropriate signals to the two valves, the right chamber of the cylinder may be connected to the low-pressurized air tank, and the left to the highly-pressurized (HP) air tank via the electronic pressure regulator (PR). Then the two electric pressure transducers (T1) and (T2) inform the CPU of the air pressure in each chamber of the cylinder. Based on this information, the CPU calculates the necessary regulated pressure and controls the proportional regulator (PR). Once a pressure differential is established across the piston, a linear acceleration on the piston-IMU assembly is induced. A measurement of the piston's position is supplied to the CPU by the magnetostrictive effect-based measuring unit. The three acceleration components and angular rates measured by the IMU are also passed to the CPU where, together with the position of the piston, the data is processed mathematically to align the IMU.

Once the piston of the pneumatic cylinder is near the end of its stroke, the CPU reverses the valves (V1 and V2) and an opposite acceleration is induced. Cushions are provided on both sides of the piston to reduce the severity of the impact with the cylinder's walls.

Eventually, the pressures in the two air tanks will equalize, limiting the number of piston cycles and thus the number of alignment data points. To restart the system, the mud-powered air pump is turned on to pressurize the HP air tank to its initial high pressure. This in turn will bring the LP tank back to its original low pressure. Air is pumped from the LP tank to the HP tank through a special one-way air valve (OWV) that will prevent air from leaking back to the LP tank through the pump P. This resetting procedure is only possible when there is mud flow. Thus, it will be performed during the drilling process. The IDA process takes place when the bottom-hole assembly is at rest.

2.4 Data Manipulation and Transmission

Since the IMU is constantly in motion during the IDA process, wiring the IMU will be impractical and will result in constant stress applied to the wires. To eliminate such problems, RF link is proposed between the IMU and a local receiving module mounted on the exterior surface of the tube through which the IMU is accelerated. Thus, the three components of acceleration and angular rate measured by the IMU are sent to a local RF receiving module and then, together with the cylinder's piston position are wired to the CPU. There, the data is mathematically processed to determine the position of the BHA in the horizontal North-East frame. It is then send to the surface by the conventional method of mud pulse telemetry [3].

2.5 Mathematical Model of the Pneumatic System

To model the pneumatic system extensively, first a model of the pneumatic cylinder for its specific application will be derived. Throughout the entire model, all pneumatic processes are assumed to be adiabatic and the fluid (gas) is treated ideally. It has been shown that such assumptions still provide excellent results for similar applications, while greatly simplifying the model [10].

Let the cylinder be divided into two separate chambers A and B. Also, assume that the piston is moving to the right with speed v, (Fig.7).

The pressure change in chamber A is described by [10]:

$$\dot{P}_a = \left(\dot{m}_a - \frac{P_a A_a}{RT_s} \dot{x} \right) \frac{c_p R T_s}{c_v \left(V_{da} + \left(\frac{x_1}{2} + x \right) A_a \right)} \tag{2}$$

where m_a and P_a are the mass of gas and pressure in chamber A respectively, and A_a is the area of the piston's surface enclosing chamber A.

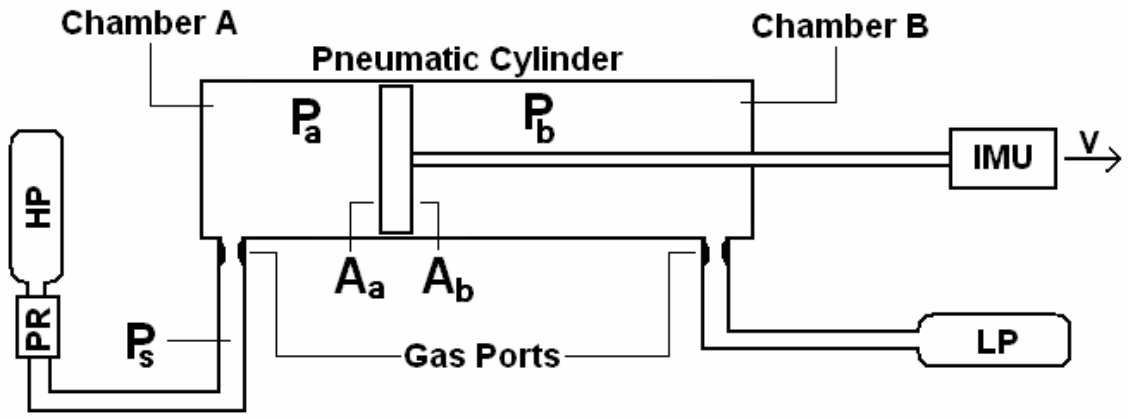


Fig.7: Supplying air to the cylinder: HP-high-pressure tank; LP-low pressure tank, PR-pressure regulator; P_a, P_b – pressure in chamber A and B respectively; P_s-supplied pressure by regulator (PR), A_a, A_b – area of piston common to chamber A and B respectively.

The position of the piston in the cylinder is denoted by x, while x_1 denotes the cylinder's stroke; Vda is the dead volume entitled to chamber A (tubing volume and unused cylinder volume). The temperature of the supplied gas is T_s, and c_p and c_v stand for the constant pressure and volume specific heats of the gas respectively; R is the gas constant. The rate of change of mass of gas in chamber A is given by [10]:

$$\dot{m}_a = \frac{c_q a P_s}{\sqrt{T_s}} \sqrt{\frac{2.8}{R(\gamma - 1)} \left[\left(\frac{P_a}{P_s} \right)^{\frac{2}{\gamma}} - \left(\frac{P_a}{P_s} \right)^{\frac{\gamma+1}{\gamma}} \right]} \quad (3)$$

In (3), c_q is the flow discharge coefficient of the pneumatic cylinder's inlet, a is the area of the inlet; and γ is the specific heat ratio. Similarly, the pressure change model for chamber B is [10]:

$$\dot{P}_b = \left(\dot{m}_b + \frac{P_b A_b}{R T_s} \dot{x} \right) \frac{c_p R T_s}{c_v \left(V_{db} + \left(\frac{x_1}{2} - x \right) A_b \right)} \quad (4)$$

where the variables correspond to the ones defined in Eq.(2), but applicable to chamber B. The rate of change of gas mass in chamber B is quantified similarly [10]:

$$\dot{m}_b = \frac{c_q a P_b}{\sqrt{T_b}} \sqrt{\frac{2.8}{R(\gamma - 1)} \left[\left(\frac{P_{ex}}{P_b} \right)^{\frac{2}{\gamma}} - \left(\frac{P_{ex}}{P_b} \right)^{\frac{\gamma+1}{\gamma}} \right]} \quad (5)$$

where, T_b is the temperature of chamber B, and P_{ex} is the exhaust pressure (pressure of LP tank).

Furthermore, the supplied pressure P_s that appears in Eq. (3) is the regulated pressure that comes from the proportional pressure regulator PR (Fig.6). However, since P_s is estimated by the CPU based only on the readings of the two pressure transducers T1 and T2 (Fig.6), it can be concluded that:

$$P_s = f(T_1, T_2) \quad (6)$$

Additionally, the motion of the IMU-piston assembly can be modeled by [10]:

$$M(\ddot{x} + g') + D\dot{x} = P_a A_a - P_b A_b + \hat{x} k \Delta \quad (7)$$

where M is the total mass of the IMU-containing capsule, piston and rod; x is the position of the piston inside the cylinder; D is some constant dependant on the materials used and the construction of the apparatus; g' is the component of Earth's acceleration parallel to the direction of induced motion on the IMU; k is the elasticity constant for the front and rear bumpers of the piston, and Δ is the change in length of the bumper. Equations 1-7 now completely define the pneumatic system for inducing a linear and angular motion on the IMU.

2.6 Materials

In order to implement the proposed design, the following materials and components were sourced.

- Pneumatic Cylinder (Cat. No. 2.00CJ2MABUS14AC20, Parker Pneumatics, Calgary, Alberta) with magnetostrictive linear position sensor (Cat. No. BTL5M1M0500RSU022KA02, Parker Pneumatics, Calgary, Alberta)
    - Cylinder Bore: 50.8mm
    - Cylinder Stroke: 508mm
    - Both sides cushioned magnetic piston:
        - Simulated Elasticity Constant(k): 20000N/m
        - Simulated Cushion Thickness: 5mm
    - Inlet/Outlet Air Ports
        - Flow Discharge Coefficient: 0.9
        - Port Cross-Section Area: $1.96*10^{-5}$ m^2
    - Dead Volumes
        - Chamber A/B : $1.96*10^{-3}$ m^3

- Electronic Proportional Pressure Regulator (Cat. No. PAR-15 W2154B179B, Parker Pneumatics, Calgary, Alberta)
    - Analog Voltage Control (0-10V)
    - Simulated Pressure Regulating Function:

      Arguments (High pressure chamber (HP), Low pressure chamber (LP))
      {
      if (HP-LP < 2000Pa AND LP+20kPa < pressure of high-pressure tank)
      {
       Regulated Pressure = LP+20kPa
       }
      else {Regulated Pressure = HP}
      }
- Micro-electromechanical (MEM) Inertial Measurement Unit (MEMSense 2693D, Rapid City, SD)
    - Accelerometers (A50)
        - Dynamic Range: ±50g
        - Drift: 0.3g
    - Gyroscopes (-1200C050)
        - Dynamic Range: ±1200º/s
    - Magnetometers (not utilized in the proposed design)
        - Dynamic Rang: ±1.9G
        - Drift: 2700ppm/ ºC
    - Absolute Maximum Ratings:
        - Operation Temperature: -40 ºC to 85 ºC
        - Acceleration (Shock): 2000g for 0.5ms

3. Results

3.1 Motion of the Piston-IMU Assembly

According to the derived model of the pneumatic system and the outlined parameters of each component, a C++ simulation (Bloodshed Dev C++, Bloodshed Software, www.bloodshed.net/devcpp.html) revealed the position of the piston in the pneumatic cylinder as a function of time (Fig.8).

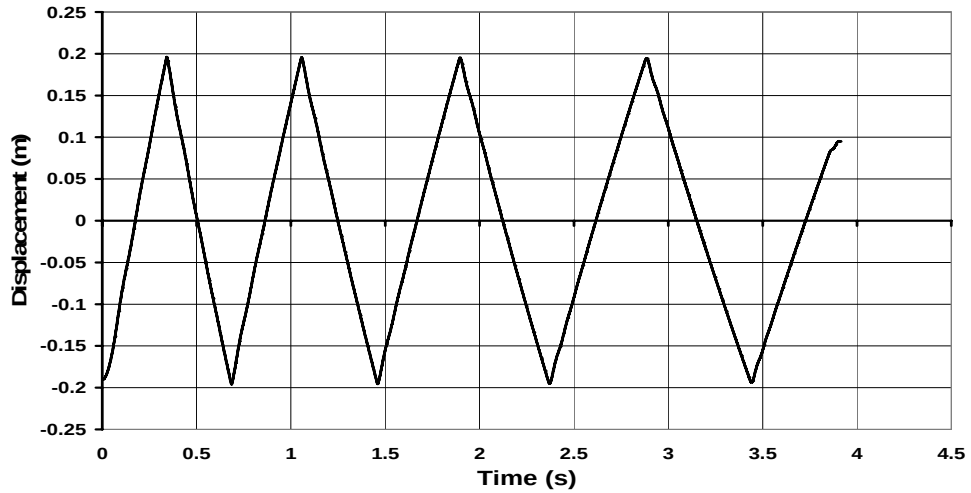


Fig.8: The displacement of the piston inside the pneumatic cylinder as a function of time. The displacement is with respect to the middle of the stroke of the cylinder.

Figure 8 demonstrates that a tank, initially pressurized to ten atmospheres will allow the completion of four full cycles in less than 3.5 seconds. The piston can be then brought to rest during the fifth cycle and locked in place by completely closing the inlet and outlet ports of the cylinder. The acceleration of the piston-IMU assembly was also simulated over the duration of a full cycle (Fig.9).

The constantly changing acceleration of the piston (Fig.9) is due to the specifically implemented function in the simulation, relating the two electronic pressure transducer outputs to the regulated pressure adjusted by the proportional pressure regulator. For a sampling rate of 400Hz, the time intervals of 0 to 0.3 seconds and 0.35 to 0.6 seconds will be proper choices for observations source. The data obtained in these time intervals can then be utilized in aligning the IMU sensors. However, a more gradually changing acceleration of the piston is desired in order to align the IMU more accurately.

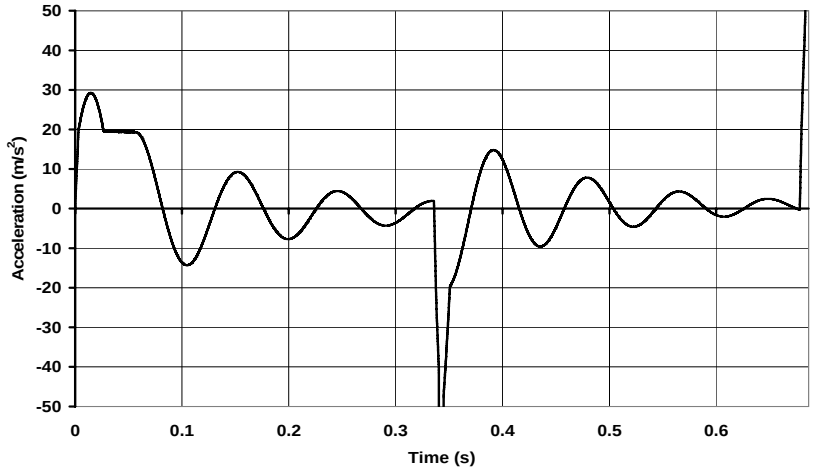


Fig.9: Piston's acceleration as a function of time during one full cycle. The acceleration peaks at 0.34s and 0.68s correspond to the accelerations experienced by the IMU-piston assembly when the piston's bumper collides with the cylinder's wall.

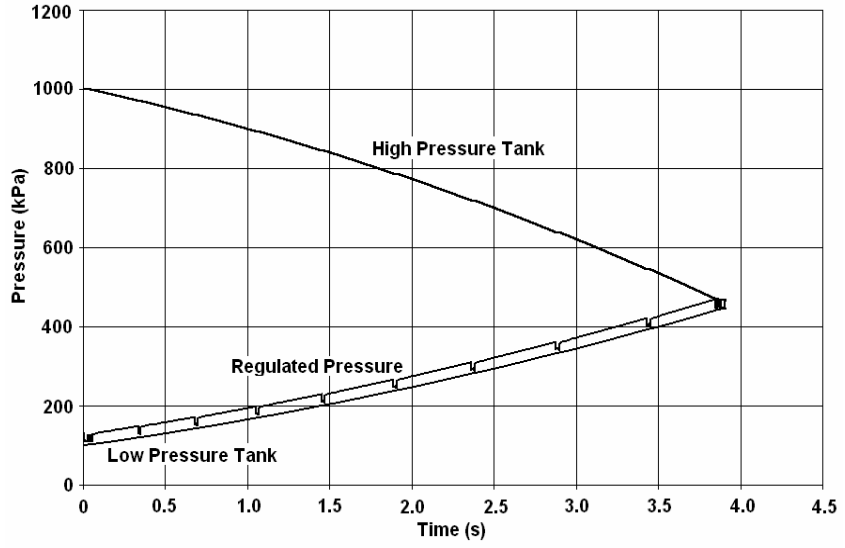


Fig.10: Output of the proportional pressure regulator, and pressures of the high and low-pressure tanks as a function of time over the entire induced motion process.

The pressure in each tank as a function of time during the entire induced motion process has also been explored (Fig.10).

It is clearly evident that after 3.8s (for the outlined system parameters), the pressures in the two tanks will equalize, and the induced motion will come to an end. At this point, the mud-powered air pump is turned on to pressurize the high-pressure tank to its initial value. Although the currently implemented pressure regulating function will yield economical use of the fluid (air), a function that will provide more gradual accelerations of the piston is desired.

4. Conclusion

This article focused on designing and quantifying an apparatus that will allow for an effective, simple and low cost aligning of the sensors of an Inertial-Measurement Unit for continuous angle attitude angle information delivery in a downhole drilling environment. A pneumatic solution was proposed, comprising an air-cylinder, two air tanks, air-pump and a proportional pressure regulator. The highly pressurized air-tank is discharged into the low-pressure tank through the air-cylinder. Correct control of the pressure on each side of the piston of the air-cylinder yields the desired accelerations of the IMU-piston assembly. The position of the piston is constantly monitored by a magnetostrictive sensor, which in turn is differentiated to give the acceleration of the IMU-piston assembly. Moreover, by moving the IMU along a helical thread, angular motion is induced on it, whose angular acceleration is a simple function of the linear acceleration. Once the IMU's angular and linear motion components are known, they are utilized in aligning the unit.

A mathematical model of the entire pneumatic system was derived and simulated with C++. It was shown that an air tank with initial pressure of ten atmospheres will yield more than four full alignment cycles of the IMU-piston assembly within a timeframe of four seconds. The induced accelerations on the IMU-piston assembly were in the range of 3g's, except during a collision with the walls of the air-cylinder, where they reach 80g's. Despite the fact that the model showed a feasible design in downhole conditions, a pressure regulating function that will allow more gradual induced accelerations is desired.

Bibliography

[1] J. Burkmann and N. Nickels, "Directional, navigational and horizontal drilling techniques," *Geothermal Resources Council Bull*., vol. V19, no.4, pp. 106-112, 1990.

[2] S.D. Joshi and W. Ding, "The cost benefits of horizontal drilling," In *Proc. American Gas Association*, Arlington, VA, Apr.-May 29-1, 1991, pp.679-684.

[3] S. D. Joshi, "Horizontal Well Technology", *Technology and Industrial Arts*, PennWell Books, 1991

[4] E.K. Fisher and M.R. French, "Drilling the first horizontal well in the Gulf of Mexico: A case history of East Cameron block 278 well B-12," in *Proc. 66th SPE Annu. Technical Conf. Exhibition*, Dallas, TX, Oct.6-9, 1991, pp.111-123

[5] C. Walker, "Drill Bit Steering," US. Patent 5311953, May 17, 1994

[6] B.A. Shelkholeslami, B.W. Schlottman, F.A. Siedel, and D.M. Button, "Drilling and production aspects of horizontal wells in the Austin Chalkl," *J. Petroleum Technol.*, pp.773-779, Jul.1991.

[7] A. Noureldin, "New measurement-while drilling surveying technique utilizing a set of fiber-optic rotation sensors," Ph.D. Dissertation, Dept. Elect. Eng., Univ. Calgary, Calgary, AB, Canada, 2002.

[8] A. Noureldin, D. Irvine-Halliday, and M.P. Mintchev, "Accuracy Limitations of FOG-Based Continuous Measurement-While-Drilling Surveying Instruments for Horizontal Wells," *IEEE Trans. On Instr. And Meas.*, vol.51, no.6, Oct. 2002.

[9] E. Pecht, "INS In-Drilling Alignment for improving Observability in Horizontal-Directional Drilling,"Ph.D. Dissertation, Dept. Elect. Eng., Univ. Calgary, Calgary, AB, Canada, 2005.

[10] R. Richardson, A.R. Plummer, M. Brown, "Modeling and simulation of pneumatic cylinders for a physiotherapy robot," School of Mech. Eng., University of Leeds, UK,

[11] O. Sound, "Linear Position Sensor Option for Series 2MA Cylinder," Parker Hannifin Corporation, Des Plaines, IL USA

[12] P. Tubel, C. Bergeron, S. Bell, "Mud pulse telemetry system for downhole measurement-while-drilling," *IEEE Instr. And Meas. Tech. Conf.*, 1992, p 219-23

[13] J.L. Thorogood and D. R. Knott, "Surveying techniques with a solid state magnetic multi-shot device," in *Proc. SPE/IADC Drilling Conf.*, New Orleans, LA, Feb. 28-March 3, 1989, pp.841-856.

Authors' Information

Alexander Djurkov – *Department of Electrical and Computer Engineering, University of Calgary, Alberta, Canada, T2N 1N4. Phone: (403) 244-2298; e-mail: alexsd_bg@yahoo.co.uk*

Justin Cloutier – *Imperial Oil Ltd., Calgary, Alberta, Canada; Department of Electrical and Computer Engineering, University of Calgary, Alberta, Canada, T2N 1N4; Phone: (403) 220-2191.*

Martin P. Mintchev – *Prof., Dr., Department of Electrical and Computer Engineering; University of Calgary; Calgary, Alberta, Canada, T2N 1N4; Department of Surgery, University of Alberta; Edmonton, Alberta T6G 2B7; Phone: (403) 220-5309; Fax (403) 282-6855; e-mail: mintchev@ucalgary.ca*

VLSI WATERMARK IMPLEMENTATIONS AND APPLICATIONS

Yonatan Shoshan, Alexander Fish, Xin Li, Graham Jullien, Orly Yadid-Pecht

Abstract*: This paper presents an up to date review of digital watermarking (WM) from a VLSI designer point of view. The reader is introduced to basic principles and terms in the field of image watermarking. It goes through a brief survey on WM theory, laying out common classification criterions and discussing important design considerations and trade-offs. Elementary WM properties such as robustness, computational complexity and their influence on image quality are discussed. Common attacks and testing benchmarks are also briefly mentioned. It is shown that WM design must take the intended application into account. The difference between software and hardware implementations is explained through the introduction of a general scheme of a WM system and two examples from previous works. A versatile methodology to aid in a reliable and modular design process is suggested. Relating to mixed-signal VLSI design and testing, the proposed methodology allows an efficient development of a CMOS image sensor with WM capabilities.*

Keywords: *Watermark, CMOS sensors, image sensors, VLSI, mixed-signal circuits, fast prototyping.*

1. Introduction

The field of digital imaging and its subsidiaries has been going through a continuous and rapid growth during the last decade. Research activity has been extensive in both the academic and commercial communities, and significant advances and breakthroughs are being constantly published [1]. Cost reductions and miniaturization enabled by major developments in VLSI fabrication technologies, CMOS in particular, are making high quality digital imaging products widely accessible, thus effectively taking traditional analog imaging out of the picture. Digital imaging has become a standard in almost all imaging applications, from professional photography and broadcasting to the everyday consumer digital camera. The ease of integrating CMOS imagers with supporting peripheral elements together with a significant reduction in power consumption introduced a variety of new portable products such as imagers on cell phones and narrow-band web cameras [2].

Since digital images are very susceptible to manipulations and alterations, a variety of security problems are introduced. For example, a security centre may wish to authenticate the data received from sensors spread across a facility it is supposed to protect. Another common application is resolving ownership disputes when copyrighted material is distributed illegally. Those problems and needs can be treated by embedding a secret, invisible watermark (WM) in images. A WM is an additional, identifying message, covered under the more significant image raw data, without perceptually changing it. By adding a transparent WM to the image, it can be

made possible to detect alterations inflicted upon the image, such as cropping, scaling, covering, blurring and many more.

The WM can be added on either a software platform or a hardware platform, each having some benefits and some drawbacks. Although WM implementation on a hardware platform suffers from limited processing power, compared to the software implementation, it features real time capabilities and compact implementations. The advantages of hardware WM implementations are especially enhanced in CMOS imagers, where it is possible to integrate the WM embedder monolithically with the sensor array on the same die.

Many WM implementations both in software and in hardware have been proposed in the literature [3]-[11]. In 1990 the modern study of steganography and digital WM was started by Tanaka et al. [10]. They suggested hiding information in multi-level dithered images as a form of secured military communications. Following that work, digital WM arose, and the development of WM algorithms became a growing field of research. Some of the proposed algorithms were relatively simple and weak, merely substituting image least significant bits with WM data [12],[13]. Others had a similar approach but selectively chose the pixels that were to be modulated – either by a random choice to enhance security or according to image quality considerations such as the variance of luminosity. In [14], the WM was embedded in the coefficients of the discrete cosine transform (DCT) of the image to allow better robustness against JPEG lossy compression. Later algorithms modulated only middle-band DCT coefficients [15] to avoid image quality degradation while maintaining a high level of robustness. During the second decade of digital WM research, much thought has been given to the methods in which the WM is implemented and some hardware specific algorithms have been presented [16],[17]. These implementations are usually optimized versions of the former software implementations as will be shown in later sections.

This paper aims to achieve two main objectives. First, the reader is introduced to basic principles and terms in the field of image WM. The paper presents different classification criterions and elementary WM properties such as robustness, computational complexity and influence on image quality. The second goal is to discuss a versatile methodology to design and test hardware implemented WM algorithms, integrated with an image sensor. The proposed methodology speeds up the development process while enhancing reliability.

Section 2 reviews the theory of WM algorithms. Watermark implementations in software and hardware are presented in Section 3. Hardware implementation development methodologies are discussed in Section 4. Section 5 concludes the paper.

2. Theory and implementation of watermark algorithms

2.1 Watermarks Classification

Different applications require utilization of WM with different properties, and no universal WM algorithm that can satisfy the requirements for all kinds of applications has been presented in the literature. WM can be classified into different categories according to various criterions. Figure 1 shows general classification of existing WM algorithms. First of all, all WM can be divided into two main categories: *visible* and *invisible*. The invisibility of a WM is determined by how it affects the image perceptually. Sometimes a WM is intentionally visible, in which case, the identifying image is embedded into the original one and both are visually noticeable. Figure 2 shows examples of the original image and the image with embedded visible WM. Generally, most WM algorithms aim for the WM to be as invisible as possible. Invisible WM has the considerable advantage of not degrading the host data and not reducing its commercial value. For that reason a lot of research has been carried out in this field, while visible WM has received substantially less attention.

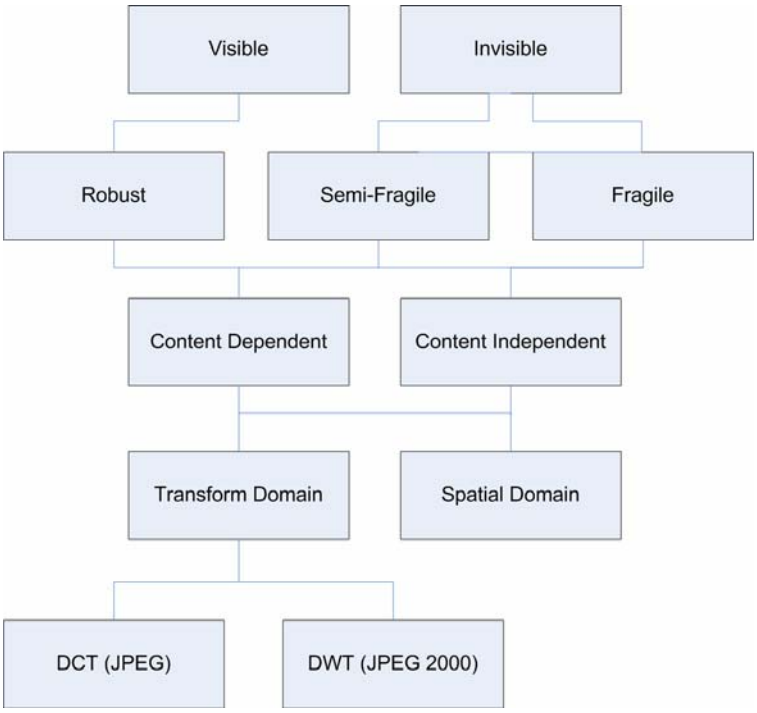


Figure 1. General classification of existing watermark algorithms

Figure 2. Examples of (a) the original image and (b) the image with embedded visible watermark

WM can also be classified according to the level of robustness to image changes and alterations. Three main categories of WM can be identified: *Fragile*, Semi-fragile and *Robust*, though no standard definition exists to explicitly determine which is which. Different applications will have different requirements, while one would need the algorithm to be robust as possible the other may be designed to detect even the slightest modification made to an image - such a WM is called fragile. A fragile WM is practical when a user wishes to directly authenticate that the image he is observing is exactly the same as it was when the WM was first embedded. This might be the case in applications where raw data is used. However, in most existing applications such modifications as lossy compression and mild geometric changes are inherently performed to the image. For those applications it is most efficient to use a semi-fragile algorithm which is designed to withstand certain legitimate modifications, but to detect malicious ones. Finally, some applications, such as copyright protection, require that the WM would be detectable even after an image goes through severe modifications and degradation, including digital-to-analog

and analog-to-digital conversions, cropping, scaling, segment removal and all sorts of attacks. A WM that answers these requirements would be called robust.

Whether or not the algorithm is content dependent is another important distinction. Making the algorithm depend upon the content of the image, is good against counterfeiting attacks, however it complicates the algorithm implementation and therefore the embedding and extracting processes.

An additional classification relates to the domain in which the WM is performed. The most straight forward and simple approach is a WM implementation in the spatial domain that relates to applying the WM to the original image, for example by replacing the least significant bit (LSB) plane with an encoded one [12],[13]. Two other common representations are the discrete cosine (DCT) and the discrete wavelet transforms (DWT) [18],[19] in which the image first goes through a certain transformation, the WM is embedded in the transform domain and then it is inversely transformed to receive the watermarked image.

2.2 Watermark Design Considerations

In order to discuss WM design considerations a number of WM properties should be introduced: (1) *Capacity* (the term is adopted from the communications systems field [21]): in a watermarking system the cover image can be thought of as a channel used to deliver the identifying data (the watermark). The capacity of the system is defined as the amount of identifying data contained in the cover image, (2) *False detection ratio:* this ratio is characterized according to the probability of issuing the wrong decision. It is comprised of the probability to falsely detect an unauthentic WM (false positive), and the probability to miss a legitimate one (false negative). It is possible to manipulate the detection algorithm in order to minimize one or the other, according to the application. The value of this ratio is usually determined experimentally, (3) *Image quality degradation*: the embedding of foreign contents in the image has a degrading effect on image quality. That parameter is relatively hard to quantize and different measures such as peak signal-to-noise ratio (PSNR) or a subjective human perception measure may be applied.

These properties are elementary in every WM system and need to be carefully appreciated. The following subsections show how they are considered from different design point of views and indicate several trade-offs between them.

2.2.1 Robustness to Attacks

A good attack on fragile and semi-fragile WM will attempt to modify the perceptual content of the image, without affecting the WM data embedded in it. Knowledge of the embedding and extracting methods is assumed. There are two approaches for an attack; while the first approach requires the decryption of the encoded mark in order to produce a suitable WM on an unauthentic image, the second one aims to maintain the original mark on a modified image without knowing the mark itself. Decrypting the original WM is a cryptographic computational problem and is directly related to the capacity of the WM system. In WM however, the potential for such an attack is even greater (compared with the cryptographic case) as the attacker does not have to find the exact key, but only one that would be close enough to pass over the detector's threshold. And still, if the capacity of the WM is large enough - using a key of several hundred bits, this attack may not be computationally tractable.

There are numerous attacks that take advantage of the existing image to create a forged one. The most intuitive one is the cover up attack [17]. This attack can be used when the mark is embedded independently to a block divided image. If the image contains homogeneous areas such as a wall, or a floor and the attacker wishes to hide a smaller object, he may do so by copying other blocks in such a way that the change would be perceptually un-noticeable, but the detector would still recognize a valid WM on the copied block. A possible counter measure for such an attack is to complicate the scheme, while increasing system complexity, and create dependencies between the marks of neighboring blocks – if a block is placed in the wrong place, detection would be false.

Attacks on copyright protection WM are designed to cause defects to the embedded WM so that it will be undetectable, while still maintaining reasonable image quality. Such attacks may include one or more of the

following: (1) A geometric attack such as cropping, rotation, scaling etc, (2) A Digital-to-Analog conversion, such as printing and then Analog-to-Digital conversion by scanning (can also be done by re-sampling), (3) Lossy compression and (4) Duplicating small segments of the picture and deleting others (jitter attack) [9].

It is shown then, that several parameters must be considered for each application, in order to optimize the use of counter-measures. The goal is to maintain the required image quality desired for each application and still be robust to potential attacks. That trade-off is discussed in the next two subsections.

2.2.2 Image quality

As mentioned, an important objective of a good WM is minimizing image quality degradation. Recently we have shown [20], that for a blind content-independent algorithm, the trade-off between the security (capacity) of the mark and the negative affect on image quality is straight-forward. There, the WM is embedded by adding a pseudo-random noise to each pixel. Increasing the bit size of the mark is equivalent to increasing the variance of the noise, which is the measure for the capacity of that algorithm. It relates directly to better false detection ratio. However, it also adds significant high frequency values to the original image, affectively degrading its quality, especially in homogeneous parts of the picture.

To avoid such a significant degradation, it is possible to increase the security of the mark by making it content dependent [16],[18]. In a content dependent WM system, the embedded data is also some function of the cover image. The decoder would need the cover image data in order to extract the correct WM, making it more difficult, and sometimes even impossible to use for marking unoriginal content. This introduces higher computational complexity, but features a more secured mark without influencing the cover data severely.

2.2.3. Computational complexity

Intuitively, it is obvious that in order to apply a more complicated algorithm, more complex embedding and detecting blocks would be required. The motivation to keep the computational complexity low depends on the application and on the method of implementation. In real time applications, computations must be done in a very short time period. The speed and processing power of the computational platform at hand, limit the algorithm level of complexity that can be computed in a given time frame. When implementing in hardware, higher complexity requires additional hardware which means more area and additional costs.

In [20] we have introduced a scheme that is very easily implemented in hardware. In this implementation, computation time and hardware requirements are almost negligible, however compromising the performance achieved. As previously mentioned, in order to provide high detection rates perceptual effects are inevitable. In addition, this scheme is not able to detect local modifications. A potential attacker may take advantage of this inability, to cover parts of the picture he is trying to hide.

Depending on the intended application, more complicated schemes can be implemented to withstand expected attacks. If, for instance, localization of the changes made is important, a partition of the image into blocks may be of use. If the marked image is expected to go through lossy compression, one may consider embedding the WM in the frequency domain, as will be described in the next section. Other algorithms employ global and local mean values, temporal dependencies (in video WM) and variety of extra features to enhance their performance. However, each additional feature, added to the algorithm, increases the computational effort and hardware resources (such as memory and adders\multipliers) used. Therefore, an optimized scheme will be comprised of the minimum number of features needed to satisfy the needs of the application it is designed for.

2.3 Discussion

It has been shown that during the design of a WM system, many trade-offs are taken into account. How can one evaluate the overall quality of the final outcome? Although there is no accepted standard to uniformly asses the quality of WM [21], there are a few popular benchmarks available. A designer can use the evaluated system to embed a WM in a series of test images, and then run them through the benchmark and asses the performance by

observing the quality of detection. The *StirMark* code, which is used for evaluating the robustness of WM algorithms designed for copyright protection applications, applies a series of attacks on a marked image [22]. In addition, it is possible to evaluate robustness to specific attacks by manually adding them to this benchmark. The *Checkmark* benchmark provides a framework for application-oriented evaluation of WM schemes, applicable to all sorts of WM algorithms including fragile and semi-fragile [21]. The use of such independent, third party, evaluation tools provides a good perspective on how well a WM system performs.

3. Watermark implementations – Software vs. Hardware

Figure 3 shows a scheme of a general WM system. The system consists of a WM generation, embedding and detection algorithms.

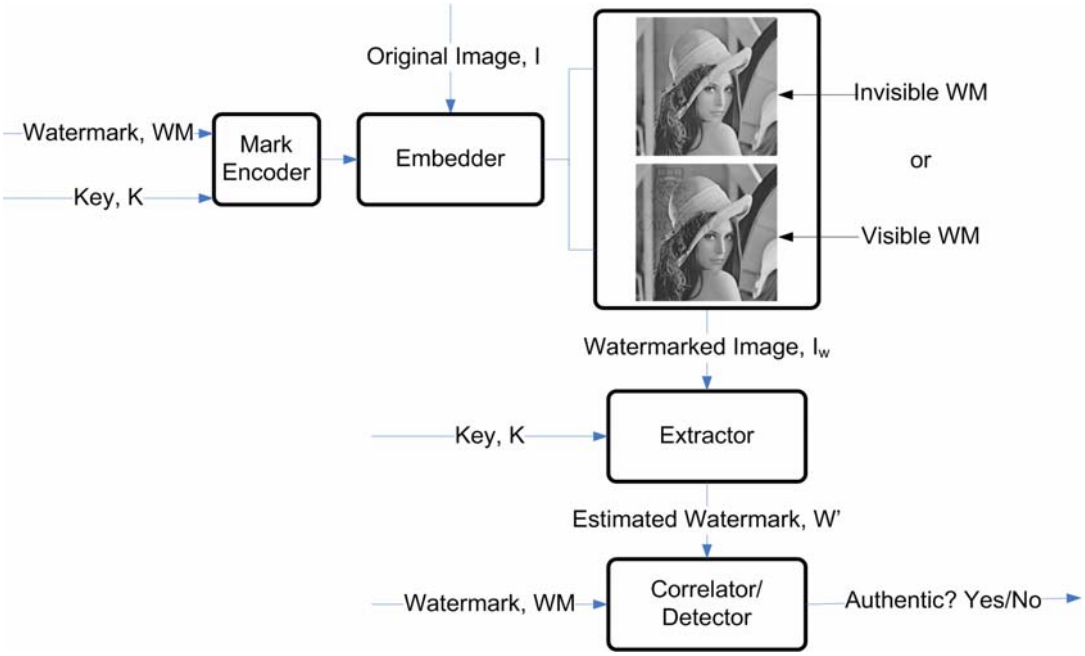


Figure 3. Scheme of general watermark system.

The identifying data (W in Figure 3) can be meaningful, like a logo, or it can just be a known stream of bits. First the identifying data is encoded using a secret key, K. Then the encoded identifying data is embedded into the original image (I in Figure 3). The result is the WM image. As previously mentioned, the WM can be visible or invisible, as shown in Figure 3. The detector part is at the receiving end. The objective is to extract the identifying data embedded in the received image, using the secret key and an inverse algorithm. Finally, a decision is made by correlating the extracted mark with the original and applying a chosen threshold.

The system can be implemented on either software or hardware platforms, or some combination of the two. A pure software WM scheme can simply be implemented in a PC environment. Such an implementation is relatively slow, as it shares computational resources and its performance is limited by the operating system. It is unsuitable for real-time applications, for it would be too slow, and it cannot be implemented on portable imaging devices that have limited processing power. On the other hand it can be easily programmed to realize any algorithm of any level of complexity, and can be used on everyday consumer PC's.

A good example of software WM solution was presented by Li [18]. In this work he proposes a software implemented fragile WM, embedded in the coefficients of the block DCT. This algorithm is designed for authentication and content integrity verification of JPEG images. The algorithm embeds the WM only in a few

selected DCT coefficients of every block in order to minimize the effect on the image. The author directly addresses known issues in similar previous works, inserting additional complexity to overcome security gaps. The system utilizes the advantages of software implementation by using resources needed to store image data, transform coefficients and WM mappings. Using a combination of different security resources, including a non-deterministic mapping of the location of coefficient modulation and block dependencies, the system succeeds in facing several attacks without changing the affect on image quality – when compared to similar works. Moreover, the computations involved in the embedding process are kept relatively basic, suggesting suitability for future hardware implementation as well.

In opposite to software solutions, hardware implementations offer an optimized specific design to incorporate a small, fast and potentially cheap WM embedder. It is most suitable for real time applications, where computation time has to be deterministic (unlike software running on a windows system for example) and short. Optimizing the marking system hardware enables it to be added into various portable imaging devices. In a full imaging system that includes both the imager and WM embedder, the system security is improved as it is certain that the data entering the system is untouched by any external party. However, hardware implementations usually limit the algorithm complexity and are harder to upgrade. The algorithm must be carefully designed, minimizing any unexpected deficiencies. For example, in [16], Mohanty et al. present an implementation of both fragile and robust invisible WM algorithms in hardware. Which WM is used will be defined by the user. The WM are embedded in the spatial domain but are designed to be robust for JPEG compression. The motivation for hardware implementation is to enable the integration of a WM module within a secure digital camera system. As JPEG is the standard data format for digital cameras it is imperative that the algorithm will not be harmed by compression. The availability of two kinds of WM algorithms corresponds to applications such as image authentication for the fragile algorithm and copyright protection for the robust. The hardware employed in this implementation is comprised of image and WM RAM memories, adders/subtractors, registers and multipliers. This is a relatively large implementation and accordingly, the algorithm is rather complex.

4. WM Hardware Implementation - A Development Methodology

In this section a development methodology to a fast mixed signal hardware design is presented. This methodology can be used for the development of an image sensor with integrated WM capabilities. Figure 3 shows different elements required for such a system. The design of this complete system is a very demanding task requiring time and financial resources. It involves hardcore analog and mixed signal design as well as complex digital architecture. Although the end goal is to implement the whole system monolithically on a single chip, it is expected, as in every development process, that more than one prototype will be designed before a final version is issued. Therefore it is worthwhile, to first focus on determining the core elements of the system which are the imager, the WM digital architecture and the interface between the two. To do so without significantly compromising the quality and performance of the peripheral elements such as the A/D converter, analog voltage biasing and memory, the design is first tested on a board utilizing commercial devices.

This specifically designed board, shown in Figure 5, emulates a System-On-a-Chip (SoC) platform, allowing the incorporation of custom VLSI designs (the imager) with peripheral elements and digital logic implemented on an FPGA. It features low-noise, separated digital and analog power supplies, 12 bit analog voltage and current biasing, 12 and 18 bit A/D converters, an SRAM memory and several I/O ports including LVDS, RS-232 and direct test points for maximum testing flexibility. The designer can choose what part of the system he wants to implement in VLSI and what elements he would use of those available on board. For the discussed WM implementation a basic imager is first designed, and then the WM logic is implemented on the FPGA, together with all other required control logic, making use of the A/D converter and SRAM memory to aid the implementation of more complex algorithms. Finally the data is read out either as a WM video stream through the LVDS interface or stored in memory and read as a WM still image.

Note the presented methodology is modular and can host various kinds of SoC designs.

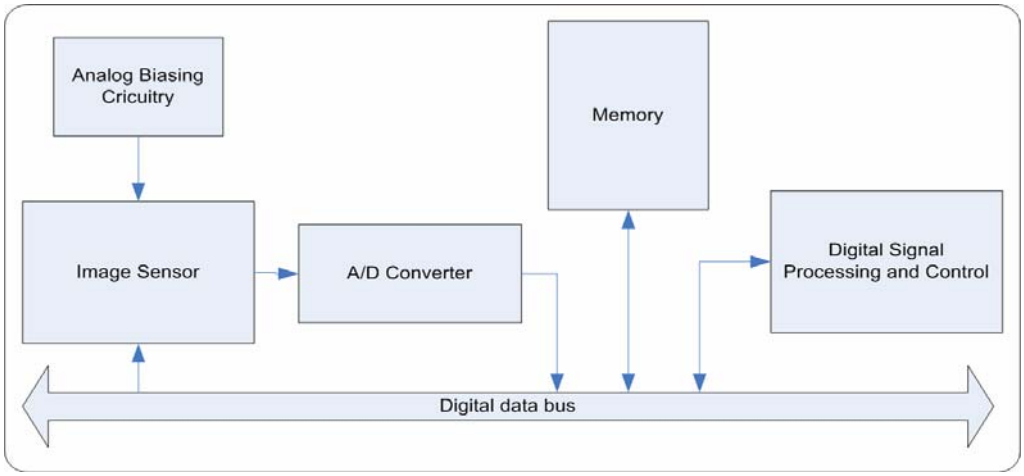


Figure 4. A general scheme of the complete image system in conjunction with digital processing.

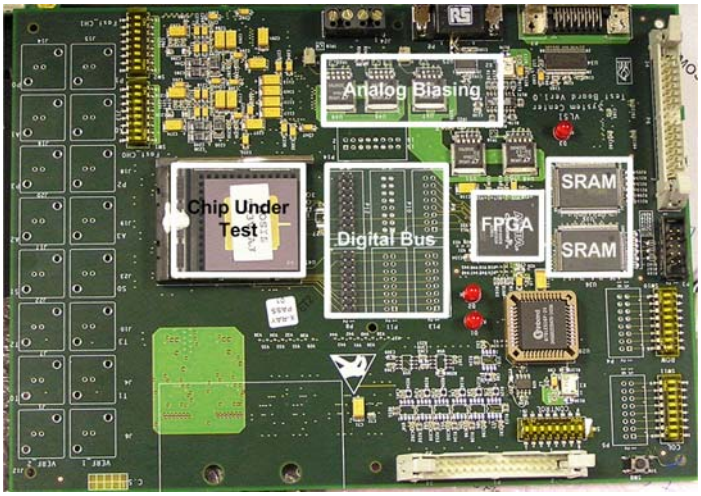


Figure 5. Mixed signal SoC fast prototyping custom development board.

5. Conclusions

Basic terms and principles in WM design and evaluation were presented. The main trade-offs and design considerations were discussed pointing out the importance of designing in light of the intended application and expected attacks. Several common attacks were also described, as well as a couple of evaluation benchmarks that facilitate the testing of different WM schemes robustness to a very large scale of attacks. The general scheme of a WM system was shown and the major benefits and shortcomings to implementations in hardware or software described. Two examples of previous works done, featuring fragile and robust WM, spatial and DCT domains, hardware and software implementations were given. In addition, a development methodology, employing custom development board for mixed signal SoC fast prototyping was shown.

Bibliography

[1] V. M. Potdar, S. Han, E. Chang, "A survey of digital image watermarking techniques", 3rd IEEE International Conference on Industrial Informatics (INDIN '05), Aug. 2005, pp. 709- 716

[2] O. Yadid-Pecht and R. Etienne-Cummings, " CMOS imagers: from phototransduction to image processing", Kluwer Academic Publishers

[3] S. P. Mohanty, "Digital Watermarking: A Tutorial Review", URL: http://www.csee.usf.edu/~smohanty/research/Reports/WMSurvey1999Mohanty.pdf

[4] F. Mintzer, G. Braudaway, and M. Yeung, “Effective and ineffective digital watermarks,” in Proc. IEEE Int. Conf. Image Process., vol. 3, 1997, pp. 9–12.

[5] C. T. Li, D. C. Lou and T. H. Chen, “Image authenticity and integrity verification via content-based watermarks and a public key cryptosystem”. Proc. IEEE Int. Conf. on Image Processing, Vancouver, Canada Sep.2000.vol. III, pp. 694-697.

[6] S. P. Mohanty, et al., "A DCT Domain Visible Watermarking Technique for Images", to appear in Proc. of the IEEE International Conference on Multimedia and Expo, July 30- August 2, 2000, Hilton New York & Towers, New York City, NY, USA.

[7] M. J. Tsai and H. Y. “Wavelet Transform Based Digital Watermarking for Image Authentication,” IEEE Proceedings of the Fourth Annual ACIS International Conference on Computer and Information Science,0-7695-2296-3/05,2005.

[8] P. Meerwald S. Pereira, “Attacks applications and evaluation of known watermarking algorithms with Checkmark”. SPIE Electron. Imaging. v4675

[9] F. A. P. Petitcolas, R. J. Anderson, and M. G. Kuhn, “Attacks on copyright marking systems,” Information Hiding: 2nd Int. Workshop, D.Aucsmith, Ed. , ser. Lecture Notes in Computer Science Berlin, Germany: Springer-Verlag, vol. 1525, pp. 218-238, 1998

[10] K. Tanaka, Y. Nakamura and K. Matsui, "Embedding Secret Information into a Dithered Multi-level Image" IEEE Military Communications Conference 1990 pp. 0216-0220

[11] F. Petitcolas, R. J. Anderson and M. G. Kuhn, "Information Hiding - A Survey" Proc. IEEE 87(7) Jul. 1999 pp. 1062-1078

[12] R. G. van Schyndel, A. Z. Tirkel, and C. F. Osborne, "A digital watermark", in Proc. IEEE Int. Conf. Image Processing, vol. 2, Austin, TX, 1994, pp. 86–90

[13] R. B. Wolfgang and E. J. Delp, "A watermark for digital images", in Proc. IEEE Int. Conf. Images Processing, Lausanne, Switzerland, Sept. 1996, pp. 219–222

[14] I. J. Cox, J. Kilian, T. Leighton, and T. Shamoon, "A secure, robust watermark for multimedia" in R. J. Anderson, Ed., "Information hiding: First international workshop", in Lecture Notes in Computer Science, vol. 1174. Berlin, Germany: Springer-Verlag, 1996, pp. 183–206

[15] C. T. Li, "Digital fragile watermarking scheme for authentication of JPEG images", IEE Proc., Vis. Image Signal Process., 2004, 151, (6), pp. 460-466

[16] S. P. Mohanty, N. Ranganathan, and R. K. Namballa, "VLSI implementation of invisible digital watermarking algorithms towards the developement of a secure JPEG encoder", in Proc. IEEE Workshop Signal Processing Systems, 2003, pp. 183-188

[17] T. H. Tsai, C. Y. Lu, "Watermark embedding and extracting method and embedding hardware structure used in image compression system", US Patent 6,993,151, 2006

[18] C. T. Li, "Digital fragile watermarking scheme for authentication of JPEG images", IEE Proc.-Vis. Image Signal Processing, Vol. 151, No. 6, December 2004, pp. 460-466

[19] M. Barni, F. Bartolini, and A. Piva, "Improved Wavelet-Based Watermarking Through Pixel-Wise Masking", IEEE Trans. Image Proc., vol. 10, no. 5, pp. 783-791, May 2001

[20] G. R. Nelson, G. A. Jullien and O. Yadid-Pecht, "CMOS image sensor with watermarking capabilities", in Proc. IEEE Int. Symp. on Circuits and Systems (ISCAS '05), vol. 5, Kobe, Japan, May 2005, pp. 5326-5329

[21] P. Meerwald and S. Pereira, "Attacks, applications and evaluation of known watermarking algorithms with checkmark", In Proceedings of SPIE, Electronic Imaging, Security and Watermarking of Multimedia Contents IV, 2002

[22] M. Kutter and F. A. P. Petitcolas, "A fair benchmark for image watermarking systems", 11th Int. Symp. Electronic Imaging, vol. 3657, San Jose, CA: IS&T and SPIE, Jan. 25-27, 1999

[23] I. J. Cox, G. Doerr, T. Furon, "Watermarking is not cryptography", in YQ Shi, B Jeon, Eds., "Digital Watermarking : 5th International Workshop", in Lecture Notes in Computer Science, vol. 4283, Berlin, Germany: Springer-Verlag, 2006, pp. 1-15

Authors' Information

Yonatan Shoshan *– ISL lab, ATIPS Lab, ECE Department, University of Calgary, Calgary AB, Canada; e-mail: shoshay@atips.ca*

Alexander Fish *– ISL lab, ATIPS Lab, ECE Department, University of Calgary, Calgary AB, Canada; e-mail: fish@atips.ca*

Xin Li *– ISL lab, ATIPS Lab, ECE Department, University of Calgary, Calgary AB, Canada; e-mail: xinli@atips.ca*

Graham Jullien *– ATIPS Lab, ECE Department, University of Calgary, Calgary AB, Canada; e-mail: jullien@atips.ca*

Orly Yadid-Pecht *– ISL lab, ATIPS Lab, ECE Department, University of Calgary, Calgary AB, Canada; The VLSI Systems Center, Ben-Gurion University, Beer-Sheva, Israel;*

e-mail: orly@atips.ca

IMAGE PARTITION TRANSFORMS FOR FAITHFUL SEGMENTATION SEARCH

Dmitry Kinoshenko, Sergey Mashtalir, Konstantin Shcherbinin, Elena Yegorova

Abstract*: The explosion of image content is closely connected with segmentations efficiency. High-level region-based interpretations are associated with some a priori information, measurable region properties, heuristics, and plausibility of computational inference. Conventional similarity analysis consists of following steps: images are segmented into disjoint regions, features are extracted from each region and the set of all features is used for high-level processing. Quite often simultaneous processing of several partitions is desired in order to produce reliable true conclusion. We propose operations with segmented images and a metric for nested partitions.*

Keywords*: image, spatial reasoning, partitions.*

ACM Classification Keywords*: I.4.6 Segmentation: region growing, partitioning*

Introduction

Efficiency of image structuring and understanding strongly depends on a segmentation as a process of separating an image into several disjoint (or weakly intersecting) regions whose characteristics such as intensity, color, texture, shape, etc. are similar [see e.g. 1, 2]. Segmentation is a key step in early vision and it has been widely investigated in the field of image processing. Nevertheless image content formal descriptions using only low-level features extracted from each region are not necessarily the case for true conclusions. We may get totally correct segmentation, but most often we obtain under-segmentation, over-segmentation, missed regions, and noise regions. It should be emphasized that a fair segmentation can be provided if and only if we know exactly what we are looking for in an image.

To obtain a reliable image interpretation we have to transform a row image into an image data structure, then into an image knowledge structure and finally into a user-specific knowledge structure. Spatial reasoning plays a most important part in decision making. In this respect partition transforms (e.g. set theoretic) are serviceable in order to find regions that are heavily correlated with significant objects in the scene and it is essential to have tools in order to compare segmented image accurately.

The three classes of distance function (point to point, point to set, and set to set) are usually discussed as measures of proximity or dissimilarity in image processing [3, 4]. It is desirable to define an image metric that can

be efficiently embedded in segmentations methods. A partition metric is consequently a candidate because it represents images as a finite subsets assemblage that takes into account mutual dependences of equivalence class corresponding to separate regions of interest. Metrics on nested partitions take on special significance since they give possibilities to define hierarchical content descriptions. We propose based on spatial relations operations with segmented images and a metric on nested partitions useful for such applications as object tracking and pattern matching.

Operations with segmented images

Let $\mathrm{B}(x)$ be an image and $x \in \mathrm{D} = \mathbb{Z}_n^+ \times \mathbb{Z}_m^+$ (here D is a viewing field). It should be noted that any faithful segmentation (a crisp clustering) generates a partition of the viewing field, i.e. $\mathrm{X} = \{[x]_1, \ldots, [x]_\alpha, \ldots, [x]_s\}$ where $[x]_\alpha \neq \varnothing$, $\mathrm{X} = \bigcup_{\alpha=1}^{s} [x]_\alpha$, $\forall \alpha \neq \beta \Rightarrow [x]_\alpha \cap [x]_\beta = \varnothing$ (hereafter α, β, γ denote all allowable indices). Suppose that a region labeling $\mathrm{F} : \mathrm{B}(x) \to \mathbb{Z}_r^+$ corresponds to obtained segmentation then arbitrary two points $x', x'' \in \mathrm{D}$ belong to the same equivalence class $x', x'' \in [x]_\alpha$ if the binary relation $\langle \mathrm{B}(x'), \mathrm{B}(x'') \rangle \in \tau \Leftrightarrow$ $\Leftrightarrow \mathrm{F}(\mathrm{B}(x')) = \mathrm{F}(\mathrm{B}(x'')) = \alpha$ is fulfilled.

Let us introduce a characteristic function on equivalence classes

$$\lambda_{[x]_\alpha}(x) = \begin{cases} 0, x \in [x]_\alpha, \\ 1, x \in \mathrm{D} \setminus [x]_\alpha. \end{cases} \tag{1}$$

It follows immediately that boundary conditions for spatial reasoning are

$$\lambda_{\mathrm{D}}(x) = 0, \ \lambda_{\varnothing}(x) = 1.$$

In addition it is reasonable to indicate the expression providing certain duality in order to analyze image contents

$$\lambda_{\mathrm{D} \setminus [x]_\alpha}(x) = 1 - \lambda_{[x]_\alpha}(x).$$

The direct check-up allows to introduce explicitly definable formulae of spatial interdependence between characteristic functions of two elements of arbitrary partitions X and Y

$$\lambda_{[x]_\alpha \cup [y]_\beta}(x) = \lambda_{[x]_\alpha}(x) \lambda_{[y]_\beta}(x), \tag{2}$$

$$\lambda_{[x]_\alpha \cap [y]_\beta}(x) = \lambda_{[x]_\alpha}(x) + \lambda_{[y]_\beta}(x) - \lambda_{[x]_\alpha}(x) \lambda_{[y]_\beta}(x), \tag{3}$$

$$\lambda_{[x]_\alpha \setminus [y]_\beta}(x) = 1 - \lambda_{[x]_\alpha}(x) + \lambda_{[y]_\beta}(x). \tag{4}$$

Appreciably intense interest consists in simultaneous transformations of equivalence class families since namely splitting and merging of partitions can get totally correct and complete segmentation of complex scenes. It easily seen that for any unions and intersections we get

$$\Xi = \bigcup_{\gamma \in \Gamma} [x]_\gamma \Rightarrow \lambda_\Xi(x) = \min_{\gamma \in \Gamma} \lambda_{[x]_\gamma}(x), \tag{5}$$

$$\Xi = \bigcap_{\gamma \in \Gamma} [x]_\gamma \Rightarrow \lambda_\Xi(x) = \max_{\gamma \in \Gamma} \lambda_{[x]_\gamma}(x). \tag{6}$$

The 169 types of spatial relations between two rectangles in 2-D space had been proposed in [2]. However, if we introduce a representation of each equivalence class as union of sets (rather points of boundaries and interior) it suffices to use combinations only of four relations in general case. Indeed, suppose that

$$\lambda_{[x]_\alpha}(x) = \partial \lambda_{[x]_\alpha} \cup \lambda^{\circ}_{[x]_\alpha}$$

where $\partial \lambda_{[x]_\alpha}$ denotes the boundary of the partition element describing by the characteristic function (1) and

$\lambda^{\circ}_{[x]_\alpha}$ corresponds to interior points of this partition element. Let us introduce relations defining spatial relationships between any two objects $[x]_\alpha$ and $[x]_\beta$, $\alpha \neq \beta$

$$\begin{cases} \langle [x]_\alpha, [x]_\beta \rangle \in \tau_{11} \Leftrightarrow \partial\lambda_{[x]_\alpha} \cap \partial\lambda_{[x]_\beta} \neq \varnothing, \\ \langle [x]_\alpha, [x]_\beta \rangle \in \tau_{12} \Leftrightarrow \partial\lambda_{[x]_\alpha} \cap \lambda^{\circ}_{[x]_\beta} \neq \varnothing, \\ \langle [x]_\alpha, [x]_\beta \rangle \in \tau_{21} \Leftrightarrow \lambda^{\circ}_{[x]_\alpha} \cap \partial\lambda_{[x]_\beta} \neq \varnothing, \\ \langle [x]_\alpha, [x]_\beta \rangle \in \tau_{22} \Leftrightarrow \lambda^{\circ}_{[x]_\alpha} \cap \lambda^{\circ}_{[x]_\beta} \neq \varnothing. \end{cases} \quad (7)$$

Consequently, the (2×2) matrix (τ_{ij}) entirely determines all eight possible mutual locations of regions, viz:

i) $[x]_\alpha$ disjoins $[x]_\beta$, i.e. all parts of $[x]_\alpha$ are separated from all parts of $[x]_\beta$ iff $\langle [x]_\alpha, [x]_\beta \rangle \notin \tau_{ij}$ $\forall i, j$;

ii) $[x]_\alpha$ contains $[x]_\beta$, i.e. all parts of $[x]_\beta$ are completely overlapping with any part of $[x]_\alpha$ iff τ_{21}, τ_{22}, are valid and the relations τ_{11}, τ_{12} are not true;

iii) similarly, $[x]_\alpha$ belongs $[x]_\beta$ iff $\langle [x]_\alpha, [x]_\beta \rangle \in \tau_{12}, \tau_{22}$ and $\langle [x]_\alpha, [x]_\beta \rangle \notin \tau_{11}, \tau_{21}$;

vi) $[x]_\alpha$ equals to $[x]_\beta$ iff $\langle [x]_\alpha, [x]_\beta \rangle \in \tau_{11}, \tau_{22}$ and $\langle [x]_\alpha, [x]_\beta \rangle \notin \tau_{12}, \tau_{21}$;

v) $[x]_\alpha$ is partly overlapping $[x]_\beta$ iff all relations τ_{ij} hold;

vi) $[x]_\alpha$ is externally bound to bound with $[x]_\beta$, i.e. there exist common points of boundaries and no part of $[x]_\alpha$ is overlapping with any part of $[x]_\beta$ iff $\langle [x]_\alpha, [x]_\beta \rangle \in \tau_{11}, \tau_{22}$ and $\langle [x]_\alpha, [x]_\beta \rangle \notin \tau_{12}, \tau_{21}$;

vii) $[x]_\alpha$ is internally bound to bound with $[x]_\beta$, i.e. there exist common points of boundaries and $[x]_\alpha$ belongs $[x]_\beta$ iff only the relation τ_{21} is not true;

viii) $[x]_\beta$ is internally bound to bound with $[x]_\alpha$, i.e. there exist common points of boundaries and $[x]_\alpha$ contains $[x]_\beta$ iff only the relation τ_{12} is not true.

All mentioned cases are illustrated by figure 1.

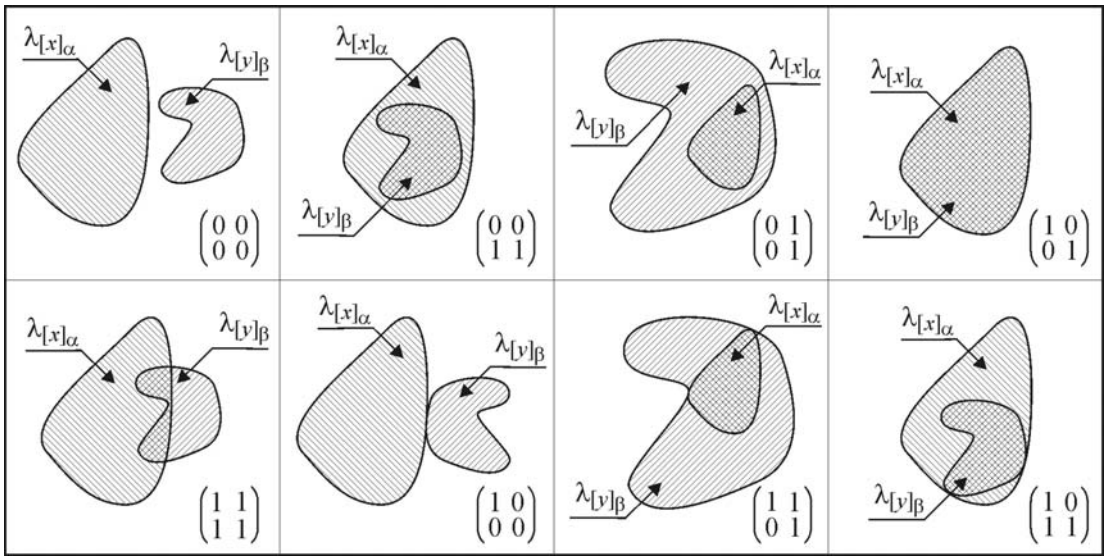


Figure 1. Possible mutual locations of equivalence classes

Now we can formalize intersection and conditional union operations with partitions. For simplicity of notations we write μ instead of a matrix (τ_{ij}) elements sum then introducing an indicator function

$$\varphi(\alpha,\beta)=\begin{cases}-1, s=1;\\ 0, s=0;\\ 1, s>1.\end{cases}$$

we get for $\mathrm{X}=\{[x]_\alpha\}$, $\mathrm{Y}=\{[y]_\beta\}$

$$\mathrm{Z}=\mathrm{X}\otimes\mathrm{Y},\ \mathrm{Z}=\{[z]_\gamma:\lambda_{[z]_\gamma}(x)=\lambda_{[x]_\alpha}(x)+\lambda_{[y]_\beta}(x)-\lambda_{[x]_\alpha}(x)\lambda_{[y]_\beta}(x)\} \tag{8}$$

and

$$\mathrm{Z}=\mathrm{X}\oplus\mathrm{Y},\ \mathrm{Z}=\{[z]_\gamma\},\ [z]_\gamma=\begin{cases}\{[x]_\alpha,[y]_\beta\}, if\ \varphi(\alpha,\beta)=0;\\ [x]_\alpha\cup[y]_\beta, if\ \varphi(\alpha,\beta)=1.\end{cases} \tag{9}$$

It is obvious evident that under $\varphi(\alpha,\beta)=-1$ a complementary analysis is required since merging of adjoining region is admissible action if features of $[z]_\gamma$ with the characteristic function $\lambda_{[z]_\gamma}(x)=\lambda_{[x]_\alpha}(x)\lambda_{[y]_\beta}(x)$ satisfy, for instance, requirements to the sought-for shape.

Thereby, expressions (2)–(4) determine operations with separate equivalence classes, relationships (5), (6) predetermine transformations of equivalence class families and (8), (9) on the base of relations (7) provide partition manipulations. The main goal of such segmented image reforming is a guaranteeing trade-off decision about regions of interest.

Results and outlook

Significant efforts are continuously being made in development of segmentation techniques. Cognitive-like approaches require obtaining of regions strongly correlated with meaningful objects in the scene. Mentioned operations create the necessary prerequisites for partitions transformations. However, efficiency of image structuring and understanding depends on the objectivity of partitions matching. Previously for finite sets we proved [5] that the functional

$$\rho(\mathrm{X},\mathrm{Y})=\sum_\alpha\sum_\beta card([x]_\alpha\Delta[y]_\beta)card([x]_\alpha\cap[y]_\beta) \tag{10}$$

(here the notation $\mathrm{X}_i\Delta\mathrm{Y}_j$ defines a symmetric difference) is a metric. Later for arbitrary measurable set with given measure $\mu(\circ)$, which can be interpreted as length, area, volume, mass distribution, probability distribution, and in special case cardinality, we had proved [6] that the functional

$$\rho(\mathrm{X},\mathrm{Y})=\sum_\alpha\sum_\beta\mu([x]_\alpha\Delta[y]_\beta)\mu([x]_\alpha\cap[y]_\beta) \tag{11}$$

is a metric also. Taking into consideration properties of nested partitions one can give concrete expression to metrics (10) and (11) for $\mathrm{X}\subseteq\mathrm{Y}$

$$\rho(\mathrm{X},\mathrm{Y})=\sum_\beta\mu([y]_\beta)^2-\sum_\alpha\mu([x]_\alpha)^2 \tag{12}$$

Figure 2. Example of nested partitions

Substantially metric (12) intends for combination of visual features and metadata analysis to solve a semantic gap between low-level visual features and high-level human concept. Figure 2 illustrates nested partitions that are generated by algorithms based on adaptive thresholding, multithresholding and band-thresholding [7]. Simple geometrical shape parameters (the area and the perimeter of region, the diameters of circles with fixed area and perimeter, orthogonal projections of the figure on axes of abscissae and ordinates, the minimal and the maximal orthogonal projections of the figure on a line, the distance between opposite sides of the figure, the distance from an origin point in the figure to its boundary point for a given direction, the same average distance for all possible directions for a given point, the lengths of the long and short semi-axes of the ellipse with given area and perimeter, drainage-basin circularity, coefficient convexity ratio, etc.) were used for split and merging procedures along with relations (7) under operations (8) and (9).

The analysis of experimental results has shown that partition transforms and unbiased partitions matching substantially meant for the use at conceptual segmentation which not only builds partitions but can also explain why a set of regions confirms a desired pixel family.

Bibliography

[1] X. Jiang. Performance evaluation of image segmentation algorithms. In: Handbook of pattern recognition and computer vision, C.H. Chen and P.S.P. Wang (Eds.), World Scientific, Singapore, 2005, pp. 525–542.

[2] S.Y. Lee, F.J. Hsu. Spatial knowledge representation for iconic image database. In: Handbook of pattern recognition and computer vision, C.H. Chen, L.F. Pau, P.S.P. Wang (Eds.), World Scientific, Singapore, 1993, pp. 839–861.

[3] B. Li E. Chang, Y. Wu Discovery of a perceptual distance function for measuring image similarity. Multimedia Systems. Vol. 8, No 6, 2003, pp. 512–522.

[4] D. Wang, X. Ma, Y. Kim. Learning Pseudo Metric for Intelligent Multimedia Data Classification and Retrieval. Journal of Intelligent Manufacturing, Vol.16, No 6, 2005, pp. 575–586.

[5] V. Mashtalir, E. Mikhnova, V. Shlyakhov, E. Yegorova. A novel metric on partitions for image segmentation. IEEE International Conference on Video and Signal Based Surveillance, avss, p. 18, 2006.

[6] D. Kinoshenko, V. Mashtalir, V. Shlyakhov. A Partition metric for clustering features analysis. International Journal "Information Theories and Applications", Vol. 14, 2007, 7p.(will be published).

[7] A. Chupikov, D. Kinoshenko, V. Mashtalir, K. Shcherbinin. Image retrieval with segmentation-based query. In: Adaptive multimedia retrieval. S. Marchand-Maillet et al. (Eds.), Springer-Verlag Berlin Heidelberg, Lecture Notes in Computer Science, Vol. 4398, 2007, pp. 208 – 222.

Authors' Information

Kinoshenko Dmitry – *Kharkov National University of Radio Electronics, Lenin Ave., 14, Kharkov, Ukraine, 61166, kinoshenko@kture.kharkov.ua*

Mashtalir Sergey – *Kharkov National University of Radio Electronics, Lenin Ave., 14, Kharkov, Ukraine, 61166, mashtalir_s@kture.kharkov.ua*

Shcherbinin Konstantin – *Kharkov National University of Radio Electronics, Lenin Ave., 14, Kharkov, Ukraine, 61166*

Yegorova Elena – *Kharkov National University of Radio Electronics, Lenin Ave., 14, Kharkov, Ukraine, 61166, yegorova@kture.kharkov.ua*

ENHANCED FEATURE VECTOR SET FOR VQ RECOGNIZER IN ISOLATED WORD RECOGNITION

Poonam Bansal, Amita Dev, Shail Bala Jain

Abstract: Speech recognition is always looked upon as a fascinating field in human computer interaction. It is one of the fundamental steps towards understanding human cognition and their behavior. This paper explicates the theory and implementation of Automatic Speech Recognition (ASR), in the form of speaker-dependent limited vocabulary isolated word recognizer (IWR). Any IWR contains two main phases, training phase and the testing phase. In the training phase feature extractor transforms the raw speech signal into a compact but effective representation and the extracted features are stored in the database. During the recognition phase the features are extracted by the same or different techniques and are compared with the stored one in the database [1]. In the purposed IWR the features of the speech are extracted as LPCC, Mel-frequency Cepstrum coefficients (MFCC), delta mfcc (DMFCC) and delta-delta mfcc (DDMFCC). Vector Quantization (VQ)[3] is used for word modeling process. The final recognition decision is made based on the matching score: word model with the smallest matching score is selected as a word of the test speech sample. Word recognition rate was observed to be 96% with the 20 MFCC coefficients and as we increase the feature vector size to 36 by including DMFCC and DDMFCC recognition rate increase to 99.3%. Better performances could be seen when applying this approach itself or mixed with Hidden Markov Model (HMM) in isolated-word speech recognition.

Keywords: *Speech recognition, Feature extraction, MFCC, DMFCC, DDMFCC and VQ*

Introduction

The Isolated Word Recognition systems were among the first speech recognition systems implemented due to rather straightforward manner in which the basic recognition units – the words can be modeled. In this paper, we will show a step-by-step approach in building such a system, the system which integrates all the stages of a speech recognition process: speech signal acquisition, parametrization, word models building and words recognition, using Vector Quantization. The paper will be structured around the two main stages of a speech recognition process: Feature extraction after the acoustic analysis of the speech signal, and the Feature matching for recognition of the basic units used in training the system (in our case, words).

Feature Extraction

The purpose of this step is to convert the speech waveform to some type of parametric representation (at a considerably lower information rate) for further analysis and processing [1, 2] which is referred as the *signal-processing front end*. The speech signal is a slowly time-varying signal (called *quasi-stationary*). When examined over a sufficiently short period of time (5 ~ 100 ms), its characteristics are fairly stationary. However, over long periods of time (on the order of 1/5 seconds or more) the signal characteristic change to reflect the different speech sounds being spoken. Therefore, the *short-time spectral analysis* is the most common way to characterize the speech signal. A wide range of possibilities exist for parametrically representing the speech signal for the speech recognition task, such as Linear Prediction cepstral coefficients (LPCC), Mel-Frequency Cepstrum Coefficients (MFCC) and others. MFCC and LPCC are well known techniques used in any ASR to describe signal characteristics, relative to the word discriminative acoustic properties. MFCC are based on the filtering of spectrum using properties of human speech perception mechanism. On the other hand, LPCC are based on the autocorrelation of the speech frame. There is no general agreement in the literature about what method is better. However, it is generally considered that LPCC are computationally less expensive while MFCC provide more precise result. The reason of such opinion is based on that all-pole model used in the LPC provides

a good model for the voiced regions of speech and quite bad for unvoiced and transient regions. The main drawback of LPCC is that unlike MFCC it does not resolve the vocal tract characteristics from the glottal dynamics, which vary from word to word and might be useful in IWR. MFCC is perhaps the best known and most popular, and it will be used in this paper. MFCC is based on the known variation of the human ear's critical bandwidths with frequencies, filters spaced linearly at low frequencies and logarithmically at high frequencies have been used to capture the phonetically important characteristics of speech. This is expressed in the *Mel-frequency* scale, linear frequency spacing below 1000 Hz and a logarithmic spacing above 1000 Hz.

Mel-frequency cepstral coefficients (MFCC)are widely used in automatic speech recognition systems. The signal is passed through a mel spaced filterbank (based on FFTs), converted to a logarithmic scale, and then submitted to a cosine transform. MFCC provide a substantial data reduction, because a few coefficients are sufficient to represent the *cepstrum* of the acoustic signal.

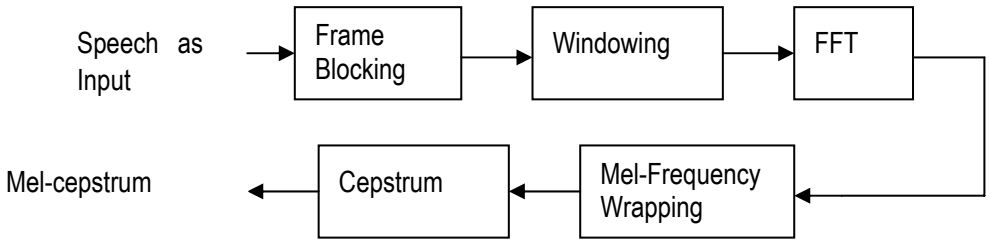


Fig.1 Block diagram of the MFCC processor

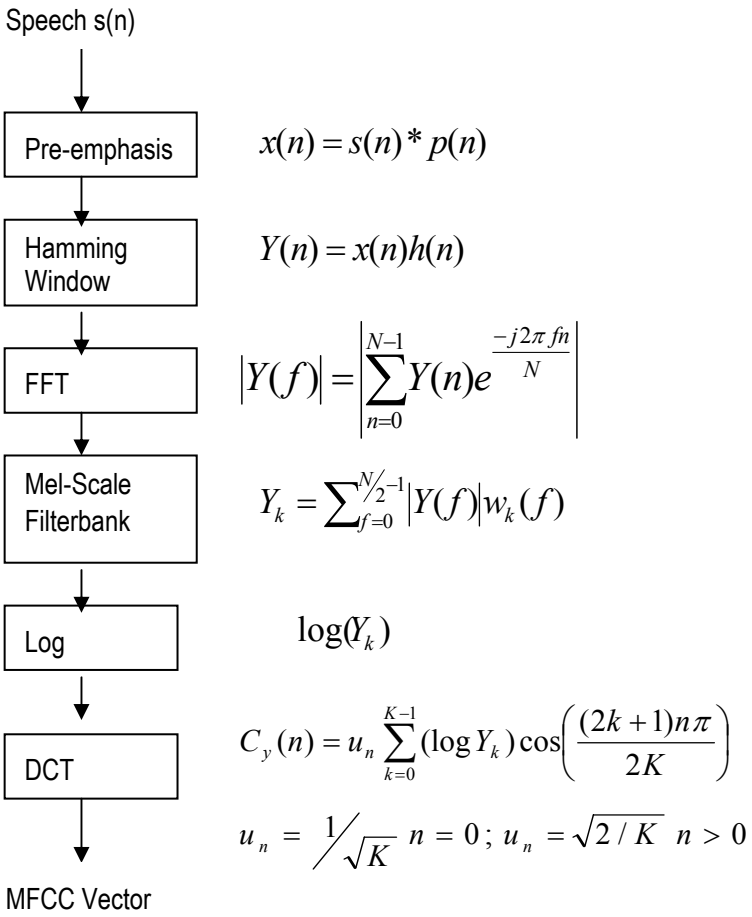


Fig.2 Block diagram for feature extraction

Fig 2 shows the details of the speech feature extraction through MFCC's. The step by step procedure is as follows.

In this step the continuous speech signal is blocked into frames of *N* samples, with adjacent frames being separated by *M* (*M* < *N*). There is a overlapping of (N-M) samples. This process continues until all the speech is accounted for within one or more frames. Typical values for *N* and *M* are *N* = 256 (which is equivalent to ~ 30 msec windowing and facilitate the fast radix-2 FFT) and *M* = 128.

Windowing

The next step in the processing is to window each individual frame so as to minimize the signal discontinuities at the beginning and end of each frame, window to taper the signal to zero at the beginning and end of each frame. A hamming window function is used.

$$W(n) = 0.54 - 0.46Cos\left(\frac{2\pi n}{N-1}\right) \qquad 0 \le n \le N-1$$

FFT

The next processing step is the Fast Fourier Transform, which converts each frame of *N* samples from the time domain into the frequency domain. The FFT is defined on a set of N samples X_n as:

$$X_n = \sum_{k=0}^{N-1} x_k e^{-2\pi jkn/N} \qquad n = 0,1,2,\ldots..,N-1$$

Mel-Frequency Wrapping

Human perception of the frequency contents of sounds for speech signals does not follow a linear scale. Thus for each tone with an actual frequency, *f*, measured in Hz, a subjective pitch is measured on a scale called the 'mel' scale. The *mel-frequency* scale is linear frequency spacing below 1000 Hz and a logarithmic spacing above 1000 Hz. Our's approach to simulate the subjective spectrum is to use a filter bank, one filter for each desired mel-frequency component. That filter bank has a triangular bandpass frequency response, and the spacing as well as the bandwidth is determined by a constant mel-frequency interval. The Mel scale filter bank is a series of L triangular bandpass filters that have been designed to simulate the band pass filtering believed to occur in the auditory system. This corresponds to series of band pass filters with constant bandwidth and spacing on a Mel frequency scale.

Cepstrum

In this final step log Mel spectrum is converted back to time. The result is called the Mel frequency cepstrum coefficients (MFCC). The Discrete Cosine Transform is done for transforming the Mel coefficients back to time domain.

$$C_n = \sum_{k=1}^{K} (\log \tilde{S}_k) Cos\left[n\left(k - \frac{1}{2}\right)\frac{\pi}{K}\right], \quad \text{n} = 1,2 \ \ldots\ldots\ \text{K}$$

where as $\tilde{S}_k$, $\text{k} = 1,2, \ldots..\text{K}$ are the outputs of the last step.

Feature Matching

The problem of word recognition belongs to a much broader topic in scientific and engineering so called *pattern recognition*. The goal of pattern recognition is to classify objects of interest into one of a number of categories or classes. The state-of-the-art in feature matching techniques used in word recognition includes Dynamic Time Warping (DTW), Hidden Markov Modeling (HMM), and Vector Quantization (VQ). As the data base used is limited in our case, we have chosen VQ approach.

Vector Quantization

VQ is a process of mapping vectors from a large vector space to a finite number of regions in that space. Each region is called a *cluster* and can be represented by its center called a *codeword*. The collection of all codewords is called a *codebook* for a known word. Vector quantization (VQ) is a lossy data compression method based on the principle of block coding. It is a fixed-to-fixed length algorithm. In the earlier days, the design of a vector quantizer (VQ) is considered to be a challenging problem due to the need for multi-dimensional integration

VQ Design

The VQ design can be stated as follows. Given a vector source with its statistical properties known, given a distortion measure, and given the number of code vectors, we can find a codebook and a partition which result in the smallest average distortion.

We assume that there is a *training sequence* consisting of M source vectors:

$$T = \left\{ x_1, x_2, x_3, \ldots\ldots, x_M \right\}$$

This training sequence can be obtained from some large database. M is assumed to be sufficiently large so that all the statistical properties of the source are captured by the training sequence. We assume that the source vectors are K-dimensional, e.g.,

$$X_m = \left\{ x_{m,1}, x_{m,2}, x_{m,3}, \ldots\ldots, x_{m,k} \right\}, \qquad m = 1, 2, 3 \ldots\ldots M$$

Let N be the number of code vectors and let $C = \left\{ c_1, c_2, c_3, \ldots\ldots, c_N \right\}$ represents the codebook.

Each code vector is K-dimensional, e.g.,

$$c_n = \left\{ c_{n,1}, c_{n,2}, c_{n,3}, \ldots\ldots, c_{n,k} \right\}, \qquad n = 1, 2, 3 \ldots\ldots N$$

Let S_n be the encoding region associated with code vector C_n and let $P = \left\{ S_1, S_2, S_3, \ldots\ldots, S_N \right\}$.

Denote the partition of the space. If the source vector X_m is in the encoding region S_n, then its approximation (denoted by Q (X_m)) is C_n: $Q(x_m) = c_n \quad , \quad \text{if } x_m \in S_n$

Assuming a squared-error distortion measure, the average distortion is given by:

$$D_{ave} = \frac{1}{Mk} \sum_{m=1}^{M} \left\| x_m - Q(x_m) \right\|^2$$

where $\left\| e \right\|^2 = e_1^2 + e_2^2 + \ldots + e_k^2$

The design problem can be succinctly stated as follows: Given T and N find C and P such that D_{ave} is minimized. If C and P are a solution to the above minimization problem, then it must satisfy the following two criteria.

Optimality Criteria

Nearest Neighbor Condition:

$$S_n = \{ x : \left\| x - c_n \right\|^2 \le \left\| x - c_{n'} \right\|^2 \ \forall n' = 1, 2, \ldots, N \}$$

This condition says that the encoding region S_n should consists of all vectors that are closer to C_n than any of the other code vectors. For those vectors lying on the boundary, any tie-breaking procedure will do.

Centroid Condition:

$$c_n = \frac{\sum_{x_m \in S_n} x_m}{\sum_{x_m \in S_n} 1} \qquad n = 1, 2, \ldots, N$$

This condition says that the code vector C_n should be average of all those training vectors that are in encoding region S_n. There is a well-know algorithm, namely LBG algorithm [Linde, Buzo and Gray], for clustering a set of *L* training vectors into a set of *M* codebook vectors.

Data Set

1. Language:	Standard Hindi
2. Vocabulary size:	A set of 1000 most frequently occurring Hindi words
3. Number of Speakers:	50 (30 Male & 20 Female)
4. Average duration of training and Testing utterances:	500-800 msec.
5. Audio recording:	S/N > 40 db
6.Sampling and quantization:	16Khz, 16-bit

Experimental Results

The performance of the word recognizer was evaluated in terms of recognition rate. We have used the following recognition measure for computing the recognition rate.-

Recognition rate (%) = $N_c / N_T * 100$-

Where N_c is the No. of words correctly recog-nized, and N_T is the Total No. of words used in the testing session.

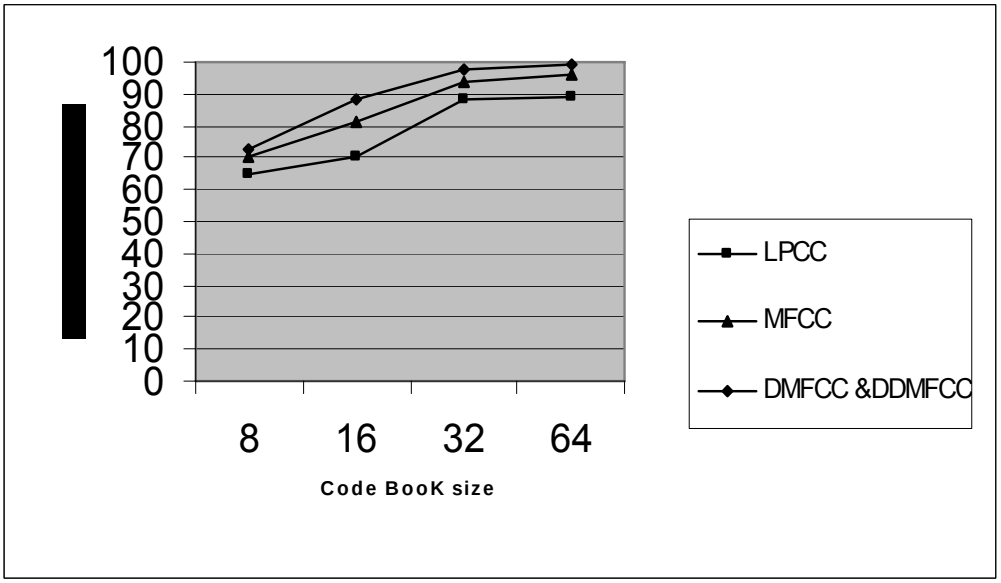


Fig. 3. Performance of Recognition rate with codebook size

It was found that out of 1000 words, which were taken in the training and testing session, our IWR was able to recognize with an efficiency of 70% having codebook size of 8 and as we go on increase the size of codebook (upto 64) the recognition efficiency goes upto 96%. Another analysis was done with improved feature vector set,

now we reduced the no. of MFCC coefficients to 12 and included DMFCC and DDMFCC to the whole feature set of each word. Although the size of feature vector increased a bit but with the same codebook sizes we were able to get the better recognition rate ranging from 73% for codebook of 8 to 99.3% for the codebook size of 64. Results compared are shown in the Fig3.

Conclusion

As MFCC take care of the vocal track characteristics from the glottal dynamics, it proved to be better in our case for recognizing isolated words, in comparison with LPCC by 7% with the codebook size of 64. By enhancing the feature vector set with DMFCC and DDMFCC performance of the recognizer gets improved by 3.3%, in comparison with when only MFCC were considered. If we want to test the viability of the IWR with the larger database for a optimum size of codebook then the recognition rate can be further improved by taking into consideration the other statistical parameters of the MFCC coefficients i.e. Average, Standard deviation etc.

Bibliography

[1] L.R. Rabiner and B.H. Juang, Fundamentals of Speech Recognition, Prentice-Hall, Englewood Cliffs, N.J.Prentice-Hall, 1993.

[2] L.R Rabiner and R.W. Schafer, “Digital Processing of Speech Signals”, Prentice-Hall, Englewood Cliffs, N.J., 1978.

[3] Y. Linde, A. Buzo & R. Gray, “An algorithm for vector quantizer design”, IEEE Transactions on Communications, Vol. 28, pp.84-95, 1980.

[4]“ The past, present and future of speech processing,” IEEE Signal Processing Magazine, May 1998.

[5] Douglas 0’Shaughnessy, “Speech Communications(Human and Machine)”, 2nd edition, University Press.

[6] A.Biem and S. Katagiri, “Cepstram-based filter-bank design using discriminative feature extraction training at varioius levels,” in Proc. IEEE Int. Conf. Acoustics, Speech, and Signal processing, 1997, pp.1503-1506

[7] S. Furui, “Speaker independent isolated word recognition using dynamic features of speech spectrum”, IEEE Transactions on Acoustic, Speech, Signal Processing, Vol. ASSP-34, No. 1, pp. 52-59, February 1986.

[8] Chulhee Lee, Donghoon Hyun, Euisun Choi, Jinwook Go and Chungyoung Lee, “Optimizing feature extraction for speech recognition”,IEEE Transactions on Speech and audio processing, Vol 11. No.1, January 2003.

[9] John G. Proakis and Dimitris G. Manolakis, “Digital signal Processing, principal, Algorithms, and applications” Prentice hall of India

[10] S. Haykin, “Adaptive filter theory”, Pearson Education publication, 4th Edition (2002).

Authors' Information

Poonam Bansal - *Department of Computer Science and Engineering, Amity School Of Engineering and Technology, 580, Delhi Palam Vihar Road, Bijwasan, New Delhi 110061, India ; e-mail: pbansal89@yahoo.co.in*

Amita Dev – *Ambedkar Institute of Technology, Madhuban, Delhi – 110092 , India;*
e-mail: amita_dev@hotmail.com

Shail bala jain – *Mahila Institute Of Technology , G.G.S.I.P. Univ., Old DCE Campus, Kashmere Gate, New Delhi -110006 , e-mail: shailbala.jain@gmail.com*

WAVELET TRANSFORMATION IN ELECTROCARDIOGRAM PROCESSING

Elena Visotska, Olga Kozina, Sophia Nuzhnova, Andrei Porvan, Constantine Chebanov, Maxim Konovalov

Abstract: *Decisions of problems during forecasting and early diagnostics of a heart attack, also in-time revealing of fatal infringements of heart rhythm, the prevention sudden coronary death are considered in the given work. Successful application biorthogonal wavelet-transform has allowed statistically characterize not only ECG-signals and wavelet-components but also to receive with authentic accuracy distinctive characteristics of two compared ECG-signals.*

Keywords: *electrocardiology, wavelet-transform.*

Introduction

In spite of current cardiology achievements in acute coronary syndrome diagnostics, mortality of patients after infarction heart attack is very high. Different methods of digital signal processing are applying in electrocardiography for discovering, selection and analysis of various parts of ECG. Among such methods wavelet transformation gives much promising results in time-frequency characteristics analysis of ECG.

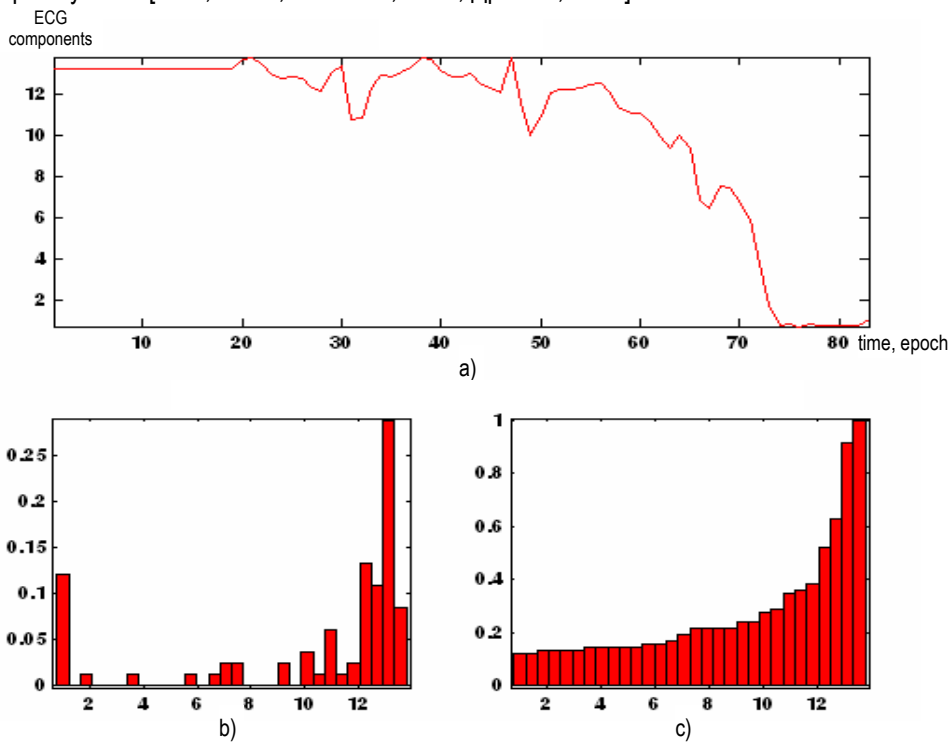
Instructions for Manuscripts Preparation

In current time is possible not only to establish and to determine of a heart attack but to make representation about size of a heart attack on an extent, and also about its "depth", i.e. about the greater or smaller distribution necrotic process in thickness of a wall of a myocardium.

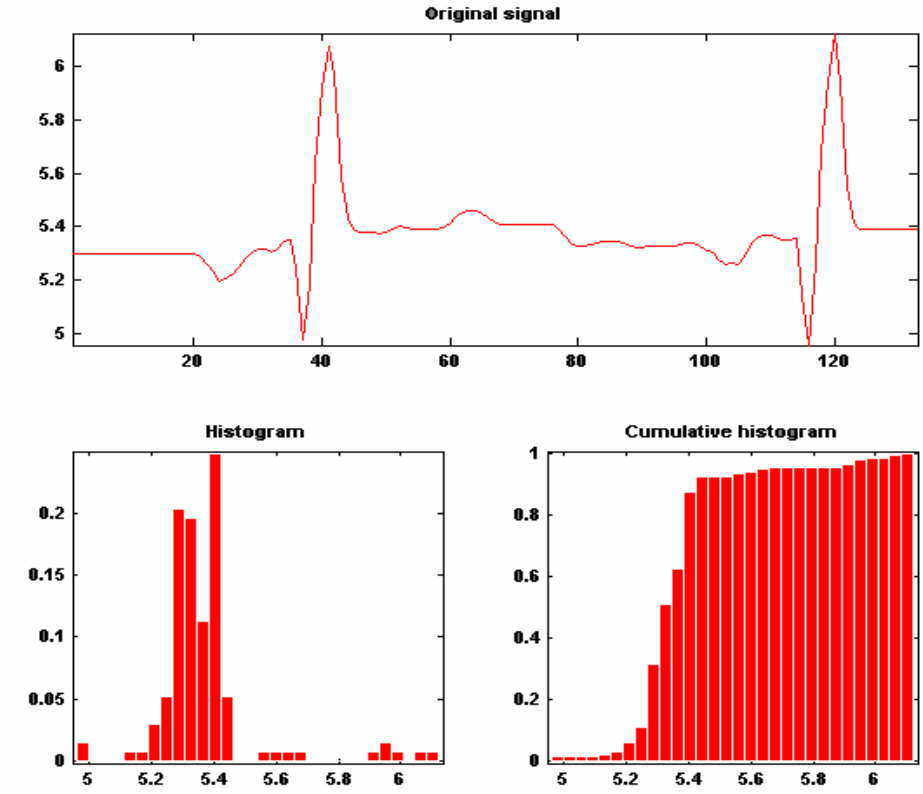
Serial ECG-research enables to watch dynamics of process and to predict a situation. However the problem consists in the prevention sudden coronary death. The classical approach in electrocardiology is use of techniques of signal analysis in time area which have various applications (standard ECG-measurement, measurement of hard rate, dispersion of repolarization). However measurements of amplitude and duration the ECG components due to methods of ECG analysis in time area are not always sufficient for the description of all features of the ECG signal. For example, definition of late potential located in QRS complex, may not be executed with use of these methods. At the same time the heart rate analysis in time area gives the full information about behavior of RR-intervals and parasympathetic influence. But sympathetic ordering can not be appreciated on the basis of heart rate measurements in time area. Thus, using of analysis in both times and frequency areas together give universal results [Simson, 1992].

Frequency representation of the signal can be received with use of various techniques, including Fourier transform. Most frequently in electrocardiology it is used fast Fourier transform (FFT) which represent a time signal (theoretically it should be periodic) on infinite number of sinusoids. This set of sinusoids then is transformed in frequency area by using of amplitude and phase of each of these functions. Thus, FFT provides relation between time and frequency representation of the signal [Zareba, 1994, Khadra, 1993]. Digital ECG-signal is finite therefore it has precise borders. That is the reason of washing out of certain frequencies. In order to avoid this, during FFT it is applied to windowed Fourier transform for smooth reduction of ECG-signal border up to zero with removal of its intermittence. Restriction of this approach is a reduction of frequency resolution and decreases quality of definition of ECG-signal frequencies. Other inevitable restriction of Fourier transform is – it does not allow determining exact position of frequency components in non-stationary ECG-signal. For example, *QRS*-complex is high-frequency component whereas *T* peak contains low-frequency components. For non-stationary signals, whose spectral content change in time, Fourier representation is not appropriate. The wavelet transform (WT) provides varying time and frequency resolutions by using windows of different lengths [Polikar, 1999]. A key advantage of wavelet techniques is the variety of wavelet functions available, thus allowing the most appropriate to be chosen for the signal under investigation. This is in contrast to Fourier analysis which is restricted to one feature morphology: the sinusoid [Watson, 2005]. Everyone wavelet has the certain duration,

time position and frequency band. Wavelet-parameters corresponds to ECG-components on a certain time interval and frequency band [Cain, 1985, Dickhaus, 1994, Дремин, 2001].



Pic.1. Scalogram a), its histogram b) and cumulative histogram c) of 1 lead on ECG of patients who will die after heart attack



Pic.2 Scalogram a), its histogram b) and cumulative histogram c) of 1 lead on ECG of patients who will survive after heart attack

72 patients in the moment before start of heart attack were selected for the analysis. From them 48 electrocardiograms was from of patients who will die after heart attack and 25 will survive with the same diagnosis. Further, each of received digital ECG-signal was transformed by seven discrete wavelet types: Daubechies wavelet, Symlet wavelet, Haar wavelet, Coiflet wavelet, biorthogonal wavelet, reverse biorthogonal wavelet, discrete approximating Meyer wavelet. Each type, in turn, is broken on seven subspecies.

The most informative WT for the analysis of 1 standard lead on ECG-signal appeared biorthogonal wavelet (see pic.1,2).

Comparison of wavelet-components dispersions allows as to distinguish ECG-signals statistically for patients who can die or can survive after heart attack and to determine a level of mortality risk from a heart attack at the earliest stages.

Conclusion

Analysis of ECG in time area by biorthogonal wavelet represents the new effective approach to definition of quantitative distinctions of signals for patients who have acute coronary insufficiency during early diagnostics and acute coronary syndrome and preventive maintenance of sudden coronary death. As application of wavelet-transform in ECG analysis – rather new area of research, many methodological aspects (mother wavelet choice and scale) demand the further researches for increase of clinical efficiency. Diagnostic and predicting importance of this technique in electrocardiology demands large clinical researches.

Bibliography

[Simson, 1992] Simson M.B. // Circulation.- 1992.- Vol. 85(Suppl).- P1145-1151.

[Zareba, 1994] Zareba W.et al.// J. Electrocardiol.- 1994.- Vol. 27(Suppl).- P. 66-72.

[Cain, 1985] Cain M.E. et al. // Am. J. Cardiol.- 1985.- Vol. 55.- P. 1500-1505.

[Khadra, 1993] Khadra L. et al. // J Med Engineering & Technology.- 1993.- Vol. 17.- P. 228-231.

[Dickhaus, 1994] Dickhaus H. et al. // Meth. Info Med.- 1994.- Vol. 33.- P. 187-195.

[Дремин, 2001] Дремин И.М. и др. // УФН.- 2001.- Т. 171, N 5.- С. 465-501.

[Polikar, 1999] Polikar R. The story of wavelets, in Physics and Modern Topics in Mechanical and Electrical Engineering, (ed. Mastorakis, N), pp. 192-197, World Scientific and Eng. Society Press, 1999.

[Watson, 2005] Watson J. N. et al. //Meas. Sci. Technol. -2005.- Vol.16. – P. L1–L6

Authors' Information

Elena Visotska *– PhD, lecturer of Biomedical Electronic Devices and Systems Department of Kharkov National University of Radio Electronics*

Olga Kozina *– PhD, lecturer of Computers and Programing Department of National Technical University 'KPI'*

Andrey Porvan *– engineer of Biomedical Electronic Devices and Systems Department of Kharkov National University of Radio Electronics*

Sophia Nuzhnova *– engineer of Economic Cybernetics Department of Kharkov National University of Radio Electronics*

Constantine Chebanov *– student of Biomedical Electronic Devices and Systems Department of Kharkov National University of Radio Electronics*

Maxim Konovalov *– student of Biomedical Electronic Devices and Systems Department of Kharkov National University of Radio Electronics*

Kharkov National University of Radio Electronics, Ukraine, 61166, Lenin Avenue, 14, Biomedical Electronics Devices and Systems Department, e-mail: diagnost@kture.kharkov.ua

SMART PORTABLE FLUOROMETER FOR EXPRESS-DIAGNOSTICS OF PHOTOSYNTHESIS: PRINCIPLES OF OPERATION AND RESULTS OF EXPERIMENTAL RESEARCHES

Volodymyr Romanov, Volodimyr Sherer, Igor Galelyuka, Marina Kachanovska, Yevgeniya Sarakhan, Oleksandra Skrypnyk

Abstract: *In the Institute of Cybernetics of National Academy of Sciences of Ukraine the smart portable fluorometer for express-diagnostics of photosynthesis was designed. The device allows easy to estimate the level of influence of natural environment and pollutions to alive plants. The device is based on real time processing of the curve of chlorophyll fluorescent induction. The principles of operation and results of experimental researches of device are described in the article.*

Keywords: *Kautsky effect, chlorophyll, chlorophyll fluorescence induction, fluorescence, fluorometer, portable device, vine plant.*

ACM Classification Keywords: *J.3 Life and Medical Sciences - Biology and Genetics*

Introduction

Development of information technologies and microelectronic circuits allows filling of the world market with portable computer devices such as handheld PCs, laptops, media players, medical devices (tonometers, glucometers, cardiographs), navigation devices and so on. The achievements of Ukrainian scientists who work in the field of biosensors combined with modern capabilities of information technologies provided development of devices for express-diagnostics of plant state, evaluation of environmental parameters, exposure of infective diseases etc.

In the context of the program of Presidium of NAS of Ukraine "Development in the field of sensor systems and technology" in Glushkov's Institute of cybernetics of NAS of Ukraine the portable computer device was developed for express-diagnostics of stress factors which influence on the plant's state. The portable device measures chlorophyll fluorescence induction (CFI) without plant destruction. Using the curve of CFI (alike the cardiogram) allows diagnosing influence of one or other influential factor on the plant's state.

Features of biological objects' luminescence

As a result of external influence, different objects, including biological ones, can generate plenty of radiation that is independent of these objects temperature.

All the types of radiation that were caused by some external sources of energy are called luminescence. Duration of luminescence after external influence stopping exceeds period of light fluctuations. Luminescence is conditioned by fluctuations of relatively small number of atoms or molecules of substance that become excited under energy source activity. Radiation is a result of transformation of atoms' or molecules' states into fundamental (unexcited) or less excited (they have less energy) states.

This is well adjusted with quantum theory, according to what every stationary orbit conforms to definite value of atom's energy (Bore's postulate). Being placed on stationary orbits an electron doesn't radiate and doesn't absorb electromagnetic waves. According to the second Bore's postulate radiation and absorption can happen only when atom changes its state from one stationary state to another:

$$h\varpi_{mn} = h\nu_{mn} = E_n - E_m, \quad (1)$$

where ϖ_{mn} or ν_{mn} – photon's frequency, E_m, E_n – energy values of the states *m* and *n*, h – Planck's constant, *m* and *n* – the numbers of energy states. At the same time electron switches from one stationary orbit to another.

Luminescence is defined by the structure of substance energy spectrum, the average time of staying in excited states and rules of selection, which allow absorption or radiation of light of defined frequency. Short-timed luminescence is also called fluorescence. Luminescence which appears during lighting of substance (phosphor) with visible or ultraviolet light is called photoluminescence. Usually process of luminescence satisfies Stocks' rule that claims that wave length λ' of radiated light is greater than wave λ of excited light. According to the quantum theory this means that photon's energy $h\varpi(h\nu)$ is used partially for non-optical processes:

$$h\varpi = h\varpi' + E, \varpi > \varpi', \tag{2}$$

where ϖ' – luminescence's frequency, E – energy waste on another process.

Luminescence is characterized by energy output which equals to ratio of luminescence energy to energy that was absorbed by substance under stationary conditions.

Energy efficiency of photoluminescence increases proportionally to wave length λ of absorbed light up to the definite maximum value at $\lambda = \lambda_{\max}$ and then rapidly decreases to zero at $\lambda > \lambda_{\max}$ (Vavilov's rule). A sharp decrease of energy at $\lambda > \lambda_{\max}$ is explained by the fact that at these wave lengths λ the energy of absorbed photons is not enough for the process of phosphor atoms and molecules transfer to the excited states.

Ratio of luminescence photons number to absorbed photons with fixed energy is called quantum yield of photoluminescence. According to Vavilov's rule, which is under Stocks' rule, quantum yield of photoluminescence doesn't depend on wave length of excited light and rapidly decreases for anti-Stocks radiation.

Intensity of luminescence *I* depends on behavior of elementary processes that causes this radiation. In case of spontaneous luminescence, when radiation starts after light absorption during which atoms or molecules are transmitted to the excited level that is placed higher than the level at which radiation takes place and then these atoms (molecules) are transmitted to the luminescence level, intensity is subordinate to exponential rule

$$I = I_0 \exp(-t/\tau), \tag{3}$$

where I – lighting intensity at the moment *t*, I_0 – lighting intensity in a moment of excited radiation stopping, $\tau \approx 10^{-9} - 10^{-8}\tilde{n}$ – an average duration of excited state of phosphor atoms or molecules. Luminescence of compound molecules and phosphorescence (after lighting) of organic substance are subordinate to the low (3).

Under influence of light there can be happened photochemical transformation of substance (including photosynthesis), which is called photochemical reactions. In a process of such reactions light absorption takes place. Energy is spent on compound molecules and polyatomic ions decomposition to component parts and creation of compound molecules of primary ones. An example of photochemical reactions is decomposition carbon dioxide under influence of light

$$2CO_2 + 2h\varpi \rightarrow 2CO + O_2 \tag{4}$$

Carbon dioxide decomposition takes place in green parts of plants under sun light influence, as photochemical process, which is a part of photosynthesis.

Principles of Operation

One of the most important properties of the molecule of chlorophyll which is the basic pigment of plant cell is ability to fluoresce. For the first time this phenomenon was researched by Kautsky [Kautsky, 1931], [Kautsky, 1937]. Dependence of chlorophyll fluorescence induction on time passed after start of lightning of plant's leaves is known as an induction curve or a chlorophyll fluorescence induction curve (Fig. 1). The form of this curve is rather

sensible to changes in the photosynthetic apparatus of plants during adaptation to different environmental conditions. This fact is a basic for extensive usage of Kautsky effect in photosynthesis research. The advantages of the method of CFI are the following: high self-descriptiveness, expressiveness, noninvasiveness and high sensibility.

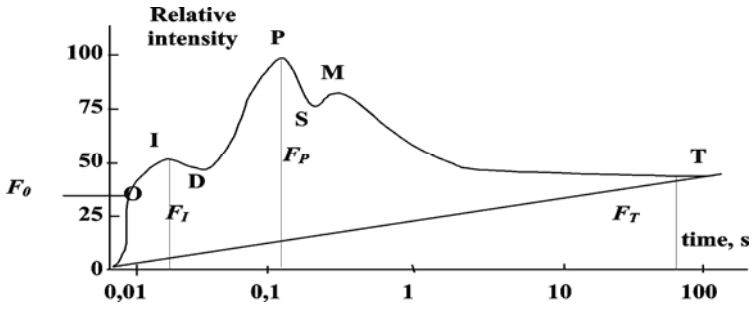


Figure 1. Chlorophyll fluorescence induction curve

The organization, scheme, basic components and advantages of the portable computer fluorometer "Floratset" are discussed in [Fedack, 2005] in detail.

Results of Experimental Research

The experimental researches of the "Floratest" were conducted in National Scientific Center "V.E. Tairov's Institute of viticulture and winemaking" of Academy of Agrarian Sciences of Ukraine. The conditions and results of the experimental researches are listed below.

Mature leaves of vine were used in the researches. Under changes of soil watering conditions there were observed sharp changes in behavior of induction transitions of chlorophyll fluorescence which were accompanied by quite essential changes of leaf tissue spectral characteristics.

Determination of fluorescence spectral characteristics was done by placing the device's sensor on the leaf's surface without integrity disturbance directly in a pot or in a field. It allowed to research on plastid and vacuolar pigments in their natural state and in that way approaching to understanding of the biophysical and physiology-biochemical processes which take place in the live leaf, and determination of important sides of photosynthetic activity.

Fluorescence intensity of the sample was determined in relative units.

It is significant that under natural conditions in the middle latitudes the drought is accompanied simultaneously by high temperatures of air, and that intensifies bad influence of ground water lack on agricultural plants.

Even in the first variant of experiment (drought) there appeared considerable changes of the behavior of fluorescence induction comparing to the control samples. Changes show in weakening of penetrability of the chloroplasts' membrane structures. That results in substantial increase of time characteristics of fluorescence induction slow decrease. At the same time noticeable variety differences become apparent. Sharp decrease of its value is typical for profound functional injuries of photosynthetic structures and cells of particular variety entirely.

Accordingly in this stage of drought influence significant variety differences in exsiccate factor resistance of both photosynthetic structures and lamina's parenchymal cells entirely became apparent.

More deep changes of destructive nature may be observed in case of high temperatures (+40 °C), which influence on leaves complementary to drought. In this case for all the varieties being studied significant and almost irreversible functional changes of plastid structures are noted. These functional changes show in sharp decrease of CFI intensity.

Disastrous changes of life activity of vine leaf cells which take place during these processes show in oppression of biosynthetic processes, intensive decomposition of cytoplasmic structures and intensification of oxide

catabolism of plant cell's content. The consequence of these processes is decrease of CFI intensity as a result of its oxidizing transformation.

Diagrams of measuring of chlorophyll fluorescence intensity for vine plants under drought conditions and normal conditions accordingly are displayed on figures 2, 3.

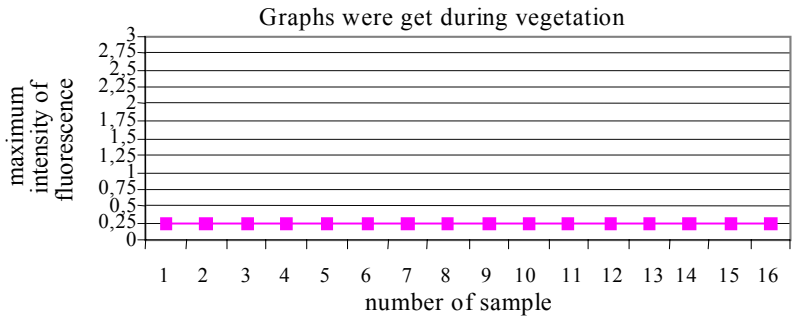


Figure 2. Maximum of CFI intensity of vine plant under drought influence (28-30 % insufficient water capacity)

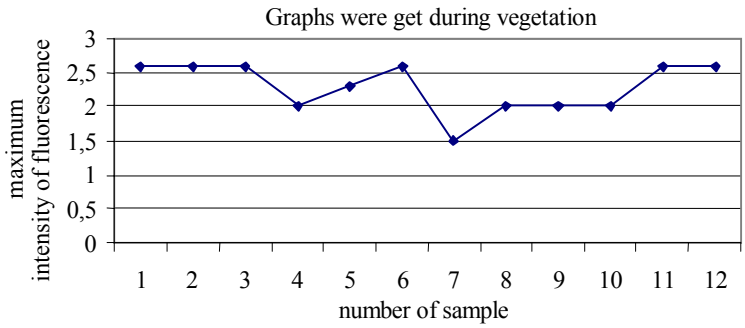


Figure 3. Maximum of CFI intensity of vine plant, control samples (68-70% insufficient water capacity)

Thus, a water deficit (WD) shows up on the Kautsky curve (figure 1) as difference of fluorescence ($F_p - F_0$) decrease. The most credible reason of this is oppression of oxygen emission which is related with slowing down of electrons transfer. Assuming that F_0 almost does not change for the test and control plants, in a maximum point the chlorophyll fluorescence intensity value can define the level of water deficit. Examples of the practical usage of fluorometer "Floratest" in the National Scientific Center "V.E. Tairov's Institute of viticulture and winemaking" and the graph of CFI on the device's display are shown on figure 4.

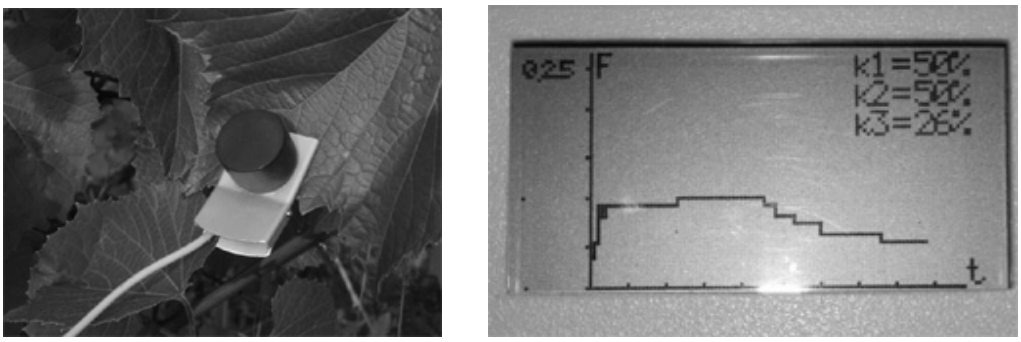


Figure 4. The sensor of the "Floratest" on the vine leaf and the image of CFI on the device's display

Experimental researches of fluorometer "Floratest" in National Scientific Center "V.E. Tairov's Institute of viticulture and winemaking" allow:

– determination of vine plants' state under the stress factor influence accordingly to the parameters of Kautsky curve;
– development of the recommendations on the fluorometer's software update and bringing output information on the device display to the recommendations, which accompany the Kautsky curve;
– development of recommendations on creation of the set of removable sensors, using which both detection of stress factors and express-diagnostics of plant disease can be performed.

Conclusions

– on the basis of modern information technologies and achievements in field of biosensorics original noninvasive portable fluorometer for express-diagnostics of plant state under stress conditions was developed;
– during the fluorometer designing and fast software and hardware tools adaptation to the conditions of exploitation the methods of virtual design created in the Institute of Cybernetics of NAS of Ukraine as a part of virtual laboratory for computer-aided design were used extensively;
– during experimental researches in National Scientific Center "V.E.Tairov's Institute of viticulture and winemaking" of Academy of Agrarian Sciences of Ukraine there were developed methodical tools which allow evaluating the state of vine plants under drought conditions and conditions of insufficient water capacity in express-mode.

Bibliography

[Kautsky, 1931] Kautsky H., Hirsch A. Neue Versuche zur Kohlenstoffassimilation // Natur wissenenschaften. – 1931. – 19. – S. 964

[Kautsky, 1934] Kautsky H., Hirsch A. Das Fluoreszenzverhalten gruner Pflanzen // Biochem Z. – 1934. – 274. – S. 422–434.

[Fedack, 2005] Fedack V., Kytaev O., Klochan P., Romanov V., Voytovych I. Portable Chronofluorometer for Express-Diagnostics of Photosynthesis // Proceeding of the Third IEEE Workshop on "Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications", IDAACS'2005. – Sofia, Bulgaria. – 2005, September 5–7. – P. 287–288.

Authors' Information

Volodymyr Romanov *- Head of department of V.M. Glushkov's Institute of Cybernetics of National Academy of Sciences of Ukraine, Doctor of technical sciences, professor; Prospect Akademika Glushkova 40, Kiev–187, 03680, Ukraine; e-mail: dept230@insyg.kiev.ua*

Volodimyr Sherer *– deputy director of National Scientific Center "V.E. Tairov's Institute of viticulture and wine-making" of Academy of Agrarian Sciences of Ukraine; Doctor of agricultural sciences; 40 let Pobeda Str., 27, Tairovo, Odessa, 65496, Ukraine; e-mail: iviv@te.net.ua*

Igor Galelyuka *– research fellow of V.M. Glushkov's Institute of Cybernetics of National Academy of Sciences of Ukraine; Prospect Akademika Glushkova 40, Kiev–187, 03680, Ukraine; e-mail: dept230@insyg.kiev.ua*

Yevgeniya Sarakhan *– research fellow of National Scientific Center "V.E. Tairov's Institute of viticulture and wine-making" of Academy of Agrarian Sciences of Ukraine; 40 let Pobeda Str., 27, Tairovo, Odessa, 65496, Ukraine; e-mail: sarakhan2006@ukr.net*

Marina Kachanovska *– software engineer of V.M. Glushkov's Institute of Cybernetics of National Academy of Sciences of Ukraine; Prospect Akademika Glushkova 40, Kiev–187, 03680, Ukraine; e-mail: dept230@insyg.kiev.ua*

Oleksandra Skrypnyk *– software engineer of V.M. Glushkov's Institute of Cybernetics of National Academy of Sciences of Ukraine; Prospect Akademika Glushkova 40, Kiev–187, 03680, Ukraine; e-mail: dept230@insyg.kiev.ua*

MODELING OPTICAL RESPONSE OF THIN FILMS: CHOICE OF THE REFRACTIVE INDEX DISPERSION LAW

Peter Sharlandjiev, Georgi Stoilov

***Abstract**: Determination of the so-called optical constants (complex refractive index N, which is usually a function of the wavelength, and physical thickness D) of thin films from experimental data is a typical inverse non-linear problem. It is still a challenge to the scientific community because of the complexity of the problem and its basic and technological significance in optics. Usually, solutions are looked for models with 3-10 parameters. Best estimates of these parameters are obtained by minimization procedures. Herein, we discuss the choice of orthogonal polynomials for the dispersion law of the thin film refractive index. We show the advantage of their use, compared to the Selmeier, Lorentz or Cauchy models.*

***Keywords**: Thin films; Materials and process characterization*

***ACM Classification Keywords**: J.2 Physical Sciences and Engineering (Physics)*

Introduction

The problem of estimation of the optical parameters of thin films: physical thickness (D) and complex refractive index N = n – i*k (real refractive index (n) and extinction coefficient (k)) is challenging from mathematical point of view and has technological and scientific importance. Usually, n and k are unknown functions of the wavelength (λ). The task is to evaluate them by the use of measurable quantities, such as film transmittance (T), front side reflectance (R) and/or backside reflectance (R′). Different methods have been proposed but no one has shown yet absolute advantage over the others. We can say that estimation of thin films optical parameters is more of an art, than scientific analysis. There are several steps that have to be fallowed: a) creation of a model, which describes the optical behavior of the film; b) collecting empirical data; c) fitting the postulated model to the data; d) evaluation of the results. The model of the wavelength dependence of the refractive index is of crucial importance: it defines the number of the unknown parameters and their functional relation. Some of the most popular models are named after the scientists that have proposed them: Cauchy, Drude, Selmeier, Lorentz, etc. Cauchy dispersion law is purely empirical:

$$n(\lambda) = A_0 + \frac{A_1}{\lambda^2} + \frac{A_2}{\lambda^4} + \dots .,$$

where A_0, A_1, A_2,… are parameters to be determined. The number of terms can reach 10 – 15. Selmeier dispersion is semi-empirical:

$$n(\lambda) = \sqrt{A_0 + \frac{A_1\lambda^2}{\lambda^2 - B_1^2} + \dots} ,$$

where A_0, A_1, B_1,… are parameters to be determined. More terms can be added for different oscillator positions.

Once the model is assumed, minimization techniques are applied to estimate the unknown parameters to the optical response of the thin film.

Here we shall consider the use of orthogonal polynomials in the dispersion law representation. We shall simulate a measurable quantity (transmittance) with predefined wavelength dependence of the complex refractive index. Then we shall fit the simulated data to different models of refractive index. Parameters in the dispersion law will be estimated, comparing Cauchy, Selmeier and orthogonal polynomials (OP) approaches.

Models and Computational Procedures

We shall consider a thin homogeneous film with wavelength dependence of the complex refractive index and physical thickness of 350 nm. The spectra of n(λ) and k(λ) are shown in Figure 1a and 1b, respectively. The choice of n(λ) and k(λ) is characteristic for many optical materials, such as amorphous semiconductors.

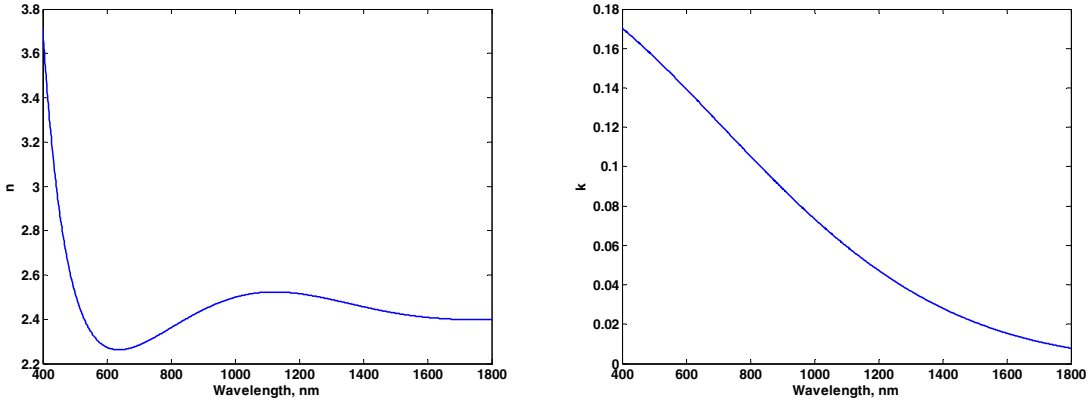


Figure 1. Spectral dependence of the refractive index (a) and extinction coefficient (b)

The simulated measurement is the spectrum of transmission in VIS at normal incidence of light calculated by the help of Abelès characteristic matrix [1], Figure 2. The measurable quantity is T ~ tt* (* stands for complex conjugate). The amplitude transmittance t is a complex quantity, related to N and D by transcendental equations [1]. During the fitting calculations, we have assumed that the physical thickness D and the extinction coefficient k(λ) are known, so that there is no fitting on these quantities. In this way we have strongly reduced parameter interactions. We have used associated Legendre functions *P(m,p;λ)* of degree *p* and order *m = 0,1,...,p*, as orthogonal polynomials. Thus, the dispersion law for n(λ) stands as:

$$n(\lambda) = A_0 P(0, p; \lambda) + A_1 P(1, p; \lambda) + A_2 P(2, p; \lambda) + \ldots.$$

For Cauchy, Selmeier or OP dispersion laws, 8-9 coefficients are need for relevant representation of the refractive index n(λ), shown in Figure1a. The nonlinear data-fitting problem is solved by the Levenberg-Marquardt method (unconstrained minimization) [2].

Results and Discussion

Calculations of the film optical response and preliminary fits showed that Selmeier model of the refractive index has to be disregarded: it cannot describe properly the 'experimental' data, shown in Figure 2. In order to compare the Cauchy and OP models, we have used 8 coefficients in their corresponding presentations, so that the degrees of freedom in the two cases are the same. Levenberg-Marquardt procedures demand initial guess of the unknown parameters. For each fit, we have put 7 of the coefficients equal to zero, while the first one is 20% up of its initial 'true' value. After the termination of the minimization, one step refinement of the estimations is undertaken as well. The results obtained with the two models differ significantly. The residual error of the fit with the Cauchy model is an order of magnitude greater. The residual error of the fit with OP reaches 0.1%, which is equal to the experimental uncertainty of high precision spectral instruments. This means that further improvement of the fit is meaningless. The minimization procedure is much faster in latter case – the consumed CPU time is ~20 times greater for the Cauchy case. In Figure 3 the relative errors of the estimated wavelength dependence of the refractive index are shown. This is an illustration of the performance of the Cauchy and OP models.

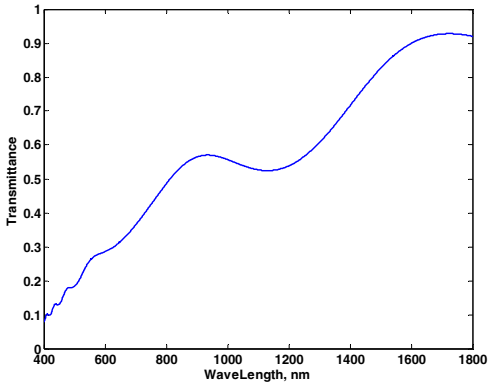


Figure 2. Calculated transmittance of the thin film as 'experimental' data for the fit

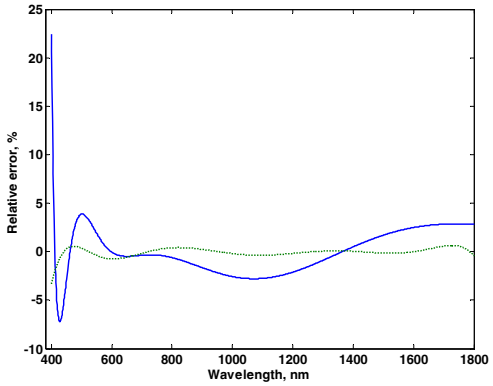


Figure 3. Relative error of refractive index fit: (dots) orthogonal polynomials; (line) Cauchy law

The advantages of fitting orthogonal polynomials to experimental data are well-known [2]. In our case, the situation is more complicated because of the nonlinear functional dependence of the target on the fitting parameters. However, the main advantage of this approach is sustained: due to the orthogonal property, each coefficient in the dispersion law representation can be determined independently from the others. If one has already obtained an evaluation of m^{-th} degree polynomial, an additional term in the dispersion law ((m+1) degree polynomial) requires only one new coefficient to be determined. The other coefficients remain the same, unlike in the Cauchy or Selmeier case. In the Cauchy case, high order polynomials may result in ill conditioned matrices. Besides, the joint confidence region in the parameter space, estimated by the covariance matrix, has minimum volume.

Conclusion

We have shown that the use of orthogonal polynomials in refractive index modeling is effective and highly productive. Although the involved coefficients have no physical meaning, this is also true for the Cauchy and Selmeier models. The OP principal feature is that the number of parameters to be fitted can be kept low at initial steps and then it can be increased, retaining the intermediate results. The orthogonal polynomials approach can be of use in many branches of material science, including photonic crystal design, optimization of elements for effective conversion of solar radiation, etc.

Acknowledgments

This work was partially supported by the National Science Foundation at the Ministry of Education of Bulgaria by grant D01-377/2006.

Bibliography

1. [Born, 1957] E. Born and P. Wolf, Principles of Optics, Ed. Princeton, New York, 1957.
2. [Himmelblau, 1970] D. Himmelblau, Process analysis by statistical methods, Ed. John Wiley & Sons, New York, 1970.

Authors' Information

Peter Sharlandjiev *– Central Laboratory of Optical Storage and Processing of Information, BAS, Acad.G.Bontchev St. bl. 101, Sofia-1113, Bulgaria; e-mail: pete@optics.bas.bg .*

Georgi Stoilov *– Central Laboratory of Optical Storage and Processing of Information, BAS, Acad.G.Bontchev St. bl. 101, Sofia-1113, Bulgaria; e-mail: gstoilov@chittacomputers.com .*

Distributed and Telecommunication Systems

GRID INFRASTRUCTURE FOR SATELLITE DATA PROCESSING IN UKRAINE

Nataliia Kussul, Andrii Shelestov, Mykhailo Korbakov, Oleksii Kravchenko, Serhiy Skakun, Mykola Ilin, Alina Rudakova, Volodymyr Pasechnik

Abstract. *In this paper conceptual foundations for the development of Grid systems that aimed for satellite data processing are discussed. The state of the art of development of such Grid systems is analyzed, and a model of Grid system for satellite data processing is proposed. An experience obtained within the development of the Grid system for satellite data processing in the Space Research Institute of NASU-NSAU is discussed.*

Keywords: *Grid system, satellite data processing, Grid services.*

ACM Classification Keywords: *H.3.4 Systems and Software - Distributed systems, H.3.3 Information Search and Retrieval.*

1 Introduction

Grid systems, originated by Ian Foster [[1], are becoming standard solutions for enabling remote computations execution and distributed data access and processing in environments of the different level of scalability.

The aim of Grid system could be formulated as connecting data, processing powers and algorithms that distributed over the network for solving particular problems. Grid system should be universal up to some degree, so these problems should not be hardcoded during its development. Instead, a set of problems being solved in a Grid environment must be open for modifications and addition of new ones. This goal is achieved by introducing standard interfaces for communicating between different kinds of Grid resources and clients.

Space agencies all over the world are successfully working on development of Grid technology for their application areas. This is due to the fact that Earth observation (EO) domain is characterized by the acquisition of large amounts of data from satellites and distributed nature of data. Furthermore, the single EO product and the data after its initial processing may easily exceed the gigabyte size. Thus, problems of storing, indexing for quick retrieval on application's demand as well as distributed computing arise within the above mentioned area. Grid technology can provide comprehensive solutions for this problem.

In this paper a brief overview of Grid systems for satellite data processing is given. Common approaches and conceptual foundations of development of Earth Observation Grid system are defined. A model of Grid system for satellite data processing is proposed and verified based on a test-bed of Grid system for satellite data processing that was developed in the Space Research Institute of the National Academy of Sciences of Ukraine and the National Space Agency of Ukraine.

2 Overview of Grid Systems for Satellite Data Processing

Nowadays Grid technology is widely applied for the solution of various problems in many domains [[2]. These applications span a wide spectrum. In this section we give a brief overview of Grid systems that are used for satellite data processing.

Earth Science GRID on Demand project [http://eogrid.esrin.esa.int/] is being developed by European Space Agency (ESA) and European Space Research Institute (ESRIN). GRID is considered as a comfortable "open platform" for handling computing resources, data, tools, etc., and not limited to only high performing computing. Online access to different data is enabled within this project, in particular to data provided by various instruments of Envisat satellite [http://envisat.esa.int/], the SEVIRI instrument onboard MSG (the Meteosat Second Generation) satellite, ozone profiles derived from GOME instrument, etc. One of the most important applications is the analysis of long-term data. For example, the analysis of 8 years of GOME on-board temperatures (overall 525 Gb of data) took less than 2 days on 40 computer elements of ESRIN "Grid-on-demand" structure (overall 38460 files were processed).

Grid Web Portal provides access to the "Grid-on-demand" [http://eogrid.esrin.esa.int/] resources enabling:
— Personal certification
— Time /space selection of data, directly from the ESA catalogue
— Data transfer from ESA data storages
— Job selection, launching and live status
— Visualization in OpenGIS Web Map and Google Earth
— Access to user products and documentation

Nowadays "Grid-on-demand" infrastructure consists of more than 150 working nodes with ability to store and handle of about 70 Gb of data. As middleware Globus Toolkit 2.4 and LCG/EGEE components are being used.

Japan Aerospace eXploration Agency (JAXA) [[3] and **KEIO University** started to establish "Digital Asia" system aimed at semi-real time data processing and analyzing. They use GRID environment to accumulate knowledge and know-how to process remote sensing data. The problems of radiometric rectification and composition of remotely sensed data are being solved.

National Aeronautics and Space Administration (NASA) have created **Information Power Grid** (IPG) [[4] targeting an operational Grid environment incorporating major computing and data resources at multiple NASA sites in order to provide an infrastructure capable of routinely addressing larger scale, more diverse, and more transient problems than is possible today. One of the problems being solved is development of techniques for satellite data fusion. Nowadays IPG have approximately 600 CPU nodes of Computing resources and 30-100 Terabytes of archival information/data storage resources.

Spatial Information Grid (SIG), a research project supported by 863 projects of China government, is a series of special grid researches in the filed of Earth Observation. SIG has been designed to be the tested of grid middleware research and grid-enable spatial information services and applications. There are 12 data centers have been involved SIG. The Web Portal has been developed in order to provide access to SIG resources (http://159.226.224.52:6140/Grid/application/index_en.jsp). This portal enables geo-data discover and processing, work monitoring, and grid resources (all service/job/node etc.) management.

3 Why EO Domain requires Grid

In particular, EO domain is characterized by the acquisition of large amounts of data from satellites. For example, an image acquired from ETM+ instrument from the Landsat-7 satellite is approximately 700 megabytes in size. NASA is planning to launch National Polar-orbiting Operational Environmental Satellite System (NPOESS) project [[5] that in 5 years will generate approximately 1 petabytes of information.

In general EO domain is characterized by:
— large amounts of data acquired from different satellites in different spectral bands that need to be integrated with aerial and in-situ components and maps;
— thematic problems solving require the use of data from multiple sources which in turn leads to the need of use complex data fusion and data mining techniques;
— long-term archives need to be created with uniform access to them.

To enable processing and management of such volumes of data sets and information flows an appropriate infrastructure is needed that will support the following functionality:

— access to distributed resources (data/services/network/computing/storage);
— high flexibility, to foster data fusion and assimilation (meteo, models, global changes, etc.);
— portal enabling easy and homogeneous accessibility;
— virtual organisation (VO) Management;
— collaborative work (e.g. sharing of data sources, tools, means, models, algorithms);
— seamless integration of resources and processes;
— allow processing of large historical archives;
— avoid unauthorised access to/use of resources.

Grid technology is an appropriate solution for solving such kind of problems.

4 The Architecture of Systems for Satellite Data Processing

Based on the existing systems and the systems that are currently being developed, it is possible to identify principal components (sub-systems) and informational flows within a system for satellite data processing (Fig. 1).

Data Storage Sub-System is intended for gathering data from multiple sources, i.e. aerial and space-borne data, in-situ data, etc. Usually, the storage system is organized as multi-layer system each level being characterized by different frequencies of data use. We will consider three-level architecture that consists of an operational archive, a short-term data archive, and a long-time archive. The operational archive contains information that was obtained recently, and there is a higher possibility of accessing this kind of data by users. To store such data hard-discs are usually used enabling minimum access time to data. The short-term archive contains data that were obtained weeks or some months ago. To store these data tape-drives are used. The long-term archive contains data obtained years ago. In some cases such kind of archives can be not automated. They can also use high level of data compression and slow recorders. Time access to these archives can be of hours or days. Such three-level architecture of data storage system is implemented in archives of NASA (USA), DLR (Germany), JAXA (Japan). Two-level architecture is used in the State Research & Productive Center “Pryroda” (Ukraine).

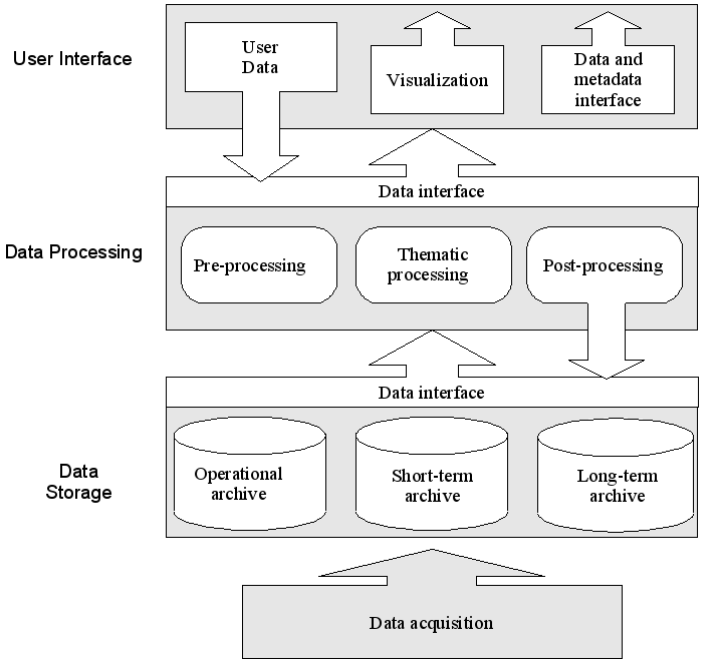


Fig. 1. Three-level architecture of system for satellite data processing

Data Processing Sub-System is intended for data pre-processing (e.g. radiometric and geometric correction of space images, filtering, etc.) and thematic problems solving based on different models and data integration from multiple sources.

User Interface Sub-System is a front-end component that allows end-users to interact with the system. This system is intended for delivering products and services (e.g. raw data and different levels of processed data delivery) to end-users on regular basis or based on their request.

5 Grid Infrastructure for Satellite Data Processing in Space Research Institute of NASU-NSAU

5.1 Grid Infrastructure

A Grid system for satellite data processing that integrates resources of the Space Research Institute of NASU-NSAU, the Institute of Cybernetics of NASU, and the State Research & Productive Center "Pryroda" has been developed. The Grid system consists of two computational SCIT-clusters (the Institute of Cybernetics), a cluster of the Space Research Institute, and an archive of the Meteosat satellite images acquiring from the data center "Pryroda". The developed infrastructure also includes works-stations and network data storage elements. Figure 2 illustrates the overall system architecture.

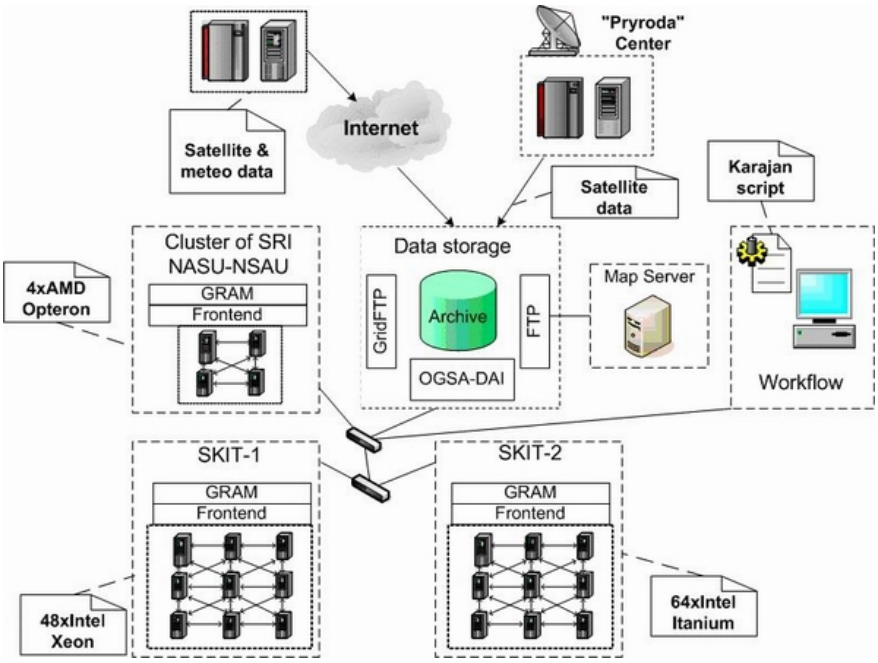


Fig. 2. Current Grid system infrastructure developed in the Space Research Institute of NASU-NSAU

The developed Grid system provides both informational and computational resources of Space Research Institute and Institute of Cybernetics. The computational resources compromise SCIT-1 (48 processors Intel Xeon) and SCIT-2 (64 processors Intel Itanium2) clusters belonging to the Institute of Cybernetics, and the cluster of the Space Research Institute that is used as testing environment. An interface between Grid system and the computational resources is enabled by Grid Resource Allocation and Management (GRAM) service of Globus Toolkit 4. GRAM enables translation of RSL-XML format that is used for job submission request in Globus Toolkit 4 in a format of local job scheduling systems (PBS, Condor, LFS, etc.). Globus provides a set of adapters for standard local job scheduling systems, and tools enabling the development of new adapters. The cluster of thq Space Research Institute uses Torque job scheduling system that is PBS-compatible. In contrast, the SCIT-clusters use its own job scheduling system. That is why, a new adapter was developed in order to integrate these resources in the Grid system.

Up to this moment informational resources consist of archive where data acquired from the “Pryroda” centre and from Internet are stored. In the near future we are planning to provide access to the Meteosat Second Generation (MSG) satellite through DVB technology. The developed archive provides FTP and GridFTP interfaces. Currently, a multi-level data access OGSA-DAI interface is under development which will enable complex distributed requests execution and results combination.

A workflow in the Grid system consisting in job submissions, data transfers, proxy certificates renewal, etc., is controlled by scripts, written in Karajan language [[6]. Karajan is developed as workflow description language for Grid environments and possesses many useful features, such as transparent scheduling and job submission, declarative parallelism and easy extendibility by Java.

5.2 Applied Services

The developed Grid system is currently used to process various satellite data, such as data acquired by the Meteosat satellite and the MODIS instrument onboard the Terra satellite.

Meteosat data in infrared spectrum are used in order to extract a cloud mask using Markov Random Field segmentation algorithm [[7] (Fig. 3). Image processing is done in three steps. First step consists in image filtration (namely, noise detection and removal) that is done using modified version of median filter [[8]. The second step is the segmentation of the image. The third step is post-processing and preparation of the data to be visualized by a map-server. The last step includes geo-reference of raw image and cloud mask, images re-projection, cloud boundary transformation in vector format, metadata creation for visualization. All these algorithms are implemented in the form of Web services available on http://www.dos.ikd.kiev.ua.

MODIS data are used for water quality monitoring in Dnieper river estuary. For this problem solving additional information is required such as in-situ measurements and a number of meteorological parameters, which are acquired using meteorological simulations. For this purpose we use WRF (Weather Research & Forecasting) mesoscale meteorological model [[9]. In order to provide initial and boundary conditions we use data produced by global meteorological model, namely Global Forecast System (GFS), and in-situ measurements. Currently we provide every 6 hours 3-day forecasts for the territory of Ukraine (Fig. 4).

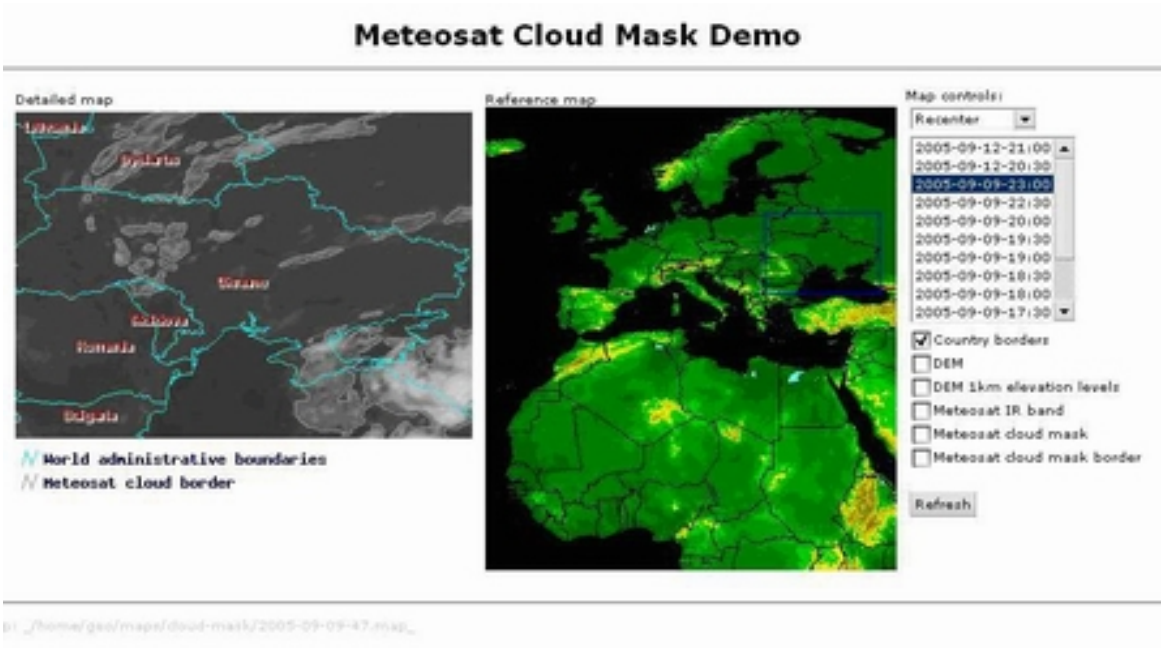


Fig. 3. Visualization of the Meteosat data processing

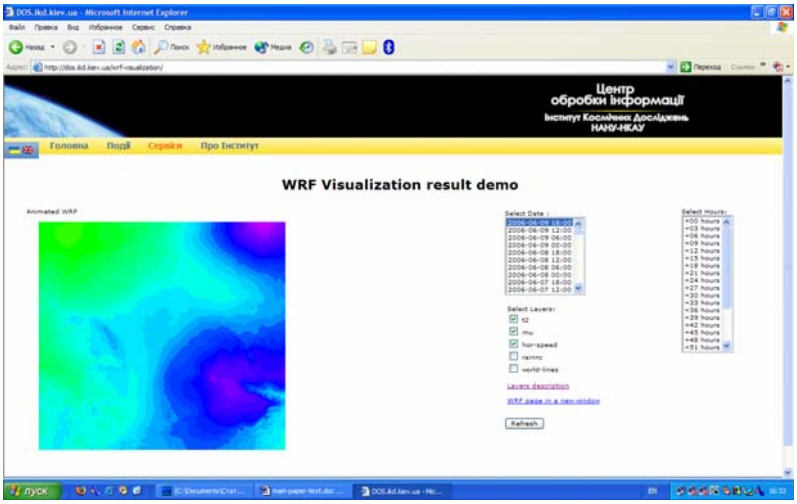


Fig. 4. WRF model visualization

The visualization of resulting data is done with the use of open-source UMN MapServer [[10] software that supports OGC (Open Geospatial Consortium) [[11] standards for spatial data representation.

6 Grid Infrastructure Simulation

Simulation is a common and useful approach for designing complex distributed systems with no exception to Grid. By using models one can decrease TCO (total cost ownership) and save funds on initial installation. Grid systems simulation requires appropriate software usage. The simulation of a Grid testbed in the Space Research Institute was performed by using GridSim [[12] modeling software. Different job scheduling algorithms were analyzed for independent tasks and for data-sharing tasks.

We used GridSim due to its ability to simulate common components of distributed systems such as heterogeneous resources, users, applications and Grid specific components including resource brokers and schedulers for single and multiple administrative domains. Within GridSim package resources can be modeled using time- and space-sharing modes, thus representing workstations, SMP systems and clusters. There are other available Grid simulation packages, such as MicroGrid [[13] and SimGrid [[14]. However, GridSim is more flexible in model design, and does not impose additional requirements, such as Globus Toolkit installation.

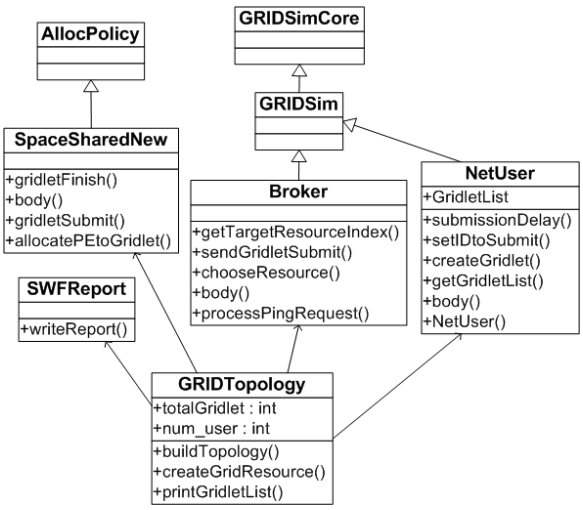


Fig. 5. GridSim Class Diagram

Figure 5 illustrates GridSim class diagram where only redefined methods are depicted. New methods were added in order to extend basic GridSim functionality for our simulations. For example, Broker class was extended for Grid infrastructure resource brokering, GRIDTopology class for Grid infrastructure resource description and presentation.

GridSim model of the developed Grid system was used to estimate different job scheduling algorithms. Two common use cases were examined:

— a large group of independent tasks

— a set of tasks that are using common data

The first use case in comparison corresponds to the ideal parallel algorithm. All branches in algorithm can be executed independently in any order. This is a common situation in Monte-Carlo simulation or pixelwise image processing. The problems of scheduling for independent tasks are well investigated [[15]. However, these investigations stay in the field of homogeneous and static heterogeneous distributed computational systems. In turns, dynamic and heterogeneous Grid environments require some modifications to existing scheduling algorithms to take advantage of full utilization of system resources. The proposed algorithm is based on weighted factoring algorithm [[16]. The proposed modifications lie in using dynamic information about system's state to take into account side load on computational resources. Fig. 6a illustrates the performance of modified algorithm (bold line) comparing with traditional weighted factoring (thin line).

On each iteration of the modified algorithm a set of tasks from the group is assigned to some computational resource. The size of set is estimated as follows:

$$k_i(t) = \alpha \hat{\omega}_i(t) K(t) ,$$

where α is granularity parameter of algorithm, $\hat{\omega}_i(t)$ is the last-known load of computational resource, and $K(t)$ is a number of uncompleted tasks at a given time.

The last-known load $\hat{\omega}_i(t)$ is non-actual by its nature. There are always some lag between present moment and the moment when the information was last updated. The proposed algorithm is quite sensitive to these lags. The performance gain over unmodified version of the algorithm is lost when this parameter grows.

The second use case is a generalization of independent tasks case. The job now consists of tasks that need the same data of considerable size (transfer time of these data is comparable to total task execution time). The data granules are stored on the servers over the network. Each server has some limited bandwidth that separated between different transfers. The developed algorithm introduces fit measure U that shows a quality of assignment of some task to specific resource:

$$U=F+Q.$$

In this expression F is the measure of unbalance and shows how balanced is the use of system resources (computational and network channels) by particular task, and Q is the measure of system resources utilization.

Fig 6b illustrates the performance of developed algorithm (bold line) comparing with random and round-robin schedulers (thin lines).

7 Conclusions

Nowadays, there is a strong interest of scientific communities from different domains in the development of distributed systems for complex problems solving with the use of high-performance computing. Grid represents an appropriate technology that enables integration and management of geographically distributed informational and computational resources. In the last years leading organizations of the NASU are involved in the research and development of Grid-based computing systems, and the first results have been already achieved. In the near future it is planned to integrate Ukrainian resources in a single infrastructure based on the high-speed network. And this infrastructure should be based on recent developments in Grid technology, high-speed networks, and multi-processors platforms.

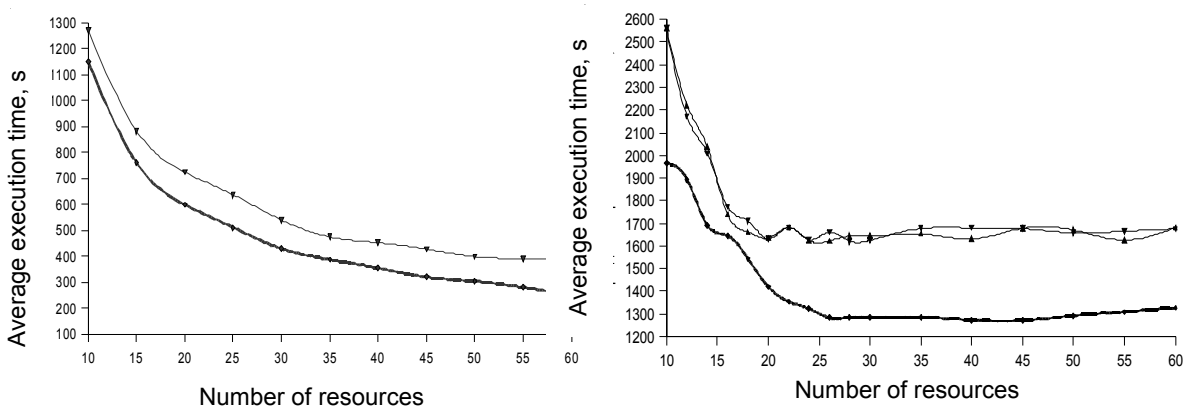


Fig. 6. Average task execution time for independent tasks
(a) and data-sharing tasks (b) depending on number of resources in Grid system

The Grid infrastructure for satellite data processing that has been developed in the Space Research Institute of NASU-NSAU will become Ukrainian segment of the GEOSS/GMES system.

Acknowledgments. The work is partly supported by the STCU and NASU Targeted Initiatives Program, project "GRID technologies for environmental monitoring using satellite data" (No. 3872); NASU grant for Young Scientists "Development of Desktop Grid system and optimization of its productivity"; and NASU Innovative Project "Development of mathematical models, methods and information technologies of satellite data processing in the framework of test-bed information system of Ukrainian segment of GEOSS/GMES".

References

[1] Foster, I., Kesselman, C., Tuecke, S.: The Anatomy of the Grid: Enabling Scalable Virtual Organizations. Int. J. of High Performance Computing Applications 15 (3) (2001) 200-222.

[2] Gentzsch, W.: Special issue on metacomputing: From workstation clusters to internet computing. Future Generation Computer Systems 15 (1999).

[3] Japan Aerospace eXploration Agency (JAXA), http://www.jaxa.jp/index_e.html.

[4] Information Power Grid (IPG), http://www.ipg.nasa.gov.

[5] National Polar-orbiting Operational Environmental Satellite System (NPOESS), http://www.ipo.noaa.gov.

[6] Karajan workflow description language, http://wiki.cogkit.org/index.php/Java_CoG_Kit_Karajan_Workflow_Reference_Manual.

[7] Kussul, N., Shelestov, A., Phuong, N., Korbakov, M., Kravchenko, A. : Parallel Markovian Approach to the Problem of Cloud Mask Extraction. In: Proc. of XI-th International Conference "Knowledge-Dialog-Solution", Varna, Bulgaria. (2005) pp. 567-569.

[8] Nguyen, T.P.: Concurrent Algorithm For Filtering Impulse Noise On Satellite Images. In: Proc. Int. Conference «Knowledge–Dialogue–Solution» (KDS–2005) (2005) 465–472.

[9] Weather Research & Forecasting model, http://wrf-model.org.

[10] University of Minnesota MapServer., http://mapserver.gis.umn.edu.

[11] OGC Standards, http://www.opengeospatial.org/specs/?page=specs.

[12] Buyya, R., Murshed, M.: GridSim: A Toolkit for the Modeling and Simulation of Distributed Resource Management and Scheduling for Grid Computing, Concurrency and Computation: Practice and Experience (CCPE), Volume 14 Issue 13-15 (2002) 1175-1220. Wiley Press, USA.

[13] Song, H., Liu, X., Jakobsen, D., Bhagwan, R., Zhang, X., Taura, K., Chien, A.: The MicroGrid: A Scientific Tool for Modeling Computational Grids, Proc. of IEEE Supercomputing (SC 2000), Dallas, USA, (2000).

[14] Casanova, H.: Simgrid: A Toolkit for the Simulation of Application Scheduling, Proc. of the First IEEE/ACM International Symposium on Cluster Computing and the Grid (CCGrid 2001), Brisbane, Australia, IEEE Computer Society Press, USA (2001).

[15] Baumgartner, K., Wah, B.W.: Computer Scheduling Algorithms: Past, Present and Future, Information Sciences, vol. 57 & 58, (1991) 319-345. Elsevier Science, Pub. Co., Inc., New York, NY.

[16] Flynn Hummel, S., Schmidt, J., Uma, R.N., Wein, J.: Load-Sharing in Heterogeneous Systems via Weighted Factoring. In: Proc. of the 8th Symposium on Parallel Algorithms and Architectures (1997).

Author's Information

Nataliia Kussul *– Professor, Senior Researcher, e-mail: inform@ikd.kiev.ua.*

Andrii Shelestov *– PhD, Senior Researcher, e-mail: inform@ikd.kiev.ua.*

Mykhailo Korbakov *– Research Assistant, e-mail: inform@ikd.kiev.ua.*

Oleksii Kravchenko *– Research Assistant, e-mail: inform@ikd.kiev.ua.*

Serhiy Skakun *- PhD, Research Assistant, e-mail: inform@ikd.kiev.ua.*

Mykola Ilin *– e-mail: inform@ikd.kiev.ua.*

Alina Rudakova *– Research Assistant, e-mail: inform@ikd.kiev.ua.*

Volodymyr Pasechnik *– e-mail: inform@ikd.kiev.ua.*

Department of Space Information Technologies and Systems, Space Research Institute of NASU-NSAU, Glushkov Ave 40, Kyiv-187, 03650 Ukraine

XML AND GRID-BASED APPROACH FOR METADATA EXTRACTION AND GEOSPATIAL DATA PROCESSING

Andrii Shelestov, Mykhailo Korbakov, Mykhaylo Zynovyev

Abstract*: The software architecture and development consideration for open metadata extraction and processing framework are outlined. Special attention is paid to the aspects of reliability and fault tolerance. Grid infrastructure is shown as useful backend for general-purpose task.*

Keywords*: Metadata, ISO19115, Grid computing, Globus Toolkit, XML.*

ACM Classification Keywords*: D.1.3 Concurrent Programming -- Distributed programming, D.2.3 Coding Tools and Techniques – Object-oriented programming, D.2.5 Testing and Debugging – Error handling and recovery, D.2.0 General – Standards.*

Introduction

Metadata extraction, indexing and querying is important task from very different points of view. Consistent and actual metadata database enables effective use of data archives for users and create capabilities to develop new high-level services taking advantage of task run time prediction or automatic composition of semantic workflows. The current activities in metadata processing tools development are mainly targeting desktop indexing and search tools (Beagle, Tracker, Strigi, libferris [Martin, 2005]). However there are no available metadata processing systems that can handle geospatial data with their complex file formats, diverse metadata structure and complex queries. There is an ISO standard for geospatial metadata (described in UML) [ISO19115, 2003] as

well as XML representation for it [ISO19139, 2007], but none of the available systems known to the authors take advantage of it. The approach described in this paper is targeting to fill this gap.

The work was supported by the INTAS-CNES-NSAU grant #06-1000024-9154 “Data Fusion Grid Infrastructure” and STCU-NASU grant #3872 “GRID technologies for environmental monitoring using satellite data”.

Objectives

The problem of extraction, storing and querying of metadata of geospatial data is very important in the context of development of distributed geoinformational systems of national or even larger scale. The large scale systems for environmental monitoring such as international GEOSS [GEOSS, 2005] or European GMES [GMES, 2005] operate on very large sets of data and need consistent and actual catalog of metadata to operate efficiently.

The development of system being described in this paper is carried within the number of national Ukrainian initiatives such as CosmoGIS and GEOUA and international grants INTAS-CNES-NSAU #06-1000024-9154 “Data Fusion Grid Infrastructure” and STCU-NASU #3872 “GRID technologies for environmental monitoring using satellite data”. These projects and initiatives are targeting on development and exploitation of Grid infrastructure in the tasks of geospatial data processing and environmental monitoring. The problems being solved in these systems involve satellite imaginary obtained via EUMETCast data dissemination system [EUMETCast, 2006], weather forecasts data and hydrological modeling. With growth of functionality of systems in development the need for metadata catalogue and queries processing engine became clean.

The Grid technology was found very suitable for development of such system being the basis for many data processing infrastructures and stated as a technological foundation for implementation of GMES system.

After analysis of requirement authors have identified a number of functions and features that must be supported by such system.

Functions:

1. **To extract metadata from files of different formats.** Scientific data often comes in quite unusual formats that require special software that can not be found on most of computers. Examples include XRIT envelopes that are used for distribution of MSG satellite or HRPT format that is used for NOAA satellite data. Data that are using standard container formats like HDF of NetCDF are using different field structures. These factors cause the need to utilize external handlers to process different kinds of data.
2. **To store the metadata in the queryable form in centralized storage.** The common problem of storing abstract metadata in common relational database is large diversity of data structures that is causing denormalization of relations. The proposed XML-based solution of this problem is described in the following sections.
3. **To provide interface to perform user queries.** The user should be able to query metadata with different match criteria or their combinations. The proposed query language is W3C recommended XQuery language [XQuery, 2007].
4. **To support geospatial queries.** This requirement originates from geospatial nature of satellite imaginary. Minimal set of geospatial queries functions is support for windows queries, spatial join by intersection and different coordinate reference systems.

Features:

1. **Distributed agent-server architecture.** The data that require processing can be collected on different sites of data retrieval, processing and storing facilities. The metadata extraction module should operate closely to the data to reduce unnecessary data transfers. This leads us to separation of metadata extraction part that should be deployed on the storage side from the metadata indexing and processing part that should be deployed on the separate server.

2. **Support for user extensions.** The extraction of metadata should be performed by the means of external handlers that must process specific file formats and communicate with main extractor process via defined interface. The communication interface must be kept as simple as possible to avoid unnecessary limitations on implementation of external handlers.
3. **Continuous processing.** The metadata extractor storage-side agent must detect and process new data files at the moment when they arrive on the storage.
4. **Fault tolerance.** The metadata extractor must properly handle different kinds of OS-level errors like file access errors or network unavailability and also isolate errors in external file formats handlers including incorrect resource deallocation.
5. **Access rights enforcement.** Distributed system requires features of authentication and authorization to prevent abuse of its services by unauthorized users.

The proposed solution for implementation of these functions and attributes is given in the following sections.

Analysis of XML databases usefulness

The XML databases appeared as a natural consequence of rapid growth of XML as the standard for information exchange and increasing volumes of XML documents. As opposed to relational database in their current state, XML databases have very different functionality, performance and query processing capabilities thus making the right choice difficult.

XML database differs from a relational database in a number of directions.

- Relational database uses a row (relation) as the basic storage unit, while an XML database uses an XML document.
- Instead of using SQL for querying and updating data XML databases use the pair of XQuery and XUpdate languages.
- Relational databases are generally inefficient if the entire document is needed as it may be split across tables. XML databases may be inefficient if document or a part is requested in a form different from which it is stored. In this respect XML databases are similar to hierarchical databases.
- Relational databases suits best to the situations with simple enough data structures that can be described in the terms of relations without denormalization. XML databases can be used in situation with complex data structures that can not be easily mapped to relations.

Among many different XML databases we can identify two large classes [Srivastava, 2004]:

- **XML-enabled Relational Databases.** These databases are natural evolution of traditional relational databases with new features that allows developers to combine SQL and XQuery queries. The most noticeable examples of this class are Oracle and MS SQL Server databases. The PostgreSQL database will implement XML features in close future.
- **Native XML Databases.** These databases were designed to store XML from the scratch. The researches are very active in this area so the databases differ greatly in sense of architecture, functionality, language support and other. There are much more representatives of this class comparing with XML-enabled relational databases. Most notable are Berkeley DB XML, eXist, Sedna, X-Hive, Xindice.

XML databases are naturally suited to storing geospatial metadata information for the reasons of already existing XML mapping of ISO19115 metadata standard and complexity of metadata structure. However the requirement for ability to process geospatial queries puts very strong restriction on the choices possible.

The only database engine available at the present moment with support for both XQuery and geospatial functions is Oracle 10g. The development of metadata indexing service uses Oracle 10g under the development license. PostgreSQL developers team claims to add XQuery features in the next releases. In this case PostgreSQL will

become an engine of choice for its long-available geospatial features implemented in PostGIS extension (http://postgis.refractions.net/).

Proposed architecture

The system being described consists of two parts:

- **The agent part**, that deployed on the side of data storage and performs continuous monitoring of new and changing data files. The results of monitoring are pushed to the server part. The term "agent" was chosen with respect to SNMP (Simple Network Management Protocol) agents and backup agents terminology. The agent is a special kind of server's client that only feeds data to it without or with very little of other kinds of interaction.
- **The server part**, that deployed on the side of metadata accepts metadata from multiple agents deployed on different hosts, stores retrieved metadata in persistent storage and handles user queries.

These parts perform very different functions and thus have different implementations. The server part is implemented as a service for Globus Toolkit 4 [Foster, 2005] container. Use of Grid approach for development of such service allows us to achieve transparent integration with other applied Grid services being developed in the scope of other projects and take advantage of consistent Grid middleware services such as authentication and authorization mechanisms, monitoring and event publication.

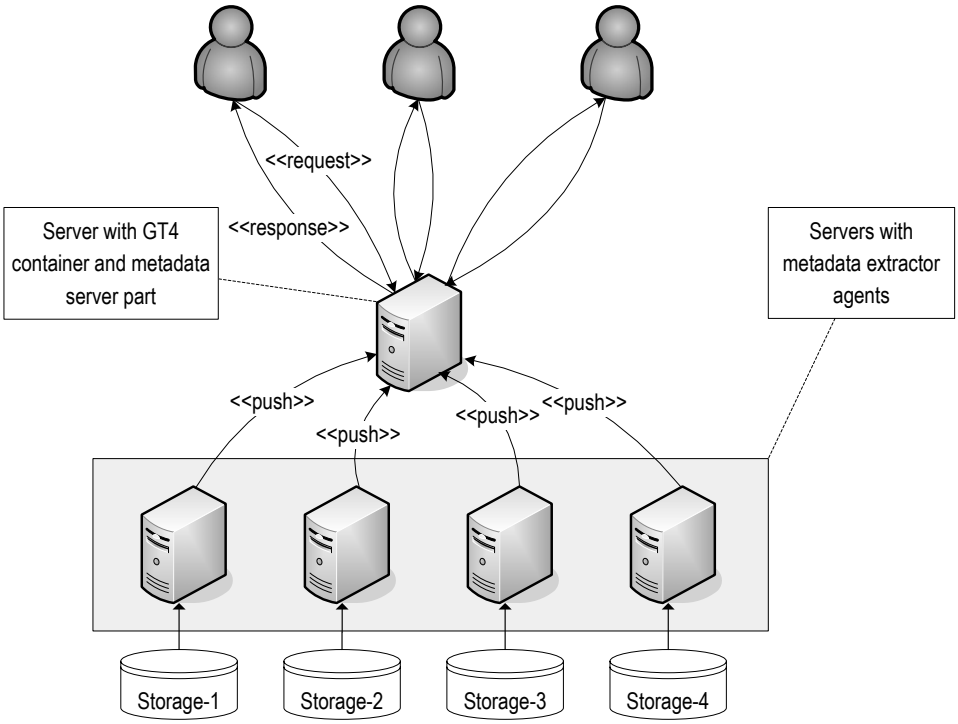


Figure 1. High-level overview of the architecture

The following features of GT4 are used:

1. **Security framework.** GT4 allows service to use PKI (Public Key Infrastructure) with loosely coupled configuration of security policy down to the granularity of specific service methods. The security framework covers "Three A" problems: Authentication, Authorization, Audit in the standard way as well as providing means for channel encryption and integration with existing security infrastructures.
2. **Index service.** This standard GT4 service allows other services to publish information about their state and to query other services information. The server part of metadata processing system using Index service to publish information about last updates and overall availability.

3. **Event publishing.** GT4 allows services to publish events and subscribe to publication in the way, similar to Observer design pattern [Gamma, 1995]. This allows other services to monitor availability and updates of metadata database and to react to specific events. Event publishing reduces the load on application server by elimination the need for periodic checks.

The Figure 1 shows the high level overview of the architecture with several metadata extractor agents, metadata server, deployed into GT4 container and some users performing requests on it.

The server receives metadata information from remote agents and puts it into persistent storage implemented as XML database. The use of XML database to store ISO19139 data granules and other metadata in form of XML documents has a number of advantages comparing with traditional relational databases systems. User requests are sent to the server in form of XQuery language. It should be mentioned that XQuery isn't very user-friendly and simple language so the optimal solution for user interface is high-level client application (either host-based or web-based) that will interpret user request in terms of ISO19115 and put it into XQuery.

The metadata extractor agent is implemented as a Python application for Linux environment and a set of extensions to handle specific file formats. The program uses FAM interface (File Alternation Monitor, http://oss.sgi.com/projects/fam/) to continuously monitor a number of directories specified by user, detects file format and then applies an appropriate handler.

To support clean shutdowns metadata extractor should support persistent storage for processing states of data files. At the present moment persistent storage is implemented by the means of the extended file system attributes that allows associating each file with a set of name-value pairs. Using of embedded database like SQLite or Embedded MySQL will improve compatibility of the software and will be implemented soon.

One of the design goals of the metadata extractor part was to avoid unnecessary limitations on implementation of external handlers. One of the possible ways to achieve it was to keep communication interface between the main program and extension as simple as possible. The current implementation allows every executable file put to special directory to be treated as an extension provided that it takes one command line parameter – the name of the file to process, returns valid XML document at the standard output and keeps the common agreements for the return code of a process. Such approach allows developer to choose most appropriate tool to develop a needed extension and shows zero learning curve.

The Figure 2 shows the class diagram of metadata extractor. Four classes implement the core functionality:

1. **Extractor** class is top of hierarchy and is responsible for startup of the program and construction of instances of other classes. The instance of this class is collecting information provided by DirectoryWatcher instances and sends it to server part.

2. **DirectoryWatcher** class is responsible for monitoring of a single directory in filesystem. The instance of this class starts FAM monitor and processes new files using FileHandler objects.

3. **FileHandler** class is encapsulation of external handler of file format. It responsible for running external process, retrieving the output and wrapping errors into exceptions.

4. **EAFileJournal** class is implementation of persistent storage for processing states of data files. It's implemented using extended attributes of file system.

Two composition relations marked as IPCObject represent wrapper objects that serve for interprocess communications implemented by the means of D-Bus protocol (http://www.freedesktop.org/wiki/Software/dbus) [Palmieri, 2005]. The details and reasoning of this design solution will be given further in "Reliability and Fault Tolerance considerations" section.

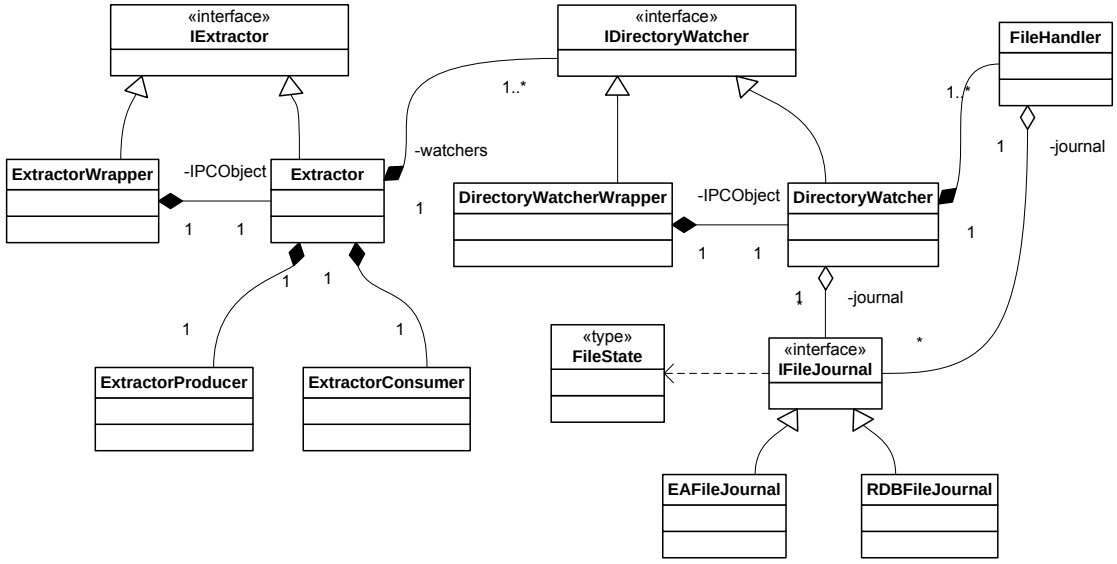


Figure 2. Class diagram of storage-side part of metadata extractor

Reliability and Fault Tolerance considerations

Reliability and fault tolerance were specifically stated in the objectives of the system. Analysis for potential faults has shown the number of sources (listed without any specific order).

- Disk read errors (SAN connectivity lost comes here too).
- Lost of network connectivity with indexing server.
- Erroneous input data.
- Exceeding of system resources (system memory is most probable candidate).
- Incorrect deallocation of resources leading to leaks.

Realization of potential fault will lead to the need of human interaction if the critical situation isn't handled in a proper way. To reduce the potential harm from these faults a number of special measures have been taken.

Disk read errors that can come either from physical hard drive fault or from lost of connectivity in storage area network do no require any special handling except proper detection of faulty situation and further periodical checks for disk availability. The same applies to the lost of network connectivity with indexing server.

Erroneous input data that is causing external file handler to crash with error should be marked as broken using persistent file state storage (FileJournal in terms of Figure 2). In those cases when input data consist of several files all of them should be declared as broken.

The exceeding of system resources and incorrect deallocation of resources is interdependent phenomena. The roots of the problem lies in the fact that the metadata extractor should run external code to retrieve metadata from different file formats. The following mechanisms were implemented to reduce the risk of crash for metadata extractor:

1. **Isolation of error.** Instead of running all of the tasks in a single process that can crash different tasks are running in separate processes. This applies both for external data format handlers and for different directories monitored by DirectoryWatcher instances (see Figure 2) running in separate process. The communication between external handlers and DirectoryWatcher instance is handled by standard output stream redirection that is very reliable. The communication between DirectoryWatcher instances and Extractor instance is performed by D-Bus interprocess message exchange system. In case of crash of process of external handler or DirectoryWatcher instance the operating system effectively frees all the

system resources that crashed process have allocated. This doesn't apply to outer resources such as database connections but the use of such resource in described scenario quite unlikely.

2. **Watchdog timer.** The parent process of Extractor instance periodically checks the availability of underlying DirectoryWatcher processes via D-Bus message exchange. In case when DirectoryWatcher process doesn't reply in the specified time the process is killed and restarted by Extractor instance.

The last mechanism for supporting uninterruptible work of metadata extractor is logging system that records all relevant activities of it and can use quick contact methods such as email, IM or SMS to notify responsible person of all of system errors.

Conclusion

The authors have described the architecture of system for metadata extraction, indexing and processing targeting the geospatial data based on Globus Toolkit and XML databases technologies. The use of Globus Toolkit for such system shows a shift in the understanding of Grid systems. Authors believes that at the present moment Grid systems should be seen not only as a mechanism for supporting large computations and data transfers but also as a useful platform for value-added tasks.

Bibliography

[Martin, 2005] Filesystem Indexing with libferris. B. Martin. Linux Journal, 2005

[ISO19115, 2003] ISO standard ISO 19115:2003. Geographic information – Metadata

[ISO19139, 2007] Draft of ISO standard ISO 19139. Geographic information -- Metadata -- XML schema implementation

[GEOSS, 2005] Global Earth Observation System of Systems GEOSS. 10-Year Implementation Plan Reference Document. Noordwijk, Netherlands: ESA Publication Division. 2005. 212 p

[GMES, 2005] GMES: From Concept to Reality. European Commission. 2005

[EUMETCast, 2006] TD 15 – EUMETCast – EUMETSAT's Broadcast System for Environmental Data. Technical Description, (http://www.eumetsat.int/idcplg?IdcService=GET_FILE&dDocName=pdf_td15_eumetcast&RevisionSelectionMethod=LatestReleased), 2006

[Xquery, 2007] XQuery 1.0: An XML Query Language, W3C Recommendation 23 January 2007 (http://www.w3.org/TR/xquery/)

[Foster, 2005] Globus Toolkit Version 4: Software for Service-Oriented Systems. I. Foster. IFIP International Conference on Network and Parallel Computing, Springer-Verlag LNCS 3779, pp 2-13, 2005.

[Welch, 2003] Security for Grid Services. V. Welch, F. Siebenlist, I. Foster, J. Bresnahan, K. Czajkowski, J. Gawor, C. Kesselman, S. Meder, L. Pearlman, S. Tuecke. Twelfth International Symposium on High Performance Distributed Computing (HPDC-12), IEEE Press, 2003.

[Gamma, 1995] Design Patterns: Elements of Reusable Object-Oriented Software. Erich Gamma, Richard Helm, Ralph Johnson, and John Vlissides. Addison-Wesley Professional, 1995.

[Palmieri, 2005] Get on D-BUS. John Palmieri. Red Hat Magazine, issue #3, January 2005.

[Srivastava, 2004] Comparison and Benchmarking of Native XML Databases. CS497 Report. Anand Vivek Srivastava. Department of Computer Science and Engineering, Indian Institute of Technology, 2004.

Authors' Information

Andriy Shelestov *– Space Research Institute of NASU-NSAU, senior scientist; Ukraine, Kiev-03680, prosp. Glushkova-40; e-mail: inform@ikd.kiev.ua*

Mykhailo Korbakov – *Space Research Institute of NASU-NSAU, junior scientist; Ukraine, Kiev-03680, prosp. Glushkova-40; e-mail: rmihael@gmail.com*

Mykhaylo Zynovyev – *Bogolyubov Institute for Theoretical Physics of NASU, engineer, Ukraine, Kiev-03680, Metrolohichna str. 14-B; e-mail: Mykhaylo.Zynovyev@cern.ch*

GEOSPATIAL DATA VISUALIZATION IN GRID SYSTEM OF UKRAINIAN SEGMENT GEOSS/GMES

Andrii Shelestov, Olexy Kravchenko, Mykola Ilin

Abstract: *Implementation of GEOSS/GMES initiative requires creation and integration of service providers, most of which provide geospatial data output from Grid system to interactive user. In this paper approaches of DOS-centers (service providers) integration used in Ukrainian segment of GEOSS/GMES will be considered and template solutions for geospatial data visualization subsystems will be suggested. Developed patterns are implemented in DOS center of Space Research Institute of National Academy of Science of Ukraine and National Space Agency of Ukraine (NASU-NSAU).*

Keywords: *data visualization.*

ACM Classification Keywords: *I.3.2 Graphics Systems - Distributed/network graphics, C.5.0 Computer system implementation – General.*

1 Introduction

Grid systems providing geospatial data are common and usually have complex visualization subsystems. Wide class of typical problems are weather prediction, satellite data processing can be solved in these systems, some of them are solved in DOS center of Space Research Institute of National Academy of NASU-NSAU. Different interfaces and architecture assumptions can make these Grid systems very hard for development and usage, lowering their value as the data source for decision making. Implementation of standards for data visualization, creation of common template solutions will simplify development and increase usability of these systems.

These approaches are used to implement distributed geospatial data visualization subsystem of national Ukrainian Earth Observation system which is developed in the frame of international program GEOSS and European program GMES.

International GEOSS program (Global Earth Observation System of Systems) is emerged to integrate national and regional Earth Observation systems [1]. One of such systems is developed within European GMES (Global Monitoring for Environment and Security) initiative. This initiative is supported by European Commission and European Space Agency and targeting on providing information services for decision making [2].

The overall structure of Ukrainian segment of GEOSS/GMES has three organizational levels. The top level is responsible for the overall management of the system, the second is responsible for integration of efforts in particular sectors of economy. At the lowest level of system's hierarchy DOS (Delivery of Service) centers are located. These centers are responsible for delivering particular services to end users [3].

To represent activities at all levels of Ukrainian segment of GEOSS/GMES the following hierarchy of Web-resources is created (Fig. 1):

- Main portal [4]
- Sectoral portals (http://cosmogis.org.ua, http://spaceweather.org.ua, …)
- Web-resource of DOS-centers

The most developed sectoral system of Ukrainian segment of GEOSS/GMES is CosmoGIS system supported by NSAU [5]. CosmoGIS is created to stimulate cooperation in the field of remote-sensing data processing and to provide end users with new quality thematic products. Environmental monitoring using remote-sensing involves execution of complex workflows of data processing and often requires computationally intensive ecological

simulations. For such applications Grid computing is desirable. Typical Web-site of DOS-center presents an interface to target Grid system and provides facilities to visualize and distribute results of processing.

Contrary to top level of Ukrainian segment of GEOSS/GMES sectoral level involves substantial interactions between components (DOS-centers). One of the goal of CosmoGIS as sectoral system consists in integration of DOS-centers, in particular providing a means for distributed data visualization and delivery. To attain this goal CosmoGIS uses open standards of geospatial data presentation. At present the most advanced standards in this area both in capabilities and available software are standards of Open Geospatial Consortium (OGC). The utilizing of OGC standards ensures possibility of integration with similar systems at national level, in particular within GMES program.

In this paper approaches of DOS-centers integration will be considered and template solutions for geospatial data visualization subsystems will be suggested. Developed patterns are used to implement DOS-center of Space Research Institute of NASU-NSAU.

2 Approaches to organization of visualization systems

One of the main obstacles on the creation of distributed systems in Ukraine is a not sufficient high throughput networks and nonuniform distribution. To account the insufficiency of high throughput networks centralized and decentralized approaches to create of distributed visualization system for geospatial data are considered [7]. The differences between these approaches consist in different traffic routing schemes between end user and DOS-centers and places where mapping products are created. Both schemes assumes that DOS-centers are using OGC Web Feature Service (WFS) [8] standard to distribute vector geospatial data and OGC Web Coverage Service (WCS) [9] to distribute raster data.

The typical structure of centralized version of the system is shown on Fig. 2a. The figure shown how the thin client accesses the portal with the directory of available services and makes a request to the specified service. This request routed to the mapping service, packed into WFS/WCS request and send to the service site (DOS).

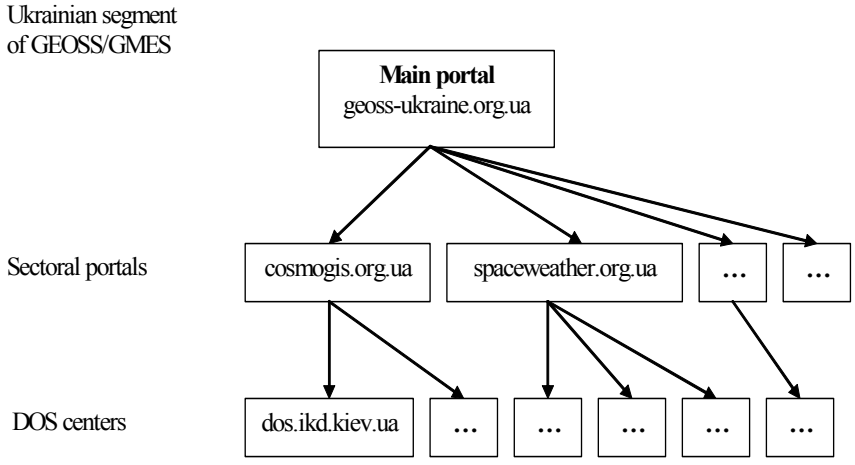


Fig. 1. Hierarchical structure of Ukrainian segment of GEOSS/GMES

The result is routed back, processed by to the cartographical service and send to client. In this case the centralized mapping service is responsible for producing cartographical output.

This first approach exhibits ability to use thin clients (and as result to serve broader range of end users), to produce high quality mapping output independent of end clients capabilities. As a drawback this scheme has potential bottlenecks in network throughput and computational power of central mapping server.

Within decentralized scheme each DOS-center uses own mapping service (Fig. 2b). Cartographical output of service center is delivered using OGS Web Map Service (WMS) [10] protocol and client software is responsible for combining created maps. Central portal only holds references to DOS-centers and routes user requests to DOS mapping services. The second approach relaxes requirements on network throughput and available computational power at the cost of using more sophisticated clients.

In both schemes dedicated software such as geoinformational systems (GIS) can call DOS directly using WCF/WFS protocols.

3 Template solutions for visualization

In this section two template solutions for visualization subsystems will be described. One solution is based on the thin client model while another utilize thick clients. Templates can be used to develop data visualization subsystems for DOS-centers and for portals at sectoral level. Both template solutions have different advantages and drawbacks. Visualization template using thin client model has a low scalability, minimalistic user interface (without navigation, dynamical scalability, etc), but does not require expensive hardware for both client and visualization server. Thick client

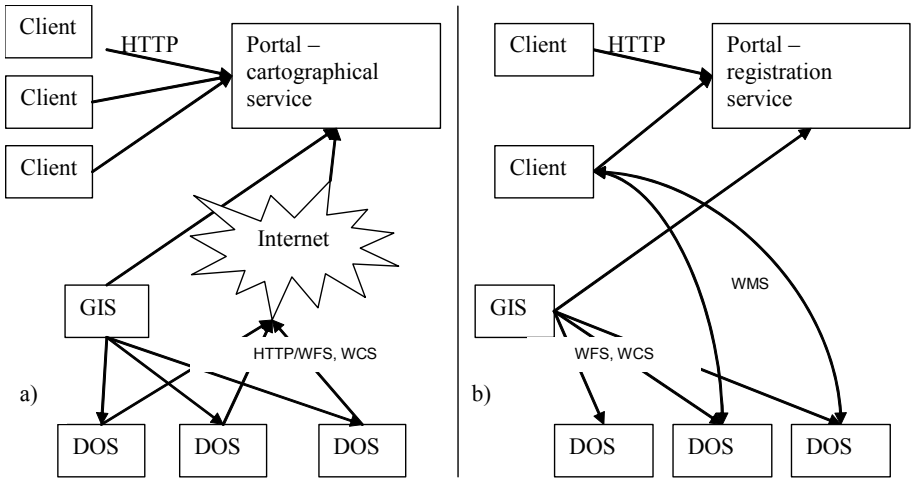


Fig. 2. (a) centralised and (b) decentralized approaches to distributed visualization of geospatial data

model visualization template has better scalability and usability, but requires expensive server or group of servers for mapping service.

3.1 Visualization systems using thin client model

First pattern of visualization system is based on the thin client model. To access the service a simple web browser is sufficient. Within this pattern, visualization system implements open standards of data presentation including OGC WMS to deliver cartographic products and OGS WFS/WCS to deliver geospatial data. Typical structure of such pattern is shown on Fig. 3

The vector and raster data produced by Grid system is visualized by mapping service. Both mapping service and Web-interface are implemented in the framework of open source software UMN Mapserver [11]. Mapping service and visualization system located on single server, this server has sufficient performance for visualization of only a few layers on the target map. Performance restrictions are critical for visualization tasks, not all available Grid systems can use this pattern because of these restrictions. On the other side simple client software, cheaper hardware for both client and server makes this pattern very suitable in Ukrainian segment of GEOSS/GMES.

The main advantage of described pattern is the standard interface of mapping service, which grants compatibility with existing and new client applications. Once developed, client applications and DOS-centers can easily switch data sources with minor or no modifications of source code. This improves scalability and availability of entire

segment. GIS can use visualization server as a client, visualization server implements standard protocols for data representation.

A typical example of thin client pattern implementation is cloud mask visualization service which is described in section 4.1.

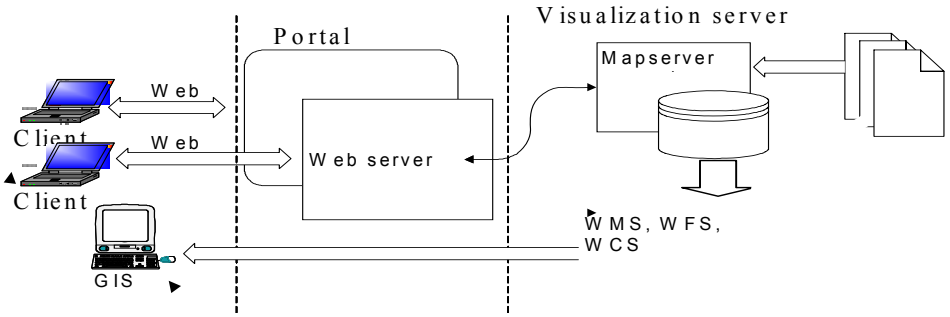


Fig. 3. Visualization system using thin client

3.2 Visualization systems using thick client model

The second pattern of visualization system is based on thick client model. To access a service a Web-browser with JavaScript is required. Key feature of proposed solution is extensibility, new versions of framework software requires more system resources. This pattern uses previous as base and adds advanced navigation capabilities for interactive users.

Another advantage of this architecture is the possibility of parallel processing of user requests allowing integration of different data sources. This feature increases the system scalability while the system remains transparent for target users. More sophisticated applications can be developed, because of performance increase, routing capabilities and load balancing. Within this template, the solution visualization subsystem is developed using open source software Cartoweb 3.2.0 [12]. In this system SOAP is used for interserver communication (among different visualization systems), allowing integration with virtually any DOS-center, even without WCS/WFS support. CartoWeb can be extended using plugin approach that makes interface modifications simple.

The main features/advantages of common Cartoweb-based interface are visible on main map control – it has scale and position arrows that can be used for scale the adjustment and navigation, this feature being commonly used with dynamic (i.e. clickable) keymap. Other commonly used features are the measuring tools for distances and surfaces measurement. The control panel has layers tree – CartoWeb supports an arbitrarily complex hierarchy of layers, with infinite depth. The interface contains a geographic query tool which can be used for geographical search. Additional features are language switch for internationalization support, users and roles support for an implementation of basic (file-based) authentication mechanism, print dialog for fully configurable PDF document production.

Typical structure of such system is shown on Fig. 4. All capabilities of previous pattern are preserved, to use new system little or no modification in existing application is required.

4 Implementation

The developed patterns for visualization subsystem were used to implement the DOS-center of Space Research Institute's of NASU-NSAU [13]. This center was created under the umbrella of CosmoGIS sectoral system of Ukrainian segment of GEOSS/GMES.

The thin client-based pattern was used to implement cloud mask visualization service which is described in section 4.1, the thick client-based approach is demonstrated on example of visualization in Numerical Weather Prediction (NWP) described in section 4.2.

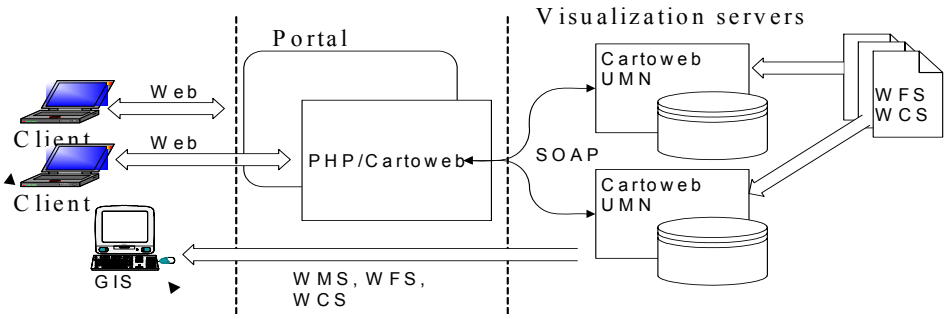


Fig. 4. Visualization system using thick client

4.1 Service of cloud mask extraction from Meteosat remote sensing data

The cloud mask visualization is a typical application of the thin client pattern. The main source of data for this service is provided by European Meteosat meteorological satellite. The cloud mask extracted from Meteosat infrared band using Markov Random Fields (MRF) approach. Cloud mask extraction is being executed on the top of Grid system developed in Space Research Institute [14, 15].

There are few layers visualized on single machine with single service, single server has sufficient performance for smooth presentation with reasonable delays. Typical service interface is shown on Fig. 5. On the left map of this figure cloud mask is shown with country names and boundaries. Middle reference map is used for navigation. Right control panel is used for map navigation, selection of the date and additional features to be included in resulting map. Available data layers include country boundaries, Digital Elevation Model data, Meteosat infrared remote sensing data, cloud mask and clouds borders.

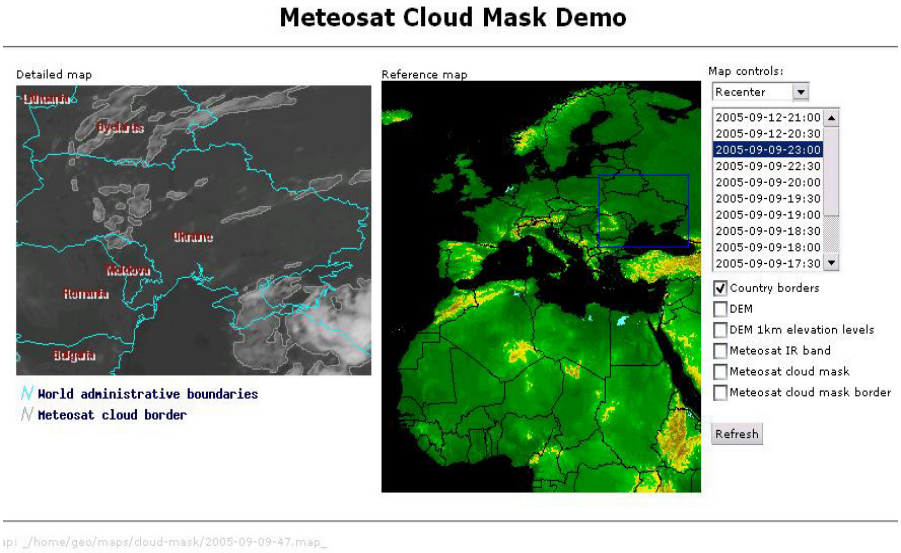


Fig. 5. Thin client system for cloud mask visualization

4.2 NWP model visualization

The visualization of NWP results is a good example of implementation of thick client-based pattern. This pattern was used to visualize results of WRF mesoscale model simulations which is regularly preformed in Space Research Institute. NWP models predicts a lot of meteorological parameters. Due to the fact that these many visualization layers have to be calculated on different servers, layer options are too complex for single mapping service.

A typical user-friendly output of visualization system is shown in the Fig. 6. The main visualization controls and options located in right panel of the figure. Currently visible tab folder shows background and geopolitical reference options. Other tabs of this panel includes options to access measurement tools, which can be used for area and length calculation, print dialog for PDF exporting support, query tools for visual MapServer query creation, and outlining tools for map annotation. On the right part of the figure WRF model temperature output is shown. In the upper left corner of this map reference keymap with relief data is shown. Visualization map has advanced zooming and navigation capabilities. Zooming can be used not only with scale switch in bottom right corner, but can be applied for user-selected region. Different line and polygon drawing tools in upper panel can also be used for map annotation.

This example refers to a thick client model because navigation interface is implemented using JavaScript, uses AJAX technology for map operations.

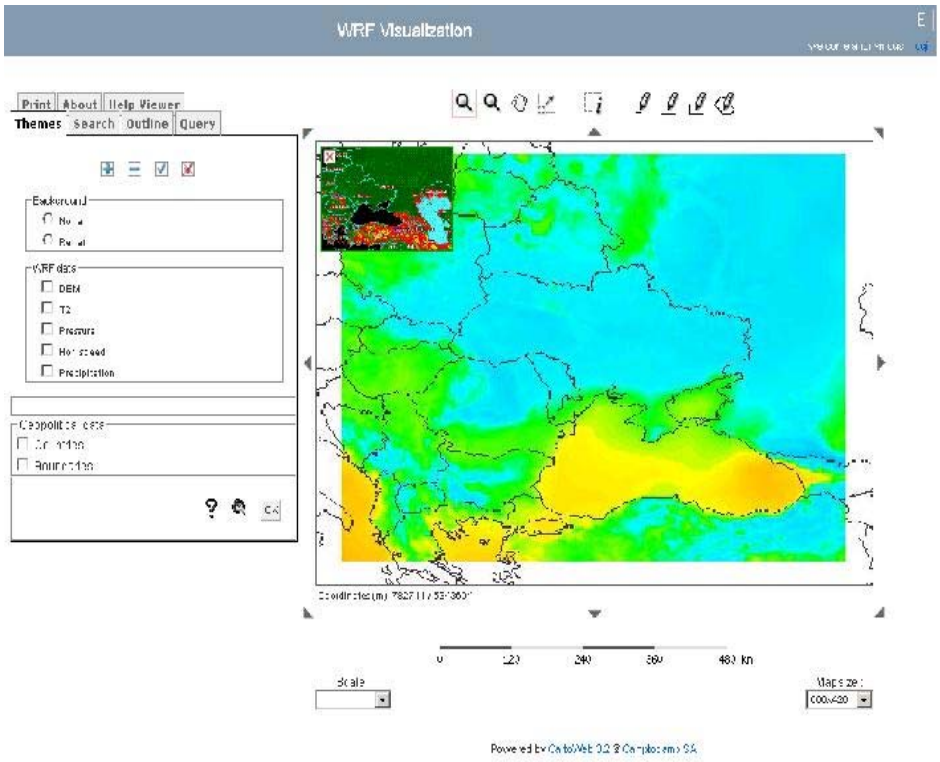


Fig. 6. Thick client system for WRF visualization

Acknowledgments. The work is partly supported by the STCU and NASU Targeted Initiatives Program, project "GRID technologies for environmental monitoring using satellite data" (No. 3872); NASU grant for Young Scientists "Development of Desktop Grid system and optimization of its productivity"; and NASU Innovative Project "Development of mathematical models, methods and information technologies of satellite data processing in the framework of test-bed information system of Ukrainian segment of GEOSS/GMES".

References

[1] Global Earth Observation System of Systems, http://www.epa.gov/geoss.

[2] Global Monitoring for Environment and Security, http://www.gmes.info.

[3] Fedorov, O.P., Kussul, N.N., Shelestov, A.Yu.: Problems and Prospects of Development of an Information Earth Observation System in the Ukraine (in russian). Journal of Automation and Information Sciences. Vol. 37, Issue 12, pp. 35-39. (2005).

[4] Ukrainian segment of GEOSS, http://geoss-ukraine.org.ua.

[5] Sectoral system CosmoGIS, http://cosmogis.org.ua.

[6] Space Research Institute's of NASU-NSAU service site, http://dos.ikd.kiev.ua.

[7] Kravchenko, O.M., Shelestov, A.U.: Using implementation of OGC standards to develop distributed systems for geospatial data visualization and delivery (in russian). Problems of programming. №2-3, pp. 135-139. (2006).

[8] WFS specification, http://www.opengis.org/techno/specs/02-058.pdf.

[9] WCS specification, http://www.opengis.org/techno/specs.

[10] WMS 1.1.1 specification, http://www.opengis.org/docs/01-068r2.pdf.

[11] UMN MapServer, http://mapserver.gis.umn.edu.

[12] Cartoweb, http://www.cartoweb.org.

[13] DOS-center of Space Research Institute of NASU-NSAU, http://dos.ikd.kiev.ua.

[14] Shelestov, A., Lobunets, A., Korbakov, M.: Grid-enabling Satellite Image Archive Prototype for UA Space Grid Testbed. International Journal "Information Theories and Applications. Volume 12, Number 4, pp. 351-357. (2006).

[15] Shelestov, A.Ju., Kussul, N.N., Skakun, S.V: Grid-infrastructure simulation. Problems of programming.Vol. №2-3, pp. 221-230. ISSN1727-4907. (2006).

Author's Information

A. Yu. Shelestov – *PhD, Senior Researcher, Department of Space Information Technologies and Systems, Space Research Institute of NASU-NSAU, Glushkov Ave 40, Kyiv-187, 03650 Ukraine, e-mail: inform@ikd.kiev.ua.*

O. M. Kravchenko *– Research Assistant, Department of Space Information Technologies and Systems, Space Research Institute of NASU-NSAU, Glushkov Ave 40, Kyiv-187, 03650 Ukraine, e-mail: inform@ikd.kiev.ua.*

M. I. Ilin *– Department of Space Information Technologies and Systems, Space Research Institute of NASU-NSAU, Glushkov Ave 40, Kyiv-187, 03650 Ukraine, e-mail: inform@ikd.kiev.ua.*

THE SPECIFICATION OF AGENT INTERACTION IN MULTI-AGENT SYSTEMS

Dmitry Cheremisinov, Liudmila Cheremisinova

Abstract: *The problem of the description of interaction between agents in the form of dialogues is explored. The concept of processes synchronization is analyzed to formalize the specification of interaction at the level of steps of the dialogue between two spatially divided agents. The approach to formalization of the description of conditions of synchronization when both the independent behavior, and the communications of agents can be presented at a logic level is offered. It is shown, that the collective behavior of agents can be specified by the synthetic temporal logic that unites linear and branching time temporal logics.*

Keywords*: multi-agent system, interaction protocol, time.*

ACM Classification Keywords*: I.2.11 [Computer Applications]; Distributed Artificial Intelligence, Multiagent systems; D.3.3 [Programming Languages]: Language Constructs and Features – Control structures, Concurrent programming structures*

Introduction

A multi-agent system can be considered as the organization of agents (by analogy to the human organization) or, in other words, as some artificial society. It is a computational system in which two or more agents interact or

work together to perform a set of tasks or to achieve a set of goals [1]. One of the core concept of multi-agent systems is *interaction*, that is the foundation for cooperative behavior among several autonomous agents. Agent interactions are established through exchanging information in the form of messages that specify the desired performatives of interacting agents. Agent system can operate if the agents have a common understanding of the possible types of messages, then they must know which messages they can expect in a particular situation and what they may do when they got some message. So messages exchanged between agents in some multi-agent system need to follow some standard patterns which are described in agent interaction *protocol*.

Protocols play the central role in agent communication. An interaction protocol defines the rules the dialog among agents conforms to. It constrains the possible sequences of messages that may occur in agent interaction. Interacting agents should comply with an interaction protocol in order to engage permissible sequences of message exchange. When agent sends a message it can expect a response to be among a set of messages indicated by the accepted protocol. The interaction protocol can be assigned by the designer of the multi-agent system otherwise an agent needs to indicate the protocol that it wants to follow before it starts to interact with other members of the system.

It is necessary for any protocol itself to be correct and verifiable. If it is not correct then the agents that follow it may perform contradictory and unexpected actions leading to possible breakdown of the interaction. The central problem of the verification of interactions (dialogues of negotiations) that take place in open (not being cooperative) systems is the problem of conformance inspection between behavior of agents and interaction protocol. That is the protocol must be understandable by all agents of the system and the they behave according to this protocol. The implementation of conformance inspection confront with a problem of identification of dialogue steps between agents. Recognition of the dialogue step which is carried out by two spatially divided agents requires analyzing the concept of interaction of processes.

At the heart of the formal models of a protocol are cooperating sequential processes. Fundamental feature, the proposed protocol models differ, is the degree of synchronization of behaviors of participants of interaction. There is still a need for a proper formalism for the process of synchronization that is suitable for human understanding and automated implementation. In this paper we focus on logical analysis of synchronization of behaviors of interacting participants. The simple yet expressive class of interactions is considered, namely dialogues consisting of separate steps. The considered dialogues involve only two agents. This restriction allows concentrating on the kernel of the problem of synchronization in different formal models of interaction protocols. The agent interaction is considered as interaction between two (or more) processes. And a special case of such interaction is considered, when one of processes outputs at the same time as the other one inputs it. The actions of message exchanging have duration.

Formalization of the concept of interaction event

Usually collective behaviour of multi-agent system is described as a dialogue of agents which communicate by means of sending and receiving messages. On each step of activity an agent carries out some action depending on its internal state and the received message. As a result of the action the agent changes its internal state and sends some messages to other agents. Speaking informally, the architecture of an agent includes 1) the internal structures of data defining internal states of the agent, 2) mail box containing messages from other agents, 3) integrity restrictions on the agent internal states, 4) actions which the agent can execute, and 5) the program that specifies the control of action execution. Execution of an action consists of 1) changing a current internal state of the agent and 2) sending a message to other agents. The current contents of a mail box consist of the messages received by the agent from other agents on the previous step. The global state of a multi-agent system consists of internal states and contents of mail boxes of all agents of the system [1].

To specify independent behaviors of agents, formalisms of high level abstractness are widely used, for example, such as temporal logic. At the same time the communications between agents are specified by means of the concepts of realization level, such as mail boxes and messages. One of the problems of such segregated

approach to interaction lies in that it is extremely difficult to simulate interactions between agents though at the same time the independent behaviour of separate agents is described completely. This problem arises due to the absence of agent model unifying all aspects of both independent behaviours and the communication. The main reason of the absence of such a general model is that there exists no general conceptual basis unifying all abstractions, connected with collective behaviour of agents.

When analyzing the behaviour of multi-agent system agents are characterized by processes. The process is specified by exhaustive description of potential behaviour of the agent. The process consists of events. Thus, to be in position to analyze the concept of interaction of processes, a suitable axiomatization of the concept of an event is required.

The concept of an event allows abstracting from physical time when describing behaviour of a system. The widespread axiomatization of an event is connected with the assumption, that events have no duration [2]. The behaviour of a multi-agent system consists of some events – steps of dialogue between agents – and is sequential in this sense. For recognition of the step which is carried out jointly by two spatially divided agents, it is impossible to bypass the concept of parallelism.

The models of parallelism known in the literature could be roughly divided into two classes: 1) the models, in which concurrent execution of two processes is described by interleaving of (atomic) events of those processes; 2) models in which causal dependencies between events are set explicitly. Interleaving models are focused on systems with events considered as instantaneous and indivisible. In this case the act of interaction is a complete event which describes participation of all processes cooperating in this act [2]. This act as the step of a dialogue is carried out by two spatially divided agents and represents the event which should have duration and structure.

There is popular opinion the concept of an event having duration is reduced to the concept of an instantaneous event. The following formulation of this assumption is taken from Hoare [3, p. 24]; “The actual occurrence of each event in the life of an object should be regarded as an instantaneous or an atomic action without duration. Extended or time-consuming actions should be represented by a pair of events the first denoting its start and the second denoting its finish.”

Now it is known that this opinion is erroneous, and often splitting, i. e. the use of pairs of instantaneous events to model events having duration, is unnatural. Mutual irreducibility of the concept of an event having duration and the concept of instantaneous event is proved formally and constructively [4]. The formal proof is based on the incomparability of these formalisms describing event systems [5]. Systems of the durational events are described by the causal relation (branching-time temporal logic), systems of instantaneous events – by relation of consequence and parallelism (linear-time temporal logic).

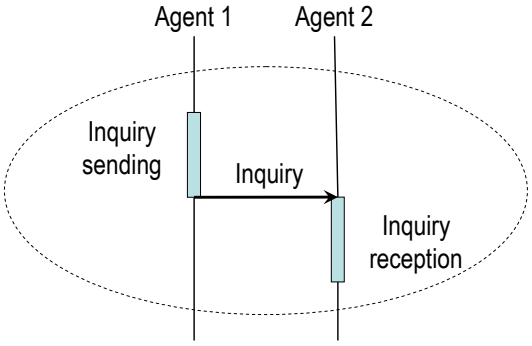


Fig. 1. Structure of the first model of a step of dialogue

The problem under discussion is how processes and events can be assembled together into a system in which the components interact with each other and with their external environment. The elementary structure of the dialogue step is a pair of durational events which constitute the step densely without a time interval between. First event of the pair can be interpreted as “pronouncing” of the message by one of the agents; the second event can be interpreted as “perception” of this message by the other participant of the dialogue. The basic feature of this structure is the assumption of density of the event composition and that members constituting the event belong to behaviours of different agents (fig. 1). Absence of a time interval between of pair of durational events designates that the event of synchronization of corresponding processes is instantaneous.

The first model of an interaction event

Ignoring the functionality of agents, it is possible to consider synchronization of their behaviour as the only goal of interaction. Thus, the dialogue step is the composition of three events, two events are durational ones, and the other is instantaneous event. However the events constituting interaction still remain occurring simultaneously in different processes.

On the one hand, synchronization of agent behaviours occur during the rare moments, in the rest of the time communicating agents behave independently from each other. On the other hand, processes should interchange information about current states to ensure synchronization. Formally it can be reached by splitting of all events constituting agent behaviour on internal and external ones. Only external events of the agent behaviour can be "visible" to the other agents. In this case the specification of the agent behaviour is the cause-effect relation on a set of possible events. In particular, this relation describes the reasons of occurrence of external and internal events.

Let a composition of a durational internal event E and instantaneous external event y is an operation. A composition $y \rightarrow E$ of durational and instantaneous events is called as a waiting operation that waits the external event y, and a composition $E \rightarrow y$ is called as an acting operation which effect is the realization of external event y. It is necessary to note, that the event sequence in both operations is the same: the first one is durational event, the second is instantaneous event. The mentioned compositions allow considering dependences between events in a composition as cause and effect because from physical reasons, event-consequence occurs behind event-reason without overlapping on time. Waiting operation is a durational event and the reason of its termination is the occurrence of y. Acting operation is a durational event too and it is the reason of occurrence of y. The symbol "$\rightarrow$" can be interpreted as cause and effect dependence between events.

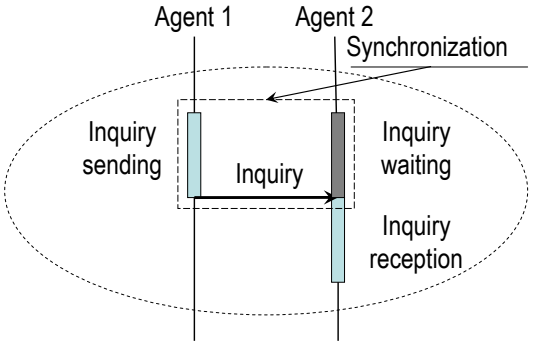


Fig. 2. Synchronization of behaviour of agents

Such treatment of waiting and acting operations is a basis of the formal semantics [6, 7] of PRALU language [8] in which conjunctions of Boolean variables describe external events of operations. PRALU language in this interpretation represents the synthetic temporal logic uniting linear and branching time temporal logics supplied by the assumption of density of time [9]. Temporal formulas of this logic are interpreted as the statements concerning event sequences of two sorts: instantaneous and durational.

The composition of events considered above allows describing independent behaviours of agents. Parallel execution of waiting operation $y \rightarrow E$ by one agent and acting operation $E \rightarrow y$ by the other one results in synchronization of behaviour of agents during the moment of occurrence of instantaneous event y (fig. 2). By the definition the effect of a waiting operation is its termination at a moment of occurrence of instantaneous external event y.

The line of "life" of the agent consists of pairs waiting and acting operations. The boundary between waiting and acting operations serves as the synchronization event. Here the action consists from "perception" of the accepted message and "pronouncing" new one. Obviously, the occurrence of synchronization depends on duration of acting operations.

The second model of an interaction event

A dialogue step can be considered also as the other composition of some three events. One of them is durational, and the others are instantaneous (fig. 3). In this model of a dialogue step the interaction itself is a durational event. This event, having duration, should have a physical basis. Without loss of a generality it is possible to

consider that event of interaction occurs in an environment of agents. In this model interaction event becomes not distributed, but a local one in the external environment.

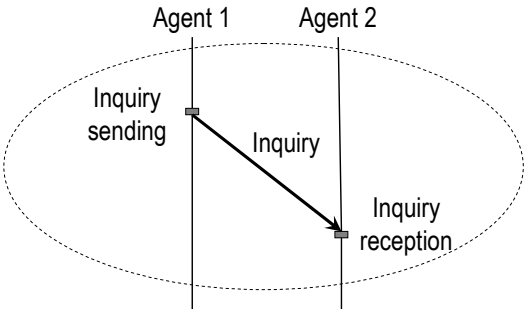


Fig. 3. Structure of the second model of an interaction event

Concept of an environment

From analysis of physical realizations of distributed systems it follows that the synchronization requires a special organization of a system of cooperating agents. Two basic types of the system organization aimed achieving synchronization are known: synchronous and asynchronous systems. The standard definition of the distinction of these types of systems declares that synchronous systems have the same shared “clock”, and in asynchronous systems each agent has its own independent clock. It is obvious, that the shared “clock” belong to an external environment of all agents of the synchronous system.

In traditional interaction theories CCS [3] or CSP [10] the concept of an environment is used implicitly, hence it is not formalized. CCS and CSP rely on the concept of an environment having the following distinguishing features. The environment is considered simply as the other agent. In other words, the environment for the given agent includes all other agents of the system that operate in parallel with this agent. In this case an agent and an environment are objects of the same nature. It is obvious, that this assumption of properties of an environment is not good enough from the point of view of specifying an agent interaction directed to achievement of synchronization. Such an approach is justified only by the following reasoning. It is considered that the concept of an environment concerns with the system realization and it is not represented at the level of agent behaviour.

Our purpose is to offer a formal model of interaction which is not concerned with a system realization. We consider an environment is essentially distinct from agents. The basis of this approach is that the interaction is considered as the communication act consisting of sending and receiving of messages. This formalization of interaction originates from Shannon’s paper about the theory of communication [11] in which interaction is considered as a way to transfer the message from a sender to a receiver through a medium, also called as transfer environment. Physical realization of an environment can be a computer program, a device or a physical environment.

Obviously, synchronization of agent behaviours is impossible without fixing data which are transferred by an environment during agent interaction [12]. Thus, environment serves as a model of transport system to deliver messages. In other words the environment can be considered as the memory that is shared with all agents. This memory is known as a global state of multi-agent system. In its most simple form, the communication can be based on the fixed set of differing signals. In the case of binary signals the representation of a global state is a set of the Boolean variables which values are possible signals. In the case of structural signals an agent environment usually refers as message passing system. The concept of an environment is closely concerned with a notion of autonomy of agents. Autonomy of agents has its focus on freely choosing between actions and on acting independently. Autonomy means also that the agents receive all information only through an environment.

System, in which the behaviour of an environment is deterministic, refers to closed system. In the case of a closed system it is supposed, that the reasons of all events are inside of the system and its behaviour is completely self controlled. If the behaviour of an environment is nondeterministic, the system refers to an opened

system. Unlike the memory considered in the theory of finite state machines, the behaviour of the memory of an environment of the open multi-agent system can depend on uncontrollable conditions.

The specification of interaction in the form of description of a message passing system does the description of autonomous behaviour of the separate agent not closed because this description is not enough for understanding of the complete behaviour of the agent. Obviously, most important property of the message passing system is restriction on length of durational events, imposed by this system.

Time as a logic concept

In the previous part of the paper time was considered as the logic concept expressed by relations between events through their sequence and order. Time is discrete, because there is an observable time quantization by the events that is fixed in behaviour of an environment. Time cannot be measured, if we do not impose some restrictions on the duration of events in all components of a multi-agent system. Time is measured if each event in a history of the system behaviour is accompanied with a number that expresses either duration of the event or specifying the moment of time when it occurs. Synchronization of behaviour of agents means that time is measured.

Measured time can be realized, if we assume, that the duration of all simultaneously executed acting operations in a multi-agent system is identical. It is natural to accept this duration as the unit of time. In this case in the closed systems the duration of waiting operations is expressed by an integer $i \geq 1$. The assumption that duration of all simultaneously executed acting operations is identical holds in synchronous systems. Obviously, this assumption specifies a pairs of interacting waiting and acting operations by the counter number of the appropriate step of time. Synchronous system keeps the assumption that the time is discrete and measured.

Other assumption that allows realizing measured time is that any operation, carried out in parallel to itself, is illegal. In this case realization of any operation in a history of the agent functioning can be accompanied with some counter number of this realization. The formal proof of this statement is in [7]. The function which calculates a counter number of the operation realization (from the start of the system) when this operation starts can be used for measurement of time. Interaction occurs only in pairs of waiting and acting operations which have the same counter number. This is known as a rendezvous condition.

An asynchronous system keeps the assumption that time is discrete and measured, but rejects the assumption that duration of all simultaneously executed acting operations is identical. The last principle of measuring time differs from that for synchronous system.

Conclusion

The independent behaviour of agents in the majority of models of multi-agent systems is described by means of formalisms of high level abstractness, but the communication is specified by the concepts close to realization. The difference of levels of the description does not allow simulating communications between agents at the level in which their independent autonomous behaviour is described. This problem arises because of absence of agent models that unify all aspects of local behaviour and the communications.

In the paper we suggest to describe the synchronization conditions by specification of event properties which have been not concerned with the realization of these events. Our approach allows specifying both the independent behaviour and the communication at a level of logic. It is shown, that the collective behaviour of agents can be described by the synthetic temporal logic that unites the linear and branching time temporal logics. Such synthetic logic is one of interpretations of the existing PRALU language.

Acknowledgements

The research was supported by the Fond of Fundamental Researches of Belarus (Project **F07-125**).

Bibliography

1. Subrahmanian V.S., Bonatti P., Dix J. et al. “Heterogeneous Agent Systems”, MIT Press, 2000.
2. Brookes S.D., Hoare C. A.R., and Roscoe A.D. “A Theory of Communicating Sequential Processes”, Journal of the ACM, no 31(3), pp. 560–599, 1984.
3. C.A.R. Hoare “Communicating Sequential Processes”, Prentice Hall International Series in Computer Science, 1985.
4. Van Glabbeek R., Vaandrager F. “The Difference between Splitting in n and n+1”, Report CS-R9553, Centre for Mathematics and Computer Science, Amsterdam 1995; Abstract in: Proceedings 3rd Workshop on Concurrency and Compositionality, Goslar, March 5-8, 1991 (E. Best & G. Rozenberg, eds.), GMD-Studien Nr. 191, Sankt Augustin, Germany 1991.
5. Cheremisinov D.I. “The Real Difference between Linear and Branching Temporal Logics”, Workshop on Discrete-Event System Design DESDes ’ 04, University of Zielona Gora Press, Poland, 2004, pp. 103–108, 2004.
6. Cheremisinov D.I. “Formal description of behaviour of the distributed systems”, Minsk: Belarus, 1991, preprints no 38. – 44 p. (in Russian).
7. Cheremisinov D.I. “The morphisms of reactive system formal languages”, Informatics, no 1 (5), pp. 76–88, 2005 (in Russian).
8. A.D. Zakrevskij, ”Parallel algorithms for logical control”, Minsk, Institute of Engineering Cybernetics of NAS of Belarus, 202 p., 1999 (in Russian).
9. Cheremisinov D.I. “About interpretation of temporal logic at symbolical verification”, Informatics, no 1, pp. 131–138, 2004 (in Russian).
10. Milner R. “A calculus of communication systems”, LNCS 92, Springer Verlag, 1980.
11. Shannon C.E. “A mathematical theory of communication”, Bell System Technical Journal, vol. 27, pp. 379–423 and pp. 623–656, 1948.
12. Odell J., Parunak H. V. D., Fleischer M., Brueckner S. “Modeling Agents and their Environment”, Proceedings of the Third International Workshop on AgentOriented Software Engineering, Lecture Notes in Computer Science, Springer Verlag (Berlin, D), vol. 2585, pp. 16–31, 2003.

Authors' Information

Dmitry Cheremisinov, Liudmila Cheremisinova *– The United Institute of Informatics Problems of National Academy of Sciences of Belarus, Surganov str., 6, Minsk, 220012, Belarus, Tel.: (10-375-17) 284-20-76, e-mail: cher@newman.bas-net.by, cld@newman.bas-net.by*

INFOSTATION-BASED PARKING LOCATOR SERVICE PROVISION WITHIN A UNIVERSITY CAMPUS

Ivan Ganchev, Máirtín O’Droma, Damien Meere

Abstract*: This paper presents an InfoStation-based multi-agent system, which facilitates the Parking Locator Service Provision within a University Campus area. The network architecture is outlined, illustrating how it functions during service provision. A description of the Parking Locator service is detailed, highlighting how the different entities interact in order to facilitate the service. Approaches to the implementation of the system are considered.*

Keywords*: InfoStations, intelligent agents, multi-agent system, JADE, LEAP*

I. Introduction

The InfoStations paradigm is an infrastructural system concept supporting "many-time, many-where" (Frenkiel and Imielinski 1996) wireless communications services. The InfoStation-based system outlined in this paper is established and operates across a University Campus area for the purpose of enhancing the mobile services experience. It allows mobile devices (mobile phones, laptops, personal digital assistants–PDAs) to communicate to each other and to a number of servers through geographically intermittent high-speed connections. In this paper, we detail the underlying network architecture and show how the different components within the architecture collaborate to facilitate one particular service, namely the Parking Locator service. This service allows registered users to locate available parking spaces throughout the campus.

The rest of the paper is organized as follows. Section II presents the InfoStation-based network architecture, illustrating how the architecture functions during service provision. Section III illustrates the Parking Locator service provision outlining sample interactions between system entities. Section IV outlines some implementation issues, and finally Section V concludes the paper.

II. InfoStation-based Network Architecture

The following InfoStation-based network architecture (Ganchev, O'Droma et al. 2003; Ganchev, Stojanov et al. 2006; Ganchev, Stojanov et al. 2006) provides access to a number of very useful services, for users equipped with mobile wireless devices, via a set of InfoStations deployed in key points around a University Campus. The 3-tier network architecture consists of the following basic building entities as depicted in Figures 1 and 2: user mobile devices, InfoStations and an InfoStation Center.

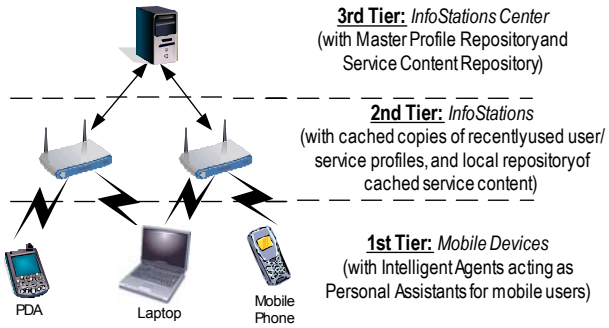


Figure 1. The 3-tier InfoStation-based network architecture

The users request services (through their mobile devices) from the nearest InfoStation via available Bluetooth (IEEE 802.15 WPAN), WiFi (IEEE 802.11 WLAN), or WiMAX (IEEE 802.16) connections. The InfoStation-based system is organized in such a way that if the InfoStation cannot fully satisfy the user request, the request is forwarded to the InfoStation Center, which decides on the most appropriate, quickest and cheapest way of delivering the service to each user according to his/her current individual location and mobile device's capabilities (specified in the user profile).

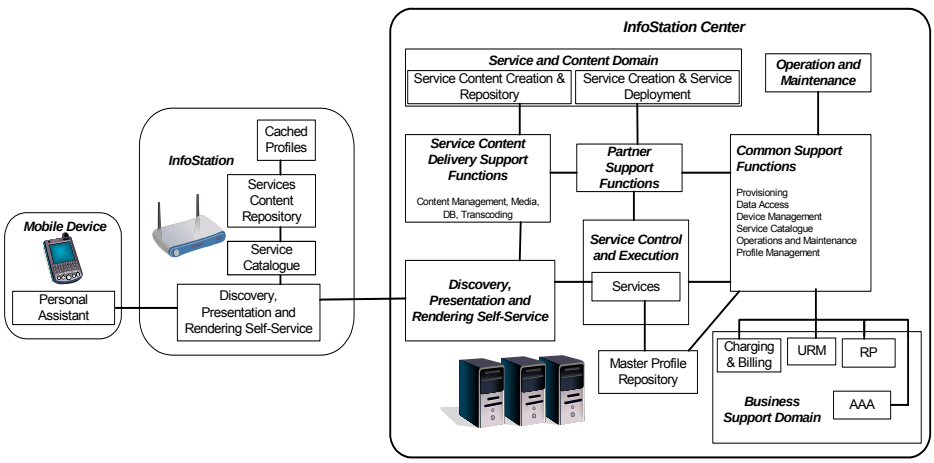


Figure 2: The InfoStation System Architecture

Figure 2 illustrates some of the main components within each entity of the architecture.

The *InfoStation Center* is concerned with the creation of service content and service creation, deployment, operation, maintenance, control and execution. In addition there are some common support functions that each service requires when initially created, for example device management, profile management, service catalogue etc. The InfoStation Center also houses a repository of all (up-to-date) master profiles relating to both users and services alike. Any changes made by the individual user to his/her own user profile and/or user service profile are forwarded on from the user mobile device, through an InfoStation to the InfoStation Center, where the repository is updated. (Each InfoStation keeps cached copies of all recently used, or updated by users, profiles.) The InfoStation Center also houses the Business Support Domain with a number of components relating to the charging and billing of users, User Relationship Management (URM), Resource Planning (RP) and indeed user Authentication, Authorization and Accounting (AAA).

When a mobile user enters within the range of an *InfoStation*, the Personal Assistant, installed in the user mobile device, and the InfoStation mutually discover each other. This process is facilitated through the Discovery, Presentation and Rendering Self-Service module within the InfoStation. The Personal Assistant sends a request to the InfoStation for user's Authorization, Authentication and Accounting (AAA). This request also includes a description of the mobile device currently being used by the user (or just the device's make and model) as well as any updates of user profile and user service profile (Figure 3). In particular with the Intelligent Parking Locator service, this process may occur a number of times as the user will, more often then not, pass through a number of InfoStation coverage areas with his/her vehicle.

The InfoStation forwards this AAA request to the InfoStation Center along with the profile updates (Figure 3).

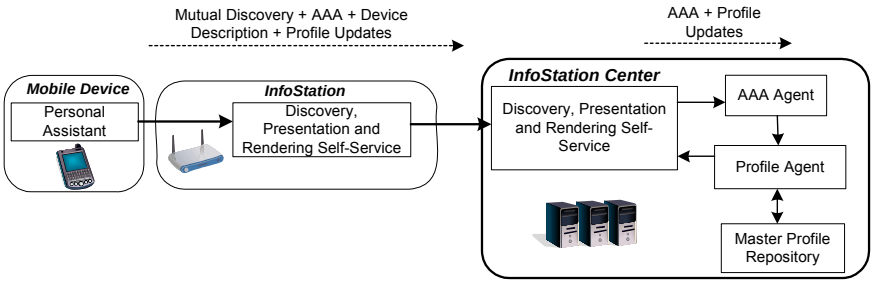


Figure 3: Step 1- Initial AAA and profile updates.

If the user is successfully authenticated and authorized to utilize the services by the AAA module within the InfoStation Center, a new account record is created for the user. The user profile is analyzed by the InfoStation Center for current user preferences (e.g. applicable services) and device capabilities (utilizing the Composite Capabilities/Preference Profile – User Agent Profile, UAProf). Then the InfoStation Center makes a service offer to the user in a form of a compiled list of applicable services from the Service Catalogue (Figure 4).

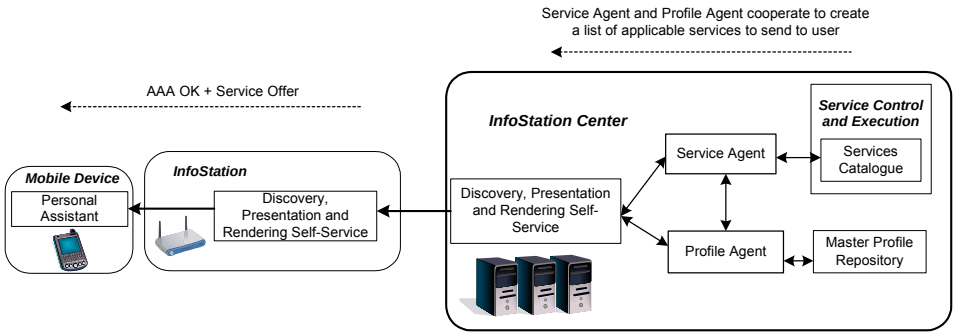


Figure 4: Step 2- Service Offer

This service offer is sent towards the Personal Assistant along with the AAA acknowledgment. The Personal Assistant displays the offer to the user who makes a choice and selects (makes a request for) the service s/he wishes to use. When the Personal Assistant forwards the user service request to the InfoStation, the latter checks its cache for the most up-to-date version of the requested service content (e.g. campus news bulletin). If the InfoStation is able to satisfy fully the user service request, it does so (Figure 5). Otherwise the InfoStation forwards this request to the InfoStation Center, which is better equipped to deal with it.

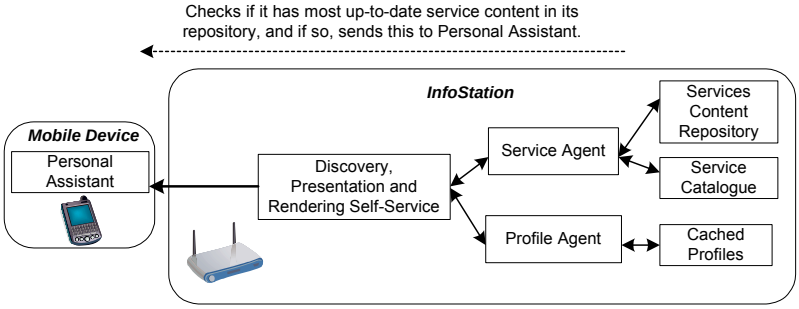


Figure 5: Step 3- Service request satisfied by InfoStation

On the user *mobile device*, the Personal Assistant (*agent*) facilitates the service utilization by the user. This is down to an agent-oriented approach to the implementation of the system. The service migrates onto the users mobile device, allowing the user unhindered access to the service even when out of range of the InfoStation. The Personal Assistant may make a service request while within the range of an InfoStation, then may pass out of the coverage area but will continue to work autonomously, adopting the functionality of the service until the user has completed his/her task. Once the mobile device comes within range of another InfoStation, the Personal Assistant updates and synchronizes the user service profile to reflect any work completed, or any new service requests made by the user while out of range.

In the following section we describe the provision of the Intelligent Parking Locator service in more detail.

III. Intelligent Parking Locator Service

A multi-agent approach (Carabelea and Boissier 2003; Ganchev, Stojanov et al. 2004; Stojanov, Ganchev et al. 2005; Adaçal and Bener 2006; Ganchev, Stojanov et al. 2006) is adopted as most suitable approach to structuring our system. In order to facilitate flexible and adaptable service provision, intelligent agents, residing within each of the three tiers of the system architecture must interact so as to satisfy, in the 'best' possible way, any user requests they might encounter. The following description outlines the entity interactions that take place during the Intelligent Parking Locator service provision. This service allows registered mobile users to gain access to information regarding available parking spaces on the University Campus and reserve a space that best suits them when approaching/entering the campus. However, visitors may also gain access to this service through prior temporary registration in the system for the duration of their stay. On accessing the service, these visitors would be directed to a visitor's car park.

In the delivery of this service, the content must be adapted and customized according to the capabilities of the user device and the user preferences. For example if the user has access to a resource-rich mobile device (e.g. a laptop or indeed a PDA), s/he may gain access to a graphical representation of the campus, which would greatly assist the user in finding the required parking space. If however the user only has access to a device with limited capabilities (e.g. a mobile phone), then the details of the available parking spaces would be specified in a simple format which 'best' suits the device (e.g. SMS/MMS). This trimming (adaptation) of the services is one way to address the shortcomings of some mobile devices, while still delivering the service.

We use the "Composite Capabilities / Preference Profile" (CC/PP) as the uniform format for the implementation of the user profiles. The master profile repository in the InfoStation Center contains descriptions of all registered

user devices, i.e. their capabilities and technical characteristics. During the initial AAA request, the user's Personal Assistant sends as parameters the make and the model of the user device. An agent working on the InfoStation (or the InfoStation Center) reads the corresponding device's description from the repository and according to this, selects and forwards the best format of the service content. However a problem arises when a user uses a non-registered device as s/he might receive the service content in unsuitable format. Thus the user needs first to register any new mobile device s/he wants to use within the system. In this case, during the initial AAA request the Personal Assistant sends a full description of the user device's capabilities towards the InfoStation Center.

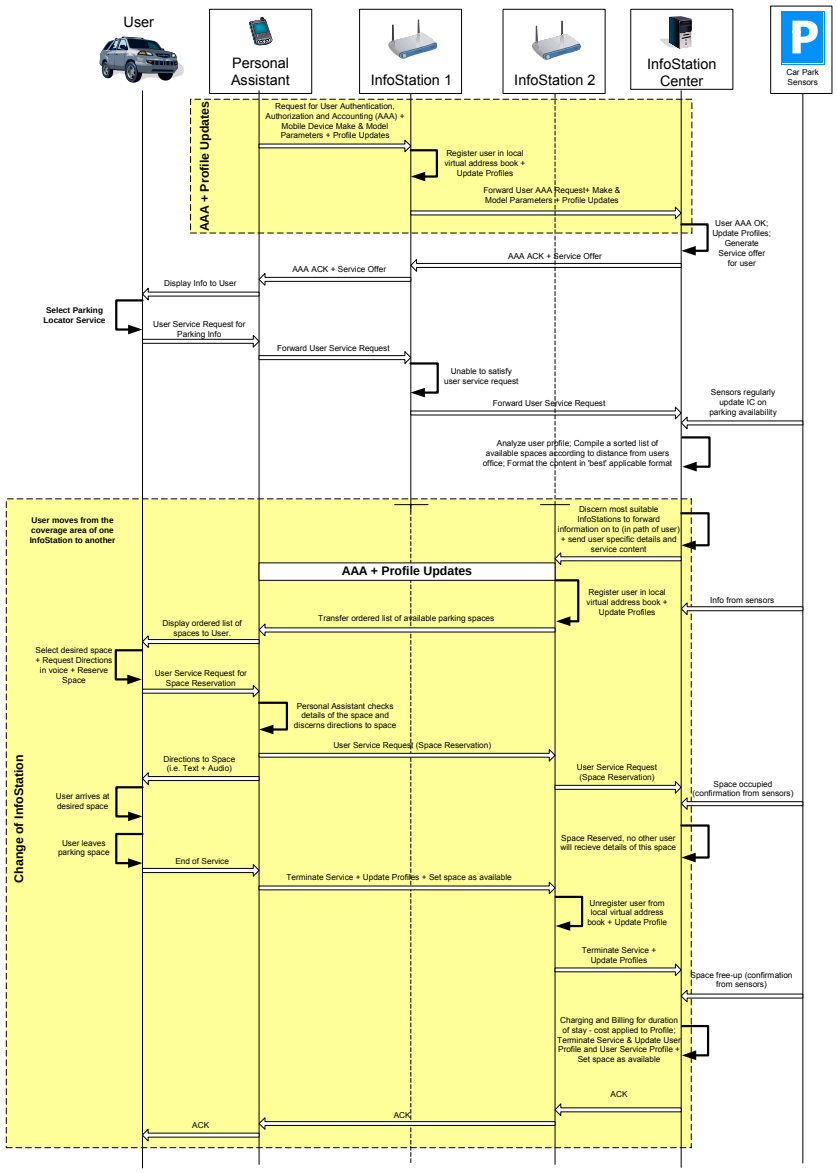


Figure 6: Intelligent Parking Locator Service: Entity Interactions

Figure 6, depicts a sample interaction between entities involved in the Intelligent Parking Locator service provision. As the user enters the campus area in a vehicle, s/he enters the coverage area of an InfoStation, positioned at the entrance to the campus. The Personal Assistant, installed in the user mobile device, and the InfoStation mutually discover each other. The Personal Assistant sends a request to the InfoStation for user's Authorization, Authentication and Accounting (AAA). During this initial AAA request, the user's Personal Assistant sends also the make and the model parameters of the user device, and any updates of user profile and user

service profile. The InfoStation registers the user in its local Virtual Address Book and updates the profile, before forwarding the user request onto the InfoStation Center along with profile updates. In the case of successful AAA, the Profile Agent within the InfoStation Center (updates and) analyses the user profile stored in its Master Profile Repository. The Service Agent, in collaboration with the Profile Agent, creates a list of services applicable to the user and makes a service offer to the user. However the user may specify in his/her profile that a request for the Parking Locator service be sent automatically after the successful AAA (and profile update) procedure. Or alternatively, if the user makes regular use of the service, the Personal Assistant could proactively anticipate the users request, i.e. once this service becomes available, the Personal Assistant automatically requests the location of parking for the user's vehicle. The InfoStation forwards on the user request to the InfoStation Center. Sensor networks within the car parks constantly update the InfoStation Center as to the availability of spaces. Different time periods of the day require more regular updates, especially from morning to mid-afternoon, as the user would require the information be as up-to-date as possible. However the updates can occur at much larger intervals during the evening and weekends when many more spaces would be available. The InfoStation Center discerns the location of the user's office from the user profile, and as such compiles a list of available parking spaces in the proximity of that office (the spaces are ordered according to their distance from the users office). The Service Content agent and Profile Agent cooperate to adapt the content to the format that best suits the current user device capabilities and user preferences (i.e. graphical representation, audio description, text). Once the content is prepared for transfer, the InfoStation Center discerns the most suitable InfoStation to forward the data on to. As the user is most probably accessing the service whilst in transit, there is a good chance the user will pass through a number of InfoStation coverage areas. The InfoStation Center makes allowances for this and forwards the information on to a number of InfoStations in the path of the user (Borràs and Yates 1999; Yuen, Yates et al. 2003), along with specific user details.

As the user moves from the coverage area of one InfoStation to another, AAA and profile update procedures are executed first. The approached InfoStation will have already received information about the user from the InfoStation Center, along with requested service content. As such the InfoStation can account for the user and immediately forward on the requisite content. This reduces the time taken for the InfoStation to provide the service content. This process may happen with a number of InfoStations as the user drives through the campus. As the user leaves the coverage area of an InfoStation, the user service profile is updated, specifying how much of the service content was transferred (if the transaction was not completed). This information is circulated around the InfoStation network, so as to ensure the user's Personal Assistant does not receive the same information a number of times.

Once the user receives the ordered list of parking spaces, s/he chooses a particular parking space, reserves it and request directions to that space. Once a space is chosen, the Personal Assistant examines the details of the space and displays precise directions. An audio explanation accompanying the text description would be best suited to this service, as it would provide the least distraction, allowing the user to concentrate on driving.

The Personal Assistant also forwards on a parking space reservation request to the InfoStation Center. Once the space has been reserved (and it's occupancy confirmed by the sensor network), no other users will be supplied with details of that space. The InfoStation Center monitors the duration the user occupies the parking space for charging and billing purposes.

When the user leaves the parking space, the sensor network confirms this to the InfoStation Center. The Charging and Billing Module within the InfoStation Center accounts for the duration of the user's stay. A corresponding charge related to parking in that car park, is charged to the user account. Once the user/ service profiles have been updated, the service is terminated.

IV. Implementation

The system is implemented in an agent-oriented manner utilizing the Java Agent DEvelopment (JADE) (JADE; Bellifemine, Poggi et al. 2001; Anghel and Salomie 2003; Bellifemine, Caire et al. 2003; Bellifemine, Caire et al.

2005; Bellifemine, Caire et al. 2006) framework. This allows for the flexible development of multi-agent systems and applications for management of network resources in compliance with the FIPA specifications. The JADE architecture is completely modular and as such, by utilizing specific modules, can be configured to adapt to the requirements of a number of different deployment environments. Within our JADE implementation, one of the most useful modules is the Lightweight Extensible Agent Platform (or LEAP) (Moreno, Valls et al. 2003; Caire and Pieri 2006) Module. This module, or add-on, replaces some parts of the JADE kernel, providing a modified run-time environment, which facilitates the implementation of agents on mobile devices with limited resources. Another very useful aspect of JADE-LEAP is its ability to support split-containers (split run-time environments) on resource-thin devices. The container can be split into two separate sections, a FrontEnd (running on the mobile device itself), and the BackEnd (running from a fixed network entity - a mediator) as illustrated in Figure 7. This mediator is charged with instantiating and maintaining the BackEnds. In our system, the InfoStations deployed throughout the campus take on these mediator roles. Each FrontEnd is connected to each BackEnd through a bi-directional connection. The splitting of the container into two separate, yet connected, entities is particularly useful in the realm of resource-constrained devices, as the FrontEnd of the container is far more lightweight in terms of the required memory and processing power then the entire container. Due to the geographically intermittent nature of the InfoStation connection, the FrontEnd and the BackEnd may undergo a loss of connection, however the Front-End can detect this and re-establish the connection as soon as possible. Any messages not transmitted due to this temporary disconnection can be buffered and delivered when the connection is re-established. This store-and-forward mechanism (implemented in both the FrontEnd and the BackEnd) is especially important to the efficient facilitation of the Parking Locator Service, where the user will pass in and out of coverage range of a number of different InfoStations, and as such data will have to be buffered and transmitted after a period of time by another InfoStation.

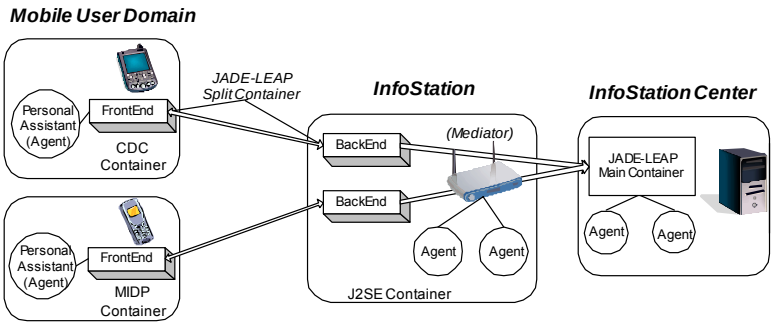


Figure 7. JADE-LEAP split-container execution

The splitting of the container has no bearing on us, as the same functionality and set of APIs are available to an agent, whether it is contained within a full container or the FrontEnd of a split container. The JADE framework also serves to shield us from the complexity of the distributed environment, allowing the concentration of our efforts on developing the application logic, rather then worrying about middleware issues such as discovery and communication of entities within the system.

V. Conclusion

The effectuation of the InfoStation-based Parking Locator service in a University Campus area has been outlined in this paper. The underlying network architecture has been described along with an illustration of how each of the different components within the architecture collaborate to facilitate mobile services. The Parking Locator service has been considered as an example. This service allows registered users to locate available parking spaces throughout the campus and reserve a space that best suits them when approaching/entering the campus area. Details of how service content is tailored to specific devices and how the duration of the user's stay affects the charging and billing for utilization of the service have been outlined.

The multi-agent structure, implemented by means of the Java Agent DEvelopment (JADE) software framework utilizing its Lightweight Extensible Agent Platform (LEAP) module in particular, has been discussed in detail due to its suitability to the proposed system. The benefits of this implementation have been also outlined in detail.

Acknowledgments

Dr. Ivan Ganchev, Dr. Máirtín O'Droma, and Damien Meere wish to acknowledge the financial support of the Ireland's HEA Strategic Initiatives Funding Program 'Technology in Education' for the development of the system.

Bibliography

Adaçal, M. and A. Bener (2006). "Mobile Web Services: A New Agent-Based Framework." IEEE Internet Computing Vol. 10(no. 3): pp. 58-65.

Anghel, C. and I. Salomie (2003). JADE Based solutions for knowledge assessment in eLearning Environments, TILAB & University of Limerick.

Bellifemine, F., G. Caire, et al. (2003). "JADE: A White Paper." exp, Telecom Italia Lab Volume 3(No. 3,).

Bellifemine, F., G. Caire, et al. (2005). JADE Programmers Guide, TILab.

Bellifemine, F., G. Caire, et al. (2006). Jade Administrator's Guide, TILab.

Bellifemine, F., A. Poggi, et al. (2001). JADE: A FIPA2000 Compliant Agent Development Environment. AGENTS '01, Montreal, Quebec, Canada.

Borràs, J. and R. D. Yates (1999). Highway InfoStations. WPMC'99, Amsterdam.

Caire, G. and F. Pieri (2006). LEAP User Guide, TILab.

Carabelea, C. and O. Boissier (2003). Multi-agent platforms on smart devices: Dream or reality? Smart Objects Conference (SOC03), Grenoble, France.

Frenkiel, R. H. and T. Imielinski (1996). "Infostations: The joy of `many-time, many-where' communications." WINLAB Technical Report(WINLAB-TR-119).

Ganchev, I., M. O'Droma, et al. (2003). A model for integration of electronic services into a distributed eLearning center. 14th EAEEIE International Conference, Gdansk, Poland.

Ganchev, I., S. Stojanov, et al. (2006). An InfoStation-Based Multi-Agent System for the Provision of Intelligent Mobile Services in a University Campus Area. IEEE-IS'06, London.

Ganchev, I., S. Stojanov, et al. (2006). An InfoStation-Based University Campus System for the Provision of mLearning Services. IEEE-ICALT '06, Kerkrade, The Netherlands.

Ganchev, I., S. Stojanov, et al. (2004). Enhancement of DeLC for the Provision of Intelligent Mobile Services. 2nd International IEEE Conference on Intelligent Systems (IS'2004), Varna, Bulgaria.

JADE Java Agent Development Framework Project - http://jade.cselt.it.

Moreno, A., A. Valls, et al. (2003). "Using JADE-LEAP to implement agents in mobile devices." TILAB "EXP in search of innovation", Italy.

Stojanov, S., I. Ganchev, et al. (2005). An Approach for the Development of Agent-Oriented Distributed eLearning Center. International Conference on Computer Systems and Technologies - CompSysTech, Varna, Bulgaria.

Yuen, W. H., R. D. Yates, et al. (2003). Effect of Node Mobility on Highway Mobile Infostation Networks. ACM MSWiM 2003, San Diego.

Authors' Information

Dr. Ivan Ganchev *– Dip. Eng. (honours), PhD, IEEE (M.), IEEE ComSoc (M.), a Lecturer and a Deputy Director of the Telecommunications Research Centre, University of Limerick, Ireland. He is currently a Track Co-chair of the 65th IEEE VTC2007 Spring conference and was a TPC member of the IEEE Globecom2006 conference. e-mail: ivan.ganchev@ul.ie*

Dr. Máirtín S. O'Droma *– B.E., PhD, C.Eng., FIEE, IEEE (SM), a Senior Lecturer and Director of the Telecommunications Research Centre, University of Limerick, Ireland. He is currently Publications Chair and Track Co-Chair of IEEE VTC 2007 Spring.*

Damien Meere – *researcheri in the Telecommunications Research Centre (TRC) in the University of Limerick. He is currently pursuing his MEng degree leading to transfer to PhD.*

ECOLOGICALLY INSPIRED DISTRIBUTED SEARCH IN UNSTRUCTURED PEER-TO-PEER NETWORKS

Li Sa, Yongsheng Ding

***Abstract**: In this paper, we reported an ecosystem inspired algorithm for searching peer-to-peer networks. An agent-based model had been developed to solve the problems caused by the blind flooding-based search, such as inefficient search and traffic, etc. Reinforcement learning and evolution of the agents can change the agents population toward sampling areas of the environment that are close to resources by increasing the density of agents near those sources.*

***Keywords**: Peer-to-Peer Search, Agent-based System, Modeling and Simulation, Reinforcement learning.*

***ACM Classification Keywords**: C.2.6 Internet working*

1 Introduction

A peer-to-peer (P2P) network is distributed systems based on the concept of resource sharing by direct exchange between peer nodes (i.e., nodes having same role and responsibility).. Exchanged resources include content, as in popular P2P files are end systems in the Internet and maintain information about a set of other nodes (called neighbors) in the P2P layer. These nodes form a virtual overlay network on top of the Internet. Each link in a P2P overlay corresponds to a sequence of physical links in the underlying network.

Early search mechanisms primarily used flooding or k-random walk [Hoile, 2002] algorithms. In the flooding approach, each node propagates the query to all its neighbors. On a receipt of a query, the node searches in its local repository. If the object is found, it informs the query originator and further search in that path terminate. If not, the node further forwards the query to all its neighbors.

The flooding method generates a large amount of network traffic. To overcome this problem, random walk algorithms are often used. In Random Walks, the requesting node sends out k query messages to an equal number of randomly chosen neighbors. Each of these messages follows its own path, having intermediate nodes forward it to a randomly chosen neighbor at each step. These queries are also known as walkers. Random Walkers cannot learn anything from its previous successes or failures, displaying high variability in all ranges of requests.

The environments in which P2P applications are deployed exhibit extreme dynamism in structure and load. In order to deal with the scale and dynamism that characterize P2P systems, a paradigm shift is required that includes self-organization, adaptation and resilience as fundamental properties. Complex adaptive systems (CAS) commonly used to explain the behavior of certain ecological and social systems can be the basis of a new programming paradigm for P2P applications. Here we are concerned with the development of an ecosystem-inspired approach to the design of agent systems for searching in Unstructured Peer-to-Peer Networks.

Ecosystems have been a source of inspiration to a number of previous developers of agent systems [Lv, 2002, Moukas, 1997]. Moukas [Moukas, 1997] employs ecosystem inspired ideas in the Amalthaea architecture for information filtering. Paul Marrow and colleagues [Lv, 2002] have advocated an “information ecosystems” approach to support a variety of information management applications.

This paper developed an ecology-based model for managing a number of search agents on the P2P networks that can provide decentralized distributed robust control of agents in the dynamic P2P network environments. This kind of agents does provide a viable means of performing network resource discovery, which makes P2P more practicable. This paper is organized as follows: Sec.2 defines the model and states the attributions of agents; a detailed description of all aspects related to the EIDS algorithm. Sec.3, next to showing the performance of the algorithm in comparison to other approaches, presents experimental results and analysis. Sec.4 provides a discussion of related work on agent based model used for searching in peer-to-peer networks.

2 System Model

This section is divided into two parts. In the first part (Section 2.1) we describe the framework chosen to model the P2P environment. In the second part (Section2.2) we describe the ecologically inspired search algorithm.

2.1 Environment Description

The model framework involves three concepts: peers, neighbors and search agents. The peers are typically computing devices that can maintain some state, or perform computations. A peer has a set of neighbors and is able to send search agents to its neighbors only.

The factors which are important for simulating P2P environments are the overlay topology. The overlay topology consists of a two-dimensional grid responsible for maintaining the neighborhood connections between the peers in the P2P network. Due to the grid structure, each peer residing in a particular node has a fixed set of eight neighbors. Each grid is conceived to be containing a heterogeneous distribution of peers of the P2P network with one or more distinct resources. We assume that there are 1024 unique resources; each of them can be represented by a 10-bit binary token as the resource information (RI). All agents take with also a 10-bit binary token as their searching information (SI) when they move across the environment. Similarity between a RI and a SI is measured by the number of bits that are identical. That is, sim(RI, SI) = d – HD(RI, SI), where HD is the Hamming distance between RI and SI. Zipf's distribution [Zipf, 1935] is chosen to distribute each of the 1024 unique resources in the network. These resources are gathered and consumed by the agent to survive. Each peer may be visited by at most one agent at any time. At any time during the evolution of the model, each agent has a distinct location on the environment, characterized by the peer the agent visits, and a distinct field of view, measured in grids. Each agent has perfect knowledge of resource levels within its field of view; an agent has no knowledge or memory of resource levels outside its field of view. All the agents have a fixed time to live. In addition, an agent has a characteristic metabolic rate for the resource it finds and consumes. Any resources found by an agent can be retained as energy without constraint. If an agent's energy diminishes to zero, the agent dies from starvation.

2.2 The EIDS Algorithm

The ecologically inspired search is a distributed algorithm in which queries are represented as agents. The agents are created at the peer who issued the query and travel over the P2P network in which peers are arranged in a grid-like topology, as in the Swarm simulator [Kaelbling, 1996]. At random times, each agent makes a random number of hops along the P2P network.

The search in our P2P network is initiated from the user peer. The user emanates search agents (message packets) to its neighbors-the packets are thereby forwarded to the surroundings. The method of spreading search agents forms the basis of the algorithm.

Ecosystems usually have several attractive qualities (such as dynamic decentralized control, self regulation, no single point of failure, robustness, and stability) that we require for P2P system. We propose a solution to the problem of controlling the number of agents appropriate for a search which is inspired by large ecosystems (Figure.1):

1. Each kind of resources (Resources could be files in a file-sharing system or CPU cycles in a computational grid.) in P2P system will be associated with energy.

2. Agents finding a resource successfully will collect the energy associated with the resource.

3. Agents consume energy over time to sustain their existence.

4. Agents that exceed the time to live or exhaust their supply of energy die.

5. Abundance of energy can cause a new agent to spawn.

All search operations are controlled by a set of micro-scale rules. The spatial distribution of resources are searched, the energy associated with the resources are gathered and consumed by the agents to survive.

An agent has two dynamic attributes behavioral attributes:

- The current energy level (EL): The current energy level of the agent. If this falls below zero the agent will die.
- The agent's age (AA): Measured in number of hops.

In addition, an agent has a number of static attributes that do not change during their lifetime:

- Search Profile (SP): Built from the informational interest of the peer by which the agent is made.
- Metabolic Rate (MR): The amount of energy it consumes during each hop.
- Energy received at Birth (EB): The energy the agent is born with.
- Energy need for reproduce (ER): The energy the agent needs to attain before it can reproduce.
- Time to Live (TTL): The agent's maximum possible age. It represents the maximum hop-distance a search agent can reach before it gets discarded.

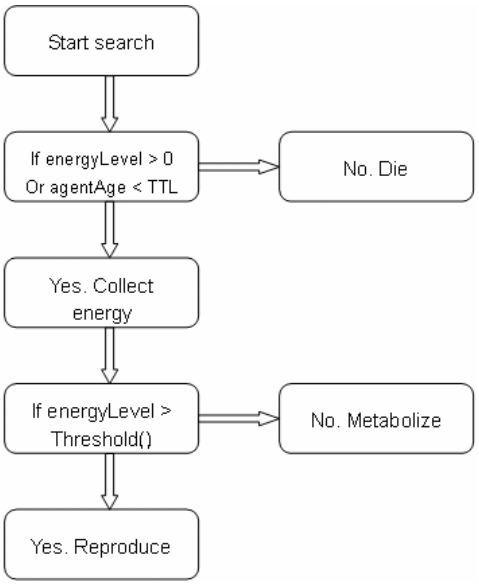


Figure.1 Flowchart of agent behavior

As time evolves, four micro-scale behavioral rules control the search agents-*Query Start Rule (QSR), Resource Search Rule (RSR), Reproduce Rule (RR)* and *Die Rule (DR)*. These rules are explained as following:

Query Rule: A query is initiated by a randomly selected peer who requests for some kinds of resources. To obtain an answer to the request, agents are generated by the peer and flooded in its neighboring peers. The **Query Start Rule** is elaborated below.

Rule 1 *QR: Generate Search Agent (*SA) /*Agents are generated by the peer in response to user requests. */

The peer emanates search agent (SA) to its neighbors.

Search Rule: Once the search agents are emanated, they hop from peer to peer in subsequent time steps. Whenever a search agent moves to a peer, it checks whether the peer has earlier been visited or not. If not, then the agent moves to the peer. In this connection,

each peer maintains a field named visit (V), a field named resource profile (RP) and a field named new energy (NE). A successful search is reported if the required resource can be found in this peer. A flag will be set to true to indicate a successful search. An algorithmic form for the resource search rule (RPR) is presented as follow.

Rule 2 *SR: If (Search agent (SA)) Start*

```
AA++;
V++;
    If ((V = = 1) AND (SP = RP) /*Report a match, V = 1 indicates first time visit by an agent.*/
flag = true;
EL = EL + NE;
Update;
```

Reproduce Rule: Once the current energy level of the search agent exceeds a threshold, the agent will spawn a new one and splitting its current energy in half.

Rule 3 *RR: If (Search agent (SA)) Start*

*If (EL >= ER) /*The agent get enough energy for reproduce.*/*

Produce a new agent;

EL = EL/2;

Die Rule: When the agent's age reach the time to live (TTL) value, or the current energy level of the agent falls below zero, the agent will die.

Rule 4 *DR: If (Search agent (SA)) Start*

If ((AA >= TTL) OR (EL <= 0))

The agent die;

2.3 Performance Metrics

In this paper we focus on efficiency aspects of the processes solely, and use the following simple metrics in our abstract p2p networks.

- Hit rate is defined as the number of resources found for each agent within a given period of time.
- Average number of hops per successful query. This parameter depends on the topology of P2P network as well as on how effectively search mechanism uses it. The less hops is required in average to find requested data, the less traffic is generated and the less time is required for search.
- Population of agents. The number of agents needs to find the requested resource according to the distribution of resource in the uncertain network environments.

Based upon the above mentioned model and metric definition, we now present the experimental results. Simulation runs on Pentium 2.3 GHz with 1GB RAM under windows XP.

3 Simulation Result and Analysis

The experimental results illustrate the efficiency of the algorithms and the effect of controlling the number of agents dynamically based on ecologically inspired control mechanism. This mechanism is completely distributed, executed locally and uses only locally available information. Thus, no globally available information is required. The emergent behavior resulting from the individual localized control decisions will yield an optimal, or sufficiently optimal, solution at the global level. For comparison, we also simulate experiments with k-random walk. The time-step experiment is elaborated next

3.1 Search Efficiency

The search is initiated by a randomly selected node and the number of resource found each time-step from the commencement of the search is calculated. The value of resource hit rate provides the indication of search efficiency. Figure. 2 shows the result of running the time-step experiment for 20 generations (1 generation is 100 time-steps). The time-to-live parameter is set to 25, and k is set to 12, grid size is 100×100.

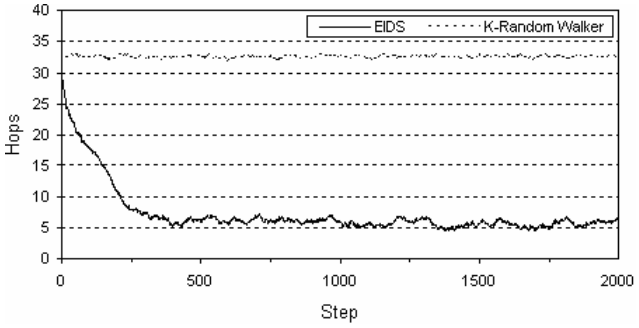


Figure.2 Average number of hops per successful query

We expect that if an agent employs spawn strategy, the total number of agents would increase, the hops required to complete a search would decrease. When EIDS and K-Random Walk are compared, EIDS requires much fewer hops (6 hops) than K-Random Walk (32 hops). In fact, K-Random Walk constantly requires a large number of hops.

The decision for agents spawn strategy is based on a parameter $\rho \in [0, 1]$. For example, if parameter is set to 0.8, each agent will employ the spawn strategy in 80% of the cases. If the parameter is set to 1, then the agent will spawn a new agent whenever its energy level is above a threshold; when the parameter is set to 0, the agent will never spawn any new one, in the fact, it employs the k-random walk strategy.

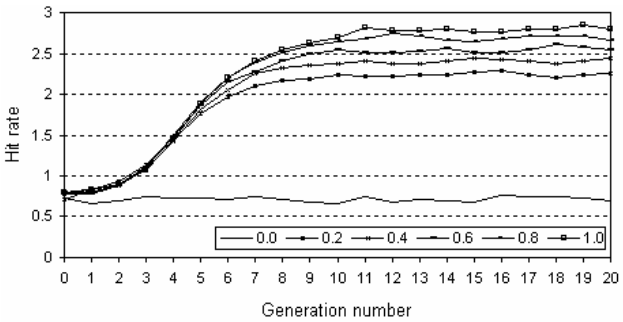


Figure.3 Hit rate per generation

The hit rate is dependent on parameter not only in the start-up phase, but also in the converged phase (Figure.3). The more search agents in the network, the higher the hit rate. The best result can be reached when setting $\rho = 1$. After ten generation, 2.8 resources on average are found in this case. All the five curves employing spawn strategy converge to the same limit over time. The worst result is obtained when setting $\rho = 0$, which is the k-random walk case, the performance stays constant with on average 0.6 resources are found. We see that search efficiency of spawn agents is almost 3-4 times higher than that of k-random walkers.

3.2 Ecosystem Inspired Control of Agent's Population

The system of search agents and the environment they inhabit, i.e., the P2P network, consist of an information ecosystem. All peers can be seen as both information producers and consumers; as consumers, peers send agents (queries) for searching information resources they required, these living agents survive in the context of limited information resources they can find in the network environment. Agent population is determined by the resources of P2P network, the "carry capacity" of the networked information environment, that is, by the size of the relevant set for the given query (Figure.4).

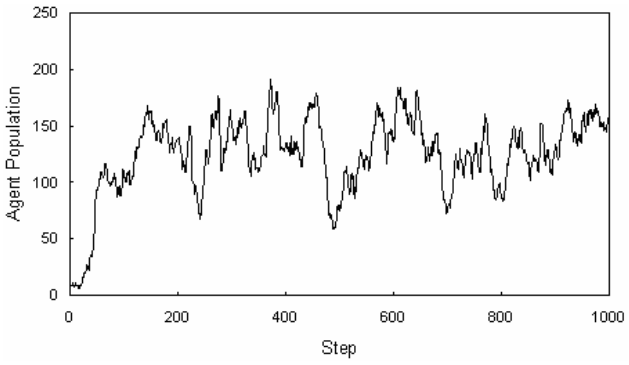


Figure.4 Agent population over time
(TTL=6, Size= 30×30)

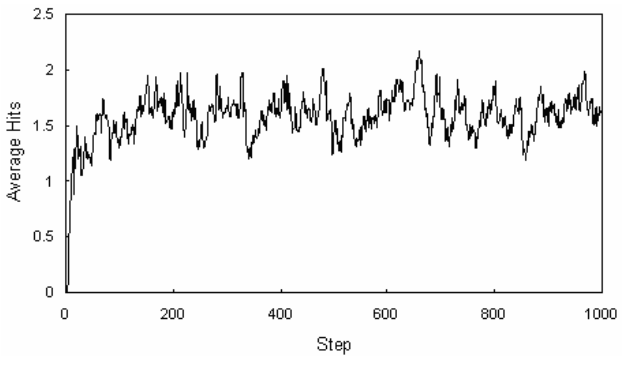


Figure.5 Average hits over time
(TTL=6, Size= 30×30)

As the number of agents in the system grows at the beginning of simulation, the resource they find increase rapidly; when the resource the agents can find reach the max mount of resource, agent population fluctuates

within a definite scope, that's to say the population reach the carry capacity of network environment, and the hit rate maintain a steady average change (Figure.5).

3.3 Reinforcement Learning

Resources from the environment that are correlated with an agent's performance can be used as reward or penalty signals to adjust the agent's behavior during its life [6]. This is the basis of the reinforcement learning framework. Such signal corresponds to the energy change computed in Rule 2 of the algorithm. It could be computed as the time derivative of the agent's energy level.

The energy level throughout the lifetime of the agent is depicted in Figure.6. The agent reproduces around time 34, giving half of its energy to the offspring. Finally the agent runs out of energy and dies shortly after time 96. The Curve 1 plots the level of accumulated energy as a function of time, resulting from the instantaneous changes in energy plotted by the Curve 2. The selection threshold is ER = 10. With the exception of the reproduction event, the energy level is the integral of the Curve 1.

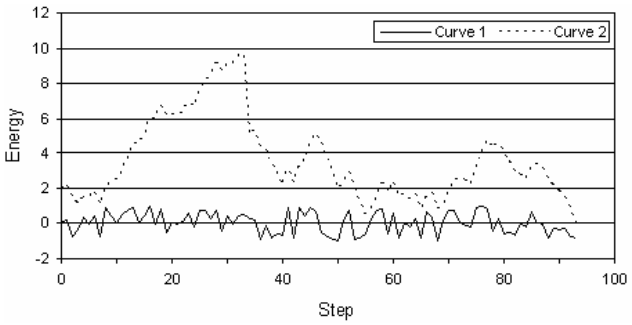


Figure.6 Illustration of energy dynamics in a typical agent's lifetime

Any reinforcement learning scheme can be used to adjust the behavior of the agent so that actions perceived to lead to rewards are reinforced, and action perceived to lead to penalties are discouraged. In this case, the prevailing penalties incurred between time 80 and 95 might have warranted changes leading to a delayed death; the reward incurred around time 34 leads to a new agent offspring, so as to find more peers with similar information resources around the neighborhood.

At the individual level, reinforcement learning biases an agent's behavior toward actions leading to a better knowledge of the environment and to increased payoff, or longer survival. At the collective level, reinforcement evolution biases the population toward sampling areas of the environment that are close to known resources, by increasing the density of agents near those sources.

4 Related Work

The majority of ecology-inspired systems are used to answer some question about real world ecosystems and its properties. Agent-based models are often used to simulating ecosystem. The most well known example is LAGER [Olson, 1995]. The term information ecosystem is used by analogy with natural ecosystems. Information ecosystems ideas can improve our understanding of information infrastructures. An example of agent based systems relevant to information ecosystems research is the InfoSpiders system developed by Menczer and Monge [Menczer, 1999]. This system implements a scalable information search algorithm by use of cooperative agents, drawing explicitly on ecological metaphors. By contrast, the Hive system [Minar, 1999] uses distributed agents to link networked resources on a local network. The MATS system developed by Ghanea-Hercock et al. [Hercock, 1999] uses mobile agents, in this case inspired by social insects, to control a distributed processing application over a network.

There are many agent based algorithms for searching in peer-to-peer networks. The famous Anthill [Babaoglu, 2002] is an open source framework for the design, implementation, and evaluation of ant algorithms in peer-to-peer networks. An Anthill system is an overlay network of interconnected nests (peers). Nests provide services

like document storage, routing table management, topology management, and generation of ants upon user requests. In addition, the Anthill API provides a basic set of actions for ants that enables them to travel from nest to nest, and to interact with the services provided by nests. The ant algorithm is not specified by Anthill, but must be designed by the user of the framework according to the application scenario. In our approach, not only those query agents behave well be rewarded to undergo offspring, but also those behave bad be punished to death. So the number of queries can be maintained on a lower level.

5 Conclusions

In this paper, we have concentrated on developing an agent-based model for controlling query messages that are represented as agent; a search algorithm which derives its inspiration from natural ecosystem is presented. Experiment results above show that this ecologically inspired algorithm is much more efficient search method than k-walker random walk. Each additional step in the search increases the number of nodes visited by only a constant. So exponentially increased over load on each visited node by flooding can be avoided. The basic strengths displayed by the EIDS algorithm need to be further explored and developed, by applying it in more realistic circumstances in the near future.

Bibliography

[Hoile, 2002] C.Hoile, F.Wang, E.Bonsma, and P. Marrow. Core specification and experiments in diet: a decentralised ecosystem-inspired mobile agent system. In Proc. of AAMAS 2002, pp. 623-630.

[Lv, 2002] Q.Lv, P.Cao, E.Cohen, and S.Shenker. Search and replication in unstructured peer-to-peer networks. In Proceedings of the 16th ACM Conference on Supercomputing, 2002.

[Moukas, 1997] A.Moukas, Amalthaea: Information Discover and Filtering using a Multiagent Evolving Ecosystem. Proc. Conf. Practical Applications of Agents and MultiAgent Technology, 1997.

[Zipf, 1935] G. K. Zipf. Psycho-Biology of Languages. Houghton-Mfflin, 1935.

[Kaelbling, 1996] P. Kaelbling, L., Littman, M.L., and Moore, A.W. Reinforcement learning: A survey. Journal of Artificial Intelligence Research, 1996, pp. 237-285.

[Olson, 1995] R. L. Olson, A. A. Sequeira. An emergent computational approach to the study of ecosystem dynamics. Ecological Modeling, 1995, pp. 95-120.

[Menczer, 1999] F. Menczer, and A. E. Monge. Scalable web search by adaptive online agents: an InfoSpiders case study. In: Intelligent Information Agents, M. Klusch, (ed.), Springer, Berlin, 1999.

[Minar, 1999] N. Minar, M. Gray, O. Roup, R. Krikorian, and P. Maes. Hive: distributed agents for networking things. In: Proceedings of ASA/MA '99, 1999.

[Hercock, 1999] R.G.Hercock, J.C. Collis, and D.T. Ndumu. Co-operating mobile agents for distributed parallel processing. In: Proceedings of Autonomous Agents 1999, Seattle, 1999.

[Babaoglu, 2002] O.Babaoglu, H.Meling, and A.Montresor, “Anthill: A Framework for the Development of Agent-Based Peer-to-Peer Systems,” in Proceedings of the 22nd International Conference on Distributed Computing Systems (ICDCS 02). July 2002, IEEE.

Authors' Information

Li Sa *- College of Information Sciences and Technology, Donghua University, Shanghai, P.R.China; e-mail: sali@ustc.edu*

Yongsheng Ding *- College of Information Sciences and Technology, Donghua University, Shanghai, P.R.China; e-mail: ysding@dhu.edu.cn*

INFLUENCE OF SOME USERS' BEHAVIOUR PARAMETERS OVER NETWORK REDIMENSIONING

Emiliya Saranova

***Abstract**: The aim of this paper is to present method of redimensioning of the network capacity (number of equivalent internal switching lines) and some of designed parameters based on consideration of detailed users behaviour and demanded QoS parameters in overall telecommunication network. We have researched dependency and sensitivity of some important parameters in the telecommunication networks redimensioning task. Comparative analyze is made and graphically shown.*

The described approach is applicable directly for every (virtual) circuit switching telecommunication system, viz.: in wireline and for wireless systems (GSM, PSTN, ISDN and BISDN). For packet - switching networks at various layers proposed approach may be used as a comparison basis and when they work in circuit switching mode (e.g. VoIP).

1. Introduction

Traffic Theory enables network designers to make assumptions about their networks based on past experience. Traffic engineering addresses service issues by definition of grade of service (GoS) parameters. A properly engineered network reflects the trade - off between low blocking and high circuit utilization, which means trade - off between service volume and the costs.

On an ongoing basis, a process of redimensioning (sometimes called "servicing") is used to ensure maximum utilization of the existing equipment and determine appropriate reallocation when service demand changes before additional equipment can be installed.

There are many different factors that we need to take into account when analyzing traffic. QoS parameters are administratively specified in Service Level Agreement (SLA) between users and operators. These QoS parameters are reflecting on GoS parameters [5].

Aim: Based on the sensitivity of designed parameters (for example according blocking probability, repeated calls and others) taking into account users' behaviour an approach for network redimensioning to be created with purpose to be maintained the administrative contractual level of QoS (in SLA) under fixed conditions.

For proposed conceptual and its corresponding analytical model [6] a network redimensioning task (NRDT) is formulated, solvability of the NRDT and the necessary conditions for analytical solution are researched as well. A system of equations based on the conceptual model and dependencies between parameters of the researched telecommunication system, is derived. Dependencies, based on the numerical - analytical results are shown graphically.

In this paper we consider model of telecommunication system with channel switching, in stationary state, with Poisson input flow, repeated calls, limited number of homogeneous terminals and losses due to abandoned and interrupted dialing, blocked and interrupted switching, not available intent terminal, blocked and abandoned ringing and abandoned conversation.

2. Conceptual model

The conceptual model [6] of the telecommunication system includes the paths of the calls, generated from (and occupying) the A-terminals in the proposed network traffic model and its environment (shown on *Fig*. 1).

The names of the devices are constructed according to their position in the model.

2.1. The comprising virtual devices

The following important virtual devices on *Fig.*1 are shown and considered:

a = comprises all the A-terminals (calling) in the system (shown with continuous line box).

b = comprises all the B-terminals (called) in the system (box with dashed line).

ab = comprises all the terminals (calling and called) in the system (not shown on *Fig.*1);

s = virtual device corresponding to the switching system. It is shown with dashed line box into the a - device. *Ns* stand for the capacity (number of equivalent internal switching lines) of the switching system.

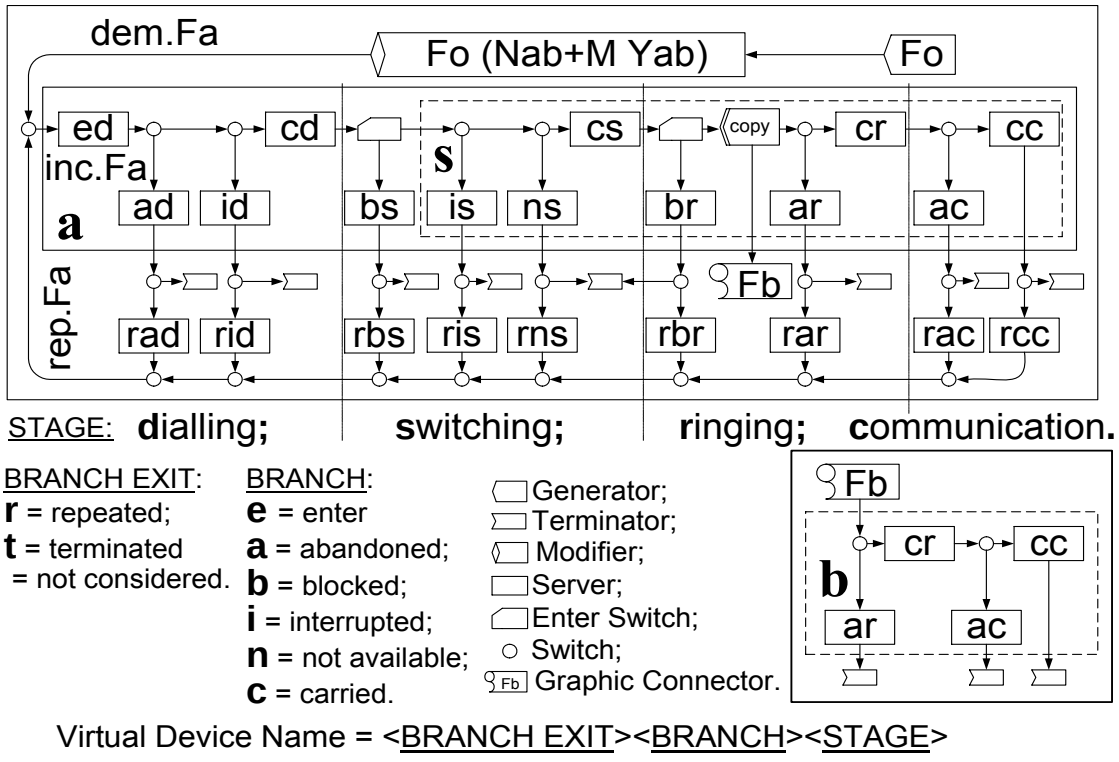


Fig. 1*. Normalized conceptual model of the telecommunication system and its environment and the paths of the calls, occupying A-terminals (a - device), switching system (s - device) and B-terminals (b - device); base virtual device types, with their names and graphic notation.*

2.2. Stages and branches in the conceptual model:

Service *stages*: **d**ialing, **s**witching, **r**inging and **c**ommunication.

Every service stage has *branches*: **e**nter, **a**bandoned, **b**locked, **i**nterrupted, **n**ot available, **c**arried (correspondingly to the modeled possible cases of ends of the calls' service in the branch considered).

Every branch has two *exits*: **r**epeated, **t**erminated (which show what happens with the calls after they leave the telecommunication system). Users may make a new bid (repeated call), or to stop attempts (terminated call).

2.3. Parameters and its notations in the conceptual model:

Letter *F* stands for calling rate (frequency) of the flow [calls/sec.], *P* = probability for directing the calls of the external flow to the device considered, *T* = mean service time, in the device [sec.], *Y* = intensity of the device traffic [Erl], *N* = number of service places (lines, servers) in the virtual device (capacity of the device). In the normalized models [10], used in this paper, every base virtual device, except the switch, has no more than one entrance and/or one exit. Switches have one entrance and two exits. For characterizing the intensity of the flow,

we are using the following notation: *inc.F* for incoming flow, *dem.F*, *ofr.F* and *rep.F* for demand, offered and repeated flows respectively [5]. The same characterization is used for traffic intensity (*Y*).

Fo is the intent intensity of calls of one idle terminal; *inc.Fa* = *Fa* is intensity of incoming flow, characterizing the flow of demand calls (*dem.Fa*).

For creating a simple analytical model, we make the system of fourteen (A-1 – A-14) assumptions made in [6].

3. Analytical model

3.1. Some general equations

For the proposed conceptual model we have derived the following system of equations [6]:

$$Yab = Fa[S_1 - S_2(1 - Pbs)\ Pbr - S_3\ Pbs] \tag{3.1.1}$$

$$Fa = dem.Fa + rep.Fa \tag{3.1.2}$$

$$dem.Fa = Fo\ Nab \tag{3.1.3}$$

$$rep.Fa = Fa\ [R_1 + R_2\ Pbr\ (1 - Pbs) + R_3\ Pbs] \tag{3.1.4}$$

$$Pbr = \begin{cases} \dfrac{Yab\text{-}1}{Nab\text{-}1} & in\ \ case\ of\ \ \ 1 \le Yab \le Nab, \\ 0 & in\ case\ of\ \ \ 0 \le Yab < 1. \end{cases} \tag{3.1.5}$$

$$Ts = S_{1Z} - S_{2Z}\ Pbr \tag{3.1.6}$$

$$ofr.Fs = Fa\ (1\text{-}\ Pad)(1\text{-}\ Pid) \tag{3.1.7}$$

$$ofr.Ys = ofr.Fs\ \ Ts \tag{3.1.8}$$

$$Pbs = Erl\_b\ (\ Ns, ofr.Ys\ ) = \frac{\dfrac{(ofr.Ys)^{Ns}}{Ns!}}{\sum_{j=0}^{Ns}\dfrac{(ofr.Ys)^j}{j!}} \tag{3.1.9}$$

$$crr.Ys = (1 - Pbs)\ ofr.Ys \tag{3.1.10}$$

The following notations are used:

$$S_1 = Ted + Pad\ Tad + (1 - Pad)[Pid\ Tid + (1 - Pid)[Tcd + Pis\ Tis + (1 - Pis)[Pns\ Tns + (1 - Pns)[Tcs + 2\ Tb]]]] \tag{3.1.11}$$

$$S_2 = (1 - Pad)(1 - Pid)(1 - Pis)(1 - Pns)[2\ Tb - Tbr] \tag{3.1.12}$$

$$S_3 = (1 - Pad)(1 - Pid)[PisTis - Tbs + (1 - Pis)[Pns\ Tns + (1 - Pns)[Tcs + 2\ Tb]]] \tag{3.1.13}$$

$$S_{1z} = Pis\ Tis + (1 - Pis)[PnsTns + (1 - Pns)(Tb + Tcs)] \tag{3.1.14}$$

$$S_{2z} = (1 - Pis)(1 - Pns)(Tb - Tbr) \tag{3.1.15}$$

$$R_1 = Pad\ \mathrm{Pr}\ ad + (1 - Pad)(Pid\ \mathrm{Pr}\ id + (1 - Pid)[Pis\ \mathrm{Pr}\ is + (1 - Pis)(Pns\ \mathrm{Pr}\ ns + (1 - Pns))Q]) \tag{3.1.16}$$

$$R_2 = (1 - Pad)(1 - Pid)(1 - Pis)(1 - Pns)(\mathrm{Pr}\ br - Q) \tag{3.1.17}$$

$$R_3 = (1 - Pad)(1 - Pid)\{\mathrm{Pr}\ bs - [Pis\ \mathrm{Pr}\ is + (1 - Pis)[Pns\ \mathrm{Pr}\ ns + (1 - Pns)Q]]\} \tag{3.1.18}$$

$$Q = Par\ Prar + (1 - Par)[Pac\ Prac + (1 - Pac)\ Prcc] \tag{3.1.19}$$

$$K = Pis\ \mathrm{Pr}\ is + (1 - Pis)[Pns\ \mathrm{Pr}\ ns + (1 - Pns)Q] \tag{3.1.20}$$

Note 1: The demand rate of calls of one idle terminal is $Fo \ge 0$.

Note 2: In (3.1.9) $Pbs = Erl\_b\ (\ Ns, ofr.Ys\ )$ is the probability of blocking due to insufficient number of lines.

3.2. The researched GoS parameters

Based on the ITU definition of blocking probability (like a GoS parameter) [5] we consider the following two GoS parameters: probability of finding B- terminal busy (*Pbr*) and probability of blocking switching (*Pbs*). The target value (*adm.Pbs*) of probability of blocking switching is administratively determined in advance in SLA (Service Level Agreement).

4. Network Redimensioning Task

4.1. Formulation of a network redimensioning task (NRDT):

Based on previous experience, to determine the volume of telecommunication resources that is enough for serving given input flow of demands with prescribed characteristics of QoS, is one of the main problems that often have to be solved by operators. It includes the following tasks:

1. To be redimensioned a network means to be found of number of necessity internal switching lines, when in advance level of QoS is administratively determined and the values of known parameters are measured and/or calculated in case of already operating network.

2. To be found the values of the designed parameters, describing the designed system state. For example, a system parameter, describing offered traffic intensity of the switching system (*dsn.ofr.Ys*), designed probability to find B terminal "busy" (*dsn.Pbr*), etc…

3. To be researched sensitivity and dependency of designed parameters regarding QoS – parameters and users' behaviour parameters in telecommunication system.

Parameters and aims in the Network Redimensioning Task (NRDT):

Given parameters:

Administrative determined parameters: *adm.Pbs, Nab = adm.Nab* (4.1.1)

Parameters with empirical values: $Fo,\ S_1,\ S_2,\ S_3,\ R_1,\ R_2,\ R_3,\ S_{1z},\ S_{2z},\ Tb$ (4.1.2)

Aim: To determine the number of equivalent internal switching lines *Ns*; and the values of following

Designed unknown parameters: *dsn.Pbr, dsn.Fa, dsn.rep.Fa* (4.1.3)

Condition: *dsn.Pbs* ≤ *adm. Pbs* (4.1.4)

4.2. Analytical solution of the NRDT:

In [8] is shown that the smaller root of eq. (4.2.1) fulfills $Pbr \in (0;1)$. When *Pbr* is determined thereby, we say that *Pbr* is determined on the base of the NRDT and we denote *dsn.Pbr*. In Theorem 1, the conditions for existence of *dsn.Pbr* in NRDT are researched and proved in [10].

Theorem 1: If $adm.Pbs \neq \frac{S_1 - S_2}{S_3 - S_2}$ then solution *Pbr* of equation (4.2.5) can be

value of *dsn.Pbr* in the NRDT

$$A\,Pbr^2 + B\,Pbr + C = 0, \tag{4.2.1}$$

where $A = R_2(1 - adm.Pbs)(Nab - 1)$

$$B = (1 - adm.Pbs)(R_2 - Fo\,S_2) - (1 - R_1 - R_3 adm.Pbs)(Nab - 1)Nab$$

$$C = Fo\,(S_1 - S_3\,adm.Pbs) - Nab\,(1 - R_1 + R_3\,adm.Pbs).$$

Proof: From (3.1.1), (3.1.2), (3.1.4) and (3.1.5) follow (4.2.1). The smaller root of eq.(4.2.1) $Pbr \in (0;1)$ is $dsn.Pbr$ in the NRDT [10].

Theorem 2: In the NRDT if $dsn.Pbr \neq \frac{S_1 - S_3\, adm.Pbs}{S_2\,(1 - adm.Pbs)}$

then
$$dsn.Fa = \frac{1 + dsn.Pbr(Nab - 1)}{S_1 - S_2(1 - adm.Pbs)\,dsn.Pbr - S_3\, adm.Pbs} \quad (4.2.2)$$

Proof: From (3.1.8), (3.1.6) and (3.1.7) follows (4.2.2).

Consequence 1: In the NRDT

$$dsn.rep.Fa = \frac{(1 + dsn.Pbr(Nab - 1))(R_1 - R_2(1 - adm.Pbs)\,dsn.Pbr - R_3 adm.Pbs))}{S_1 - S_2(1 - adm.Pbs)\,dsn.Pbr - S_3\, adm.Pbs} \quad (4.2.3)$$

Proof: Using (3.1.4), (3.1.5), (3.1.1) and Theorem 1 follows (4.2.3).

Consequence 2: In the NRDT

$$\frac{dsn.rep.Fa}{dsn.Fa} = R_1 - R_2(1 - adm.Pbs)\,dsn.Pbr - R_3 adm.Pbs \quad (4.2.4)$$

Proof: Using (3.1.4) and Theorem 1 follows (4.2.4).

There are many different traffic models used for traffic dimensioning/redimensioning.

We use Erlangs' B formulae and its recursion form because this model suits well the observed telecommunication networks.

This formulae may be used for redimensioning when the equivalent offered traffic is evaluated on the basis of [2] and the level of QoS is determined administratively in advance (for example blocking probability *adm.Pbs*). This is proved in Theorem 3 [8].

Theorem 3: There is only one solution in the NRDT through the equation

Erl_b (*Ns,ofr.Ys*) = *adm.Pbs* (4.2.5)

regarding to the number of switching lines *Ns*.

Adm.Pbs $\in$(0; 1] is in advance administratively determined value of blocking probability, providing of GoS.

In [6] is proved that only one solution of *Ns* exists, fulfilling the equation (4.2.5) and corresponding to the determined administratively in advance value of the blocking probability *adm.Pbs* $\in$(0; 1].

The number of internal switching lines *Ns* and the values of *dsn.ofr.Ys* are calculated on the conditions of the Theorem 1, Theorem 2 and Theorem 3 [10].

4.3. Research of Sensitivity and dependency of parameters in NRDT. Numerical results and conclusions

4.3.1. Sensitivity – used definitions.

As mathematical approach are used partial derivatives of researched parameters [1] and follow definitions:

Let $P(x_1, x_2, ..., x_n)$ is tuple (ordered set) consists of variables $x_1, x_2, ..., x_n$.

Tuple P_0 of empirical or evaluated parameters' values is $P_0(x_1^0, x_2^0, ..., x_n^0)$, where the parameters' values are $x_1^0, x_2^0, ..., x_n^0$.

Let parameter *A* depends on tuple *P* then $A(P) = f(x_1, x_2, ..., x_n)$ where $x_1 \in D_1$, $x_2 \in D_2 ..... x_n \in D_n$. The value of *A* in P_0 is $A(P_0) = f(x_1^0, x_2^0, ..., x_n^0)$.

Sensitivity of parameter *A* regarding x_i, $i \in [1; n]$ in P_0 is $sen(A \mid x_i)\big|_{P_0} = \left.\frac{\partial A}{\partial x_i}\right|_{P_0}$. (4.3.1)

The *sensitivity* of parameter *A* according parameter x_i in P_0 can be indirect evaluated numerically through (4.3.2) as well

$$sen(A \mid x)\big|_{P_0} = \left.\frac{\partial A}{\partial w}\frac{\partial w}{\partial x}\right|_{P_0}, \quad sen(A \mid x)\big|_{P_0} = \left.\frac{\partial A}{\partial w}\frac{\Delta w}{\Delta x}\right|_{P_0} \text{ or } sen(A \mid x)\big|_{P_0} = \left.\frac{\Delta A}{\Delta w}\frac{\Delta w}{\Delta x}\right|_{P_0} \quad (4.3.2)$$

where ΔA means variation of A concerning variation Δx_i of variable x_i, $i \in [1; n]$.

Range of parameters' values of *A* according parameter *x* is amplitude of it $Max\ A(x_k^s) - Min\ A(x_k^p)$, where $Max\ A(x_k^s)$ and $Min\ A(x_k^p)$ are absolute maximum resp. minimum received in points $x_k^s \in D_k$ and $x_k^p \in D_k$.

$$Ran(A \mid x_k) = Max\ A(x_k^s) - Min\ A(x_k^p) \quad (4.3.3)$$

4.3.2. Numerical results:

Numerical results are received using follows parameters' values:

Ted = 3 sec, *Pad* = 0.09, *Tad* = 5 sec, *Prad* = 0.95, *Pid* = 0.01, *Tid* = 11 sec, *Prid* = 0.2, *Tcd* = 12 sec, *Tbs* = 5 sec, *Pis* = 0.01, *Tis* = 5 sec, *Pris* = 0.01, *Pns* = 0.01, *Tns* = 6 sec, *Prns* = 0.3, *Tbr* = 5 sec, *Par* = 0.65, *Tar* = 45 sec, *Prar* = 0.75, *Tcr* = 10 sec, *Pac* = 0.2, *Tac* = 13 sec, *Prac* = 0.9, *Tcc* = 180 sec, *Prcc* = 0.01, *Tb* = 139.07 sec, *Tcs* = 5 sec, *emp.Fo*=0.0125714, *Nab*=7000 terminals.

4.3.3. Research of parameters' dependency and sensitivity in NRDT regarding GoS parameters (blocking probability) and some human behaviour parameters (probability of generating of repeated calls after blocking). Knowing sensitivity from a parameter, we may estimate the parameters' importance and necessary occupancy of its measurement.

4.3.3.1. Parameters' dependency in NRDT from GoS parameters *Pbs* and *Pbr*

We consider sensitivity of $\frac{rep.Fa}{Fa}$ because *Ns* and *ofr.Ys* is indirect dependent on it.

Numerical results: Dependency of number of equivalent switching lines *Ns* from probabilities for blocking switching *Pbs* or finding B – terminal busy *Pbr*:

When 0 % <*Pbr* < 100 % and *Pbs* = 1% then *Ns* grows with 4.17 % and *rep.Fa* - with 46 %. In case *Pbs* = 99% then *Ns* grows with 1.7 % and *rep.Fa* - with 0.7 %.

Conclusion: When *adm.Pbs* is fixed then we may find *dsn.rep.Fa/Fa* and Ns/Nab, based on *Pbr.*

It is shown graphically on fig.2 and fig.3.

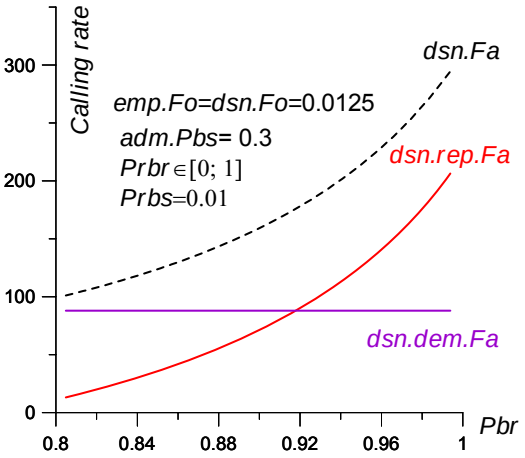


Fig.2. *Dependency of designed calling rate: all call attempts (dsn.Fa), primary (dem.Fa) and repeated (rep.Fa) regarding Pbr.*

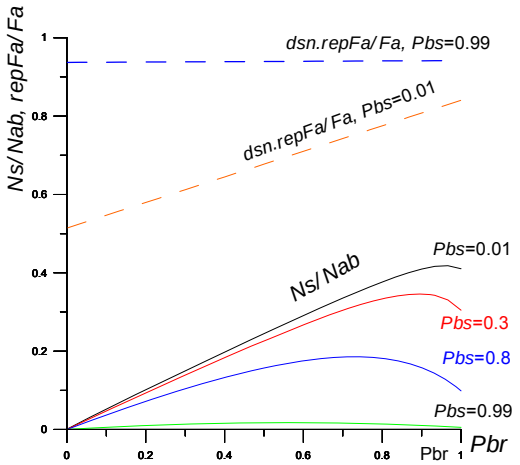


Fig.3. *Dependency of designed rep.Fa/Fa and Ns/Nab when adm.Pbs is respectively 1 %, 30 %, 80 % and 99 %.*

4.3.3.2. Users' behaviour parameters – R_1, R_2 and R_3 in NRDT

We assume that loss probabilities *Pad, Pid, Pis, Pns, Par, Pac, Pcc* are fixed.

From (3.1.16) – (3.1.20) follows that Q, K, R_1, R_2 and R_3 are dependent on human behaviour parameters and represent it in case of repeated calls.

a) Parameter R_1 represents generalized probability of generating repeated calls and is independent both of *Prbr* and *Prbs*. In fact, R_1 is dependent on the loss probability only and is evaluated through (3.1.16).

b) Parameter R_2 is dependent on blocking probability *Prbr* of repeated calls after finding B – terminal busy. R_2 is evaluated through (3.1.17) and is independent of *Prbs*.

Sensitivity of R_2 is $sen(R_2 \mid \Pr br) = \dfrac{\partial R_2}{\partial \Pr br} = k_2 = (1-Pad)(1-Pid)(1-Pis)(1-Pns) \geq 0 \Rightarrow R_2$ is linear function and increases regarding *Prbr.*

From (3.1.19) and when $\Pr br \geq Q$ follows $0 \leq R_2 \leq k_2\,(1-Q) \Rightarrow R_2 \in [0;1]$

c) Parameter R_3 is dependent on *Prbs* and is independent of *Prbr*. R_3 is evaluated through (3.1.18). Sensitivity of R_3 is $sen(R_3 \mid \Pr bs) = \dfrac{\partial R_3}{\partial \Pr bs} = k_3 = (1-Pad)(1-Pid) \geq 0 \Rightarrow R_3$ is linear function and is increases regarding *Prbs.*

From (3.1.20) and when $\Pr bs \geq K$ follows $0 \leq R_3 \leq k_3\,(1-K) \Rightarrow R_3 \in [0;1]$.

Obviously, $sen(R_3 \mid \Pr bs) > sen(R_2 \mid \Pr br)$ in the NRDT and influence of *Prbs* is bigger than of *Prbr* over R_1, R_2 and R_3, resp. over *rep.Fa*, the traffic intensity *Yab*, offered traffic *ofr.Ys*, etc...

Numerical results:

If $\Pr br \in [0;1]$ then $R_2 \in [-0.234; 0.649]$, $Ran(R_2 \mid \Pr br) = 0.883$, when *Prbs*=0.95.

If $\Pr bs \in [0;1]$ then $R_3 \in [-0.231; 0.634]$, $Ran(R_3 \mid \Pr bs) = 0.865$, when *Prbr*=0.8.

4.3.3.3. Sensitivity and dependency of $\frac{rep.Fa}{Fa}$ regarding R_2 and R_3 respectively *Prbr* and *Prbs* in the NRDT.

We consider $\frac{rep.Fa}{Fa}$ concerning (4.2.4) and (3.1.17). Let *adm.Pbs* is fixed. Then $sen\left(\frac{rep.Fa}{Fa} \mid \Pr br\right)\Big|_{adm.Pbs} = (1-adm.Pbs)\,dsn.Pbr\,(1-Pad)(1-Pid)(1-Pis)(1-Pns) \geq 0$.

On analogy, $sen\left(\frac{rep.Fa}{Fa} \mid \Pr bs\right)\Big|_{adm.Pbs} = adm.Pbs\,(1-Pad)(1-Pid) \geq 0$.

Therefore, $\frac{rep.Fa}{Fa}$ regarding *Prbr* and respectively *Prbs* is increasing in NRDT and the influence over $\frac{rep.Fa}{Fa}$ of *Prbs* is bigger than of *Prbr* as well.

In NRDT when *adm.Pbs* is determined in advance then common sensitivity of $\frac{rep.Fa}{Fa}$ would be

$$sen\left(\frac{rep.Fa}{Fa} \mid \Pr bs, \Pr br\right)\Big|_{adm.Pbs} = sen\left(\frac{rep.Fa}{Fa} \mid \Pr bs\right)\Big|_{adm.Pbs} + sen\left(\frac{rep.Fa}{Fa} \mid \Pr br\right)\Big|_{adm.Pbs} \quad (4.3.4)$$

Numerical results: If *adm.Pbs* = 40 %, $\Pr br \in [0;1]$, $\Pr bs \in [0;1]$ and *emp.Fo* = 0.0125 then regarding

$\Pr br \in (0.3;1) \Rightarrow 39\% < \frac{rep.Fa}{Fa} < 64\%$ and $Ran\left(\frac{rep.Fa}{Fa}\middle| \Pr br\right) = 25$ %

$\Pr bs \in (0.3;1) \Rightarrow 40\% < \frac{rep.Fa}{Fa} < 60\%$ and $Ran\left(\frac{rep.Fa}{Fa}\middle| \Pr bs\right) = 20$ %.

Conclusion: The influence over $\frac{rep.Fa}{Fa}$ of *Prbr* is bigger than *Prbs* when $\Pr br \in [0;1]$ and $\Pr bs \in [0;1]$.

4.3.3.4. Sensitivity and dependency and sensitivity of *Ns* regarding R_2 and R_3 respectively *Prbr* and *Prbs*

Based on eq. (3.1.1), (3.1.5), (3.1.4), (3.1.7) and (3.1.8) follows that dependency of *ofr.Ys* resp. *Ns* regarding *Prbs* and *Prbr* is resulted from influence of it over probability *Pbr* of finding B – terminal busy.

$$\begin{aligned} sen\,(Ns \mid \Pr br)\big|_{adm.Pbs} &= \frac{\Delta Ns}{\Delta ofr.Ys}\left(\frac{\partial ofr.Ys}{\partial \Pr br}\right)\Big|_{adm.Pbs} = \frac{\Delta Ns}{\Delta ofr.Ys} sen\,(ofr.Ys \mid \Pr br)\big|_{adm.Pbs} \\ sen\,(Ns \mid \Pr bs)\big|_{adm.Pbs} &= \frac{\Delta Ns}{\Delta ofr.Ys}\left(\frac{\partial ofr.Ys}{\partial \Pr bs}\right)\Big|_{adm.Pbs} = \frac{\Delta Ns}{\Delta ofr.Ys} sen\,(ofr.Ys \mid \Pr bs)\Big|_{adm.Pbs} \end{aligned} \quad (4.3.5)$$

Numerical results: When $adm.Pbs = 0.3$, $emp.Fo = 0.0125$ and regarding $\Pr br \in [0;1]$, $\Pr bs \in [0;1]$ then $\Rightarrow \frac{Ns}{Nab} \in [0.305;0.346]$, $Ran\left(\frac{Ns}{Nab}\right) = 0.041$ and $\Rightarrow \frac{rep.Fa}{Fa} \in [0.13;0.715]$, $Ran\left(\frac{rep.Fa}{Fa}\right) = 0.585$, $dsn.Pbr \in [0.805;0.999]$, $Ran(dsn.Pbr) = 0.195$, $\frac{Ts}{Tb} \in [0.078;0.262]$, $Ran\left(\frac{Ts}{Tb}\right) = 0.184$. It is shown graphically on *Fig.* 4.

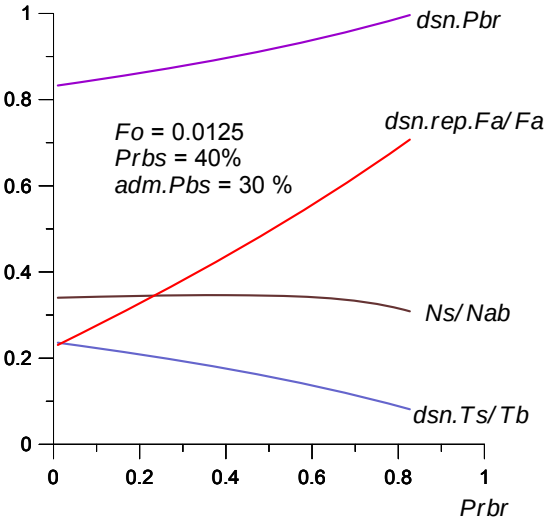


***Fig. 4**. Dependency of dsn.Pbr, dsn.rep.Fa, Ns, dsn.ofr.Ys and Ts/Tb on Prbr. The values of adm.Pbs = 30 %, Nab = 7000 terminals*

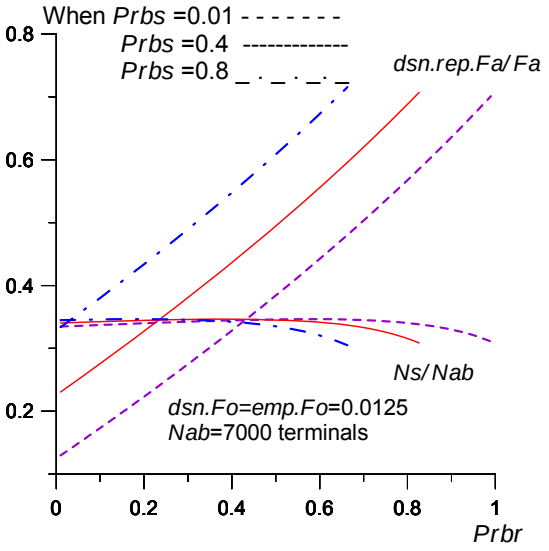


***Fig. 5**. Dependency of designed rep.Fa/Fa, Ns/Nab on Prbr. The values of Prbs are 1 %, 40 % and 80 %, adm.Pbs = 30 %.*

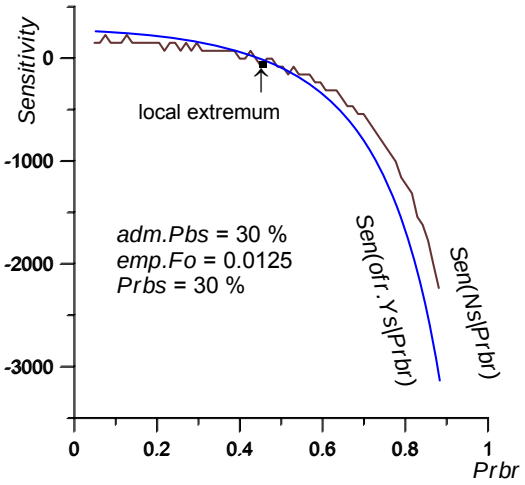


Fig. 6. Sensitivity of the designed number of internal switching lines Ns and dsn.ofr.Ys regarding users' behaviour parameter Prbr.

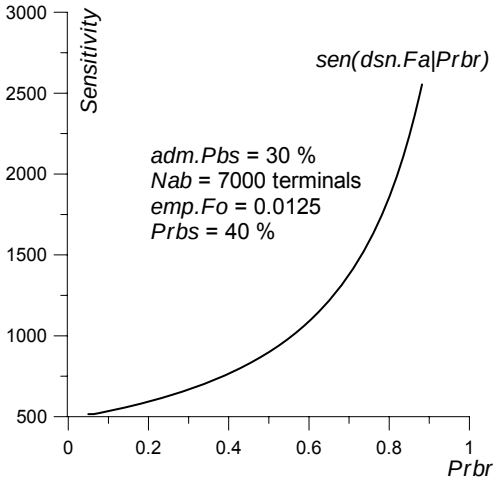


Fig. 7. Sensitivity of the designed calling rate regarding users' behaviour parameter Prbr.

Conclusion: Comprising the ranges of $\frac{Ns}{Nab}$ and $\frac{rep.Fa}{Fa}$ follow that the influence of *Prbr* and *Prbs* over $\frac{rep.Fa}{Fa}$ is bigger than over $\frac{Ns}{Nab}$. Because minimum of *dsn.Pbr* is 80 %, the influence of *Prbr* and *Prbs* over $\frac{rep.Fa}{Fa}$ is big. It is shown graphically on *Fig.* 5.

On ***Fig. 4*** are shown numerical results concerning influence of *Prbr* over designed $\frac{Ns}{Nab}$, $\frac{rep.Fa}{Fa}$, $\frac{ofr.Ys}{Nab}$, $\frac{Ts}{Tb}$ and *dsn.Pbr.*

The parameters' sensitivity research is important for the accuracy received by Network redimensioning.

The sensitivity regarding *Prbr* of number of equivalent internal switching lines *Ns,* designed offered traffic *dsn.ofr.Ys* are graphically shown on *Fig.* 6 and respectively sensitivity of *dsn.Fa* – on *Fig.7*.

5. Algorithm for calculating the unknown values of the parameters in the NRDT:

1. Input data: administrative determined parameters are *adm.Pbs, Nab* = *adm.Nab* and empirical evaluated are $Fo,\ S_1,\ S_2,\ S_3,\ R_1,\ R_2,\ R_3,\ S_{1z},\ S_{2z},\ Tb$.
2. Based on Theorem 1, we may evaluate *dsn.Pbr* always:

$$\forall\, adm.Pbs \Rightarrow \exists\, dsn.Pbr \qquad (5.1.1)$$

Dsn.Pbr is the smaller root of equation (4.2.5) and belongs to $(0;1)$. We use *dsn.ofr.Ys* = *ofr.Ys (dsn.Pbr, adm.Pbs).*

3. On the basis of *emp*. *Fo* = *dsn.Fo* and eq. (3.1.3), we calculate values *dsn.dem.Fa*.
4. Using equation (4.2.15) we find *dsn.ofr.Ys*.
5. For finding of the number of internal switching lines *Ns,* the recursion Erlangs'B – formula (5.1.2). From numerical point of view, the following linear form is the most stable:

where $I(Ns, ofr.Ys) = 1 + \dfrac{Ns}{ofr.Ys} I(Ns-1, ofr.Ys),$ where $I(0, ofr.Ys) = 1$ (5.1.2)

and $I(Ns, ofr.Ys) = \dfrac{1}{Pbs\,(Ns, ofr.Ys)}$. This recursion formula is exact, and for large values of (*Ns*, *ofr.Ys*) there are no round of errors [5].

The received results for numerical inversion of the Erlang's formula (for finding the number of switching lines *Ns*) was confirmed with results of others commercial computer programs.

Note: If $Pbr \neq \dfrac{S_1 - S_3\, adm.Pbs}{1 - adm.Pbs}$ then the NRDT is solvable and there is proposed algorithm for its solution.

When *Pbr* = 0 the network loading is rather low and it is not of great practical interest, but in this case a mathematical research is made also.

6. Conclusions

1. Detailed normalized conceptual model, of an overall (virtual) circuit switching telecommunication system (like PSTN and GSM) is used. The model is relatively close to the real-life communication systems with homogeneous terminals.
2. An approach for network redimensioning of telecommunication system is described. Network Redimensioning Task (NRDT) with condition to be support administrative preassigned QoS level is formulated and is solved analytical. The network redimensioning task (NRDT) is formulated on the base of preassigned values of QoS parameter *adm.Pbs* and its corresponding GoS – parameters.
3. General blocking probability *adm.Pbs* and probability of finding B – terminal busy (*Pbr*) as GoS – parameters are used.
4. An analytical solution of the NRDT is shown. The received solutions are researched concerning GoS parameters blocking probability of finding B – terminal busy (*Pbr*).
5. An algorithm and a computer program for a calculation the values of the offered (*ofr.Ys*), calling rate (*Fa, rep.Fa*) and the number of equivalent switching lines *Ns*, are proposed. The results of numerical solution are derived and graphically shown.

6. New numerical results of the human behaviour characteristics R_1, R_2 and R_3 regarding *Prbr* and *Prbs* are obtained.
7. Sensitivities of the received results regarding some users' behaviour parameters (*Prbr* and *Prbs*) are researched. The influence of repeated calls after blocking due to lack of resources (*Prbs*) or/and due to find B – terminal "busy" (*Prbr*) over *rep.Fa/Fa* and *Ns/Nab* is investigated.
8. Numerical experiments are made and the results are graphically shown as well.
9. The received results, in NRDT, make the network redimensioning, based on QoS requirements easily, due to clear methodology presented and estimation of important parameters sensitivity shown.

The described approach is applicable directly for every (virtual) circuit switching telecommunication system (like GSM and PSTN) and may help considerably for ISDN, BISDN and most of core and access networks dimensioning. For packet switching systems, like Internet, proposed approach may be used as a comparison basis and when they work in circuit switching mode.

Bibliography

1. Dorn, William S., Daniel McCracken, 1972 – Numerical Methods with FORTRAN IV case Studies –John Wiley &Sons, Inc. New York (translated in Bulgarian 1977);
2. ITU-T Recommendation E.501: Estimation of traffic offered in the network. (Previously CCITT - Recommendation, revised 26. May 1997);
3. ITU-T Recommendation E.600: Terms and Definitions of Traffic Engineering (Melbourne, 1988; revised at Helsinki, 1993);
4. CISCO: Traffic Analysis for Voice over IP (Version Number 1 – created 06/25/2001 and Version Number 2 incorporated editorial comments – 11/01/2001);
5. Iversen V. B., 2004. Teletraffic Engineering and Network Planning, Technical University of Denmark, pp.125, 127;
6. Poryazov S. A, Saranova E. T. Some General Terminal and Network Teletraffic Equations in Virtual Circuit Switching Systems. Symposium "Modeling and Simulation Tools for Emerging Telecommunications Networks: Needs, Trends, Challenges, Solutions", (Proceedings of the European COST-285 Telecommunications Symposium) Munich, Germany, 8-9 September 2005, Institut für Technische Informatik, Universität der Bundeswehr München. Springer Sciences+Business Media, LLC 2006, ISBN 0-387-32921-8, pp. 471-505;
7. Poryazov S. A. 2005b. What is Offered Traffic in a Real Telecommunication Network? COST 285 TD/285/05/05; 19th International Teletraffic Congress, Beijing, China, August 29- September 2, 2005, accepted paper No 32-104A.;
8. E. T. Saranova. Redimensioning of Telecommunication Network based on ITU definition of Quality of Services Concept, In: Proceedings of the International Workshop "Distributed Computer and Communication Networks", Sofia, Bulgaria, 2006, Editors: V. Vishnevski and Hr. Daskalova, Technosphera publisher, Moscow, Russia, 2006, p. 12;
9. ITU-T Recommendation E.734 (10/96), Methods for allocating and dimensioning Inteligent Network (IN) resources;
10. Saranova E. T.. Network Redimensioning Sensitivity from Users' Behaviour Parameters, Symposium "Modeling and Simulation Tools for Emerging Telecommunications Networks: Needs, Trends, Challenges, Solutions", (Proceedings of the European COST-285 Telecommunications Symposium) Surrey, UK, 28 - 31 Mart, 2007, Centre for Communications Systems Research, University of Surrey, UK.

Authors' Information

Emiliya Saranova *– e-mail: saranova@hctp.acad.bg*

Institute of Mathematics and Informatics - Bulgarian Academy of Science, Sofia, Bulgaria

High College of Telecommunication and Posts, Sofia, Bulgaria

VOIP TRAFFIC SHAPING ANALYSES IN METROPOLITAN AREA NETWORKS

Rossitza Goleva, Mariya Goleva, Dimitar Atamian, Tashko Nikolov, Kostadin Golev

Abstract: This paper represents VoIP shaping analyses in devices that apply the three Quality of Service techniques – IntServ, DiffServ and RSVP. The results show queue management and packet stream shaping based on simulation of the three mostly demanded services – VoIP, LAN emulation and transaction exchange. Special attention is paid to the VoIP as the most demanding service for real time communication.

Keywords: Packet network, IP, Quality of Service, VoIP, shaping.

ACM Classification Keywords: C.4 Performance of Systems, C. Computer Systems Organization, C.2 Computer-Communication Networks

Introduction

IP networks and their Quality of the Service are challenging area for investigation. In spite of the fact that they are easy for use, enough cheap and quite useful in human life there is recently high demand for IP network use instead of all other kind of communication. Real time and non real time services and applications interwork on the same infrastructure. Different services have different quality requirements. The quality offered by the network depends on the traffic. In this paper we analyze the traffic shaping effect of the three mostly used techniques – IntServ, DiffServ, and RSVP. The analyses are made on the basis of the three popular services – VoIP, LAN emulation, transaction exchange [Jha], [Janevski], [Pitts], [Ralsanen]. The shaping effect is estimated under typical queueing circumstances. The model uses queues and priorities specific for the IntServ, DiffSerf, and RSVP. The reason is to investigate the effect that can be reached without implementation of the expensive shaping devices. This fractional shaping phenomenon is important in small to medium wire and wireless Metropolitan Area Networks (MAN) that grow rapidly. Changing circumstances in ad hoc networks also can apply the results presented.

Traffic sources

The traffic sources generate combination of three types of services in the network – Voice over IP, LAN emulation and transaction exchange. The size of the example network is typical for the business area. Some assumptions are made for every traffic source. In Voice over IP (VoIP) service silence and talk intervals are exponentially distributed with equal mean values [Jha], [Pitts]. There are authors who use talk to silence ratio of ½. Others do prefer to use on-off model for voice service. The behavior of the VoIP end-user is supposed to be similar to the phone user. The limits for waiting times are calculated under consideration of end-to-end delay bounds for every service [Lavenberg], [Iversen]. The same is valid for queue length. Servicing times per packets are fixed for LAN connection of 100 Mbps. Table 1 represents traffic sources parameters in the model.

LAN emulation is modeled with sessions. Sessions are established for any Internet connections. Packet rate is higher in comparison to the VoIP. Session duration is low. The traffic source is behaving as on-off model with exponential duration of the silence and transmission intervals [Lavenberg]. Transaction exchange is specific with few packets exchange. The service is not time demanding. Sessions are short and similar to the datagram exchange.

Number of traffic sources is taken from the typical business area. Packets are taken to be long. IP packets of 800 bytes carry up to 80 milliseconds voice. This means that quality voice can be transmitted only in the area with up to 2-3 hops. Therefore, we design VoIP service for regional connectivity. More precision investigation can be done with 200 bytes voice packets.

Table 1. Traffic Sources Parameters

No	Parameter	VoIP	LAN emulation	Transactions
1.	Pear rate, packets per second	10	164	0
2.	Mean call/ session duration, sec	180	20	10
3.	Mean duration between calls/sessions, sec	360	10	15
4.	Mean talk/ silence duration, sec	20	5	2
5.	Distribution of call/series duration	Exponential	Exponential	Exponential
6.	Maximal waiting time, sec	0.00072	0.6	1
7.	Maximal number of waiting packets	210	1804	2
8.	Traffic sources	5000	500	1500
9.	Priorities	High	Medium	Low
10.	Packet length, bytes	800	800	800

Integrated Services

Integrated Services (IntServ) is a complex technique that ensures Quality of Service in IP networks. It is applied usually in access routers or gateways and tried to serve packets from different services in a different ways depending on the quality requirements. IntServ classifies services into three main classes depending on the traffic requirements [Janevski]:

- Elastic application;
- Tolerant real-time applications;
- Intolerant real-time applications.

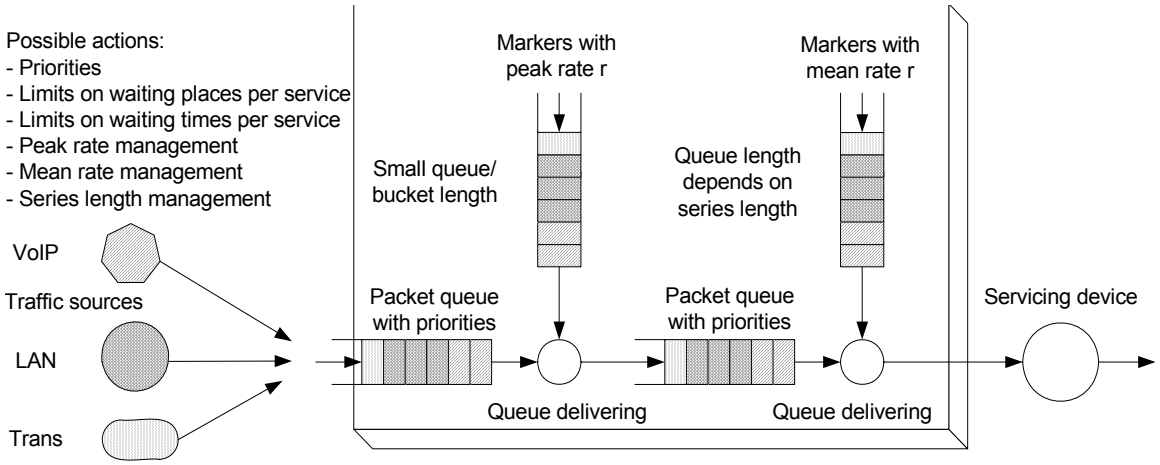


Figure 1. Black box IntServ model approximation

Elastic applications are served with “best effort” discipline [Tanenbaum]. They are served without any guarantee of quality level like transaction exchange. Tolerant real-time applications are delay sensitive and usually require high bandwidth. Token bucket model with peak rate control is a proper model for such traffic. LAN emulation is usually modeled this way. Some authors propose token bucket that controls series length and mean rate for more accuracy. Many authors propose the two token buckets to be connected in a cascade as it is shown on Figure 1

[Ralsanen]. Intolerant real-time applications require low delays and almost guaranteed bandwidth. The model with two cascaded token buckets is compulsory for such traffic [Jha]. VoIP service is intolerant to the quality degradation service. IntServ simulation model is based on two cascaded token buckets that bound peak rate, series length and mean rate of the traffic (Figure 1). The model is approximated as a black box that changes the characteristics of the data at output in specific for IntServ way. As a result after approximation and few calculations it is easy to derive simpler model with one FIFO queue, priorities, fixed rate at the output and different limits for waiting times in the queue. The resulting model is represented on Figure 2. This is the model that has been simulated further. Table 2 represents main data for model behavior.

Differentiated Services

Differentiated Services (DiffServ) is another quality management technique that is more applicable for core networks. Due to its nature DiffServ applies its rules on aggregated traffic. After appropriate marking of the aggregated packets they are gathered in the way that is defined for their class. There are three main types of services we highlighted in this paper [Pitts]:

- Premium service with low delay, low loss, guaranteed bandwidth like VoIP;

- Assured service with less requirements to the delay and loss in comparison to the premium service like LAN emulation;

- Olympic service with no time requirements at all like transaction exchange.

The model from Figure 2 with different parameters is used to represents DiffServ application. The parameters are shown on Table 2.

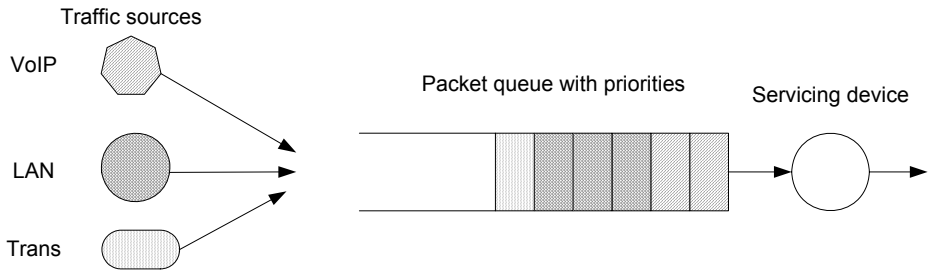


Figure 2. Final IntServ model with input data, bounds for waiting times and queue length specific for service type

RSVP

Resource Reservation Protocol (RSVP) is a technique useful for delay sensitive traffic like VoIP. Three types of services are identified for RSVP like:

- Wildcard filter that is applied to gather maximal requirements for given interface like LAN emulation;

- Shared explicit that is applied to gather maximal requirements for the interface taking into account called address. Transaction exchange is modeled as shared explicit service;

- Fixed filter that requires full reservation for quality sensitive services like VoIP.

The model simplified for IntServ and DiffServ procedures is applied with specific parameters for RSVP. Characteristics of the derived model are shown on Table 2.

Table 2. Model characteristics

No	Parameter	IntServ	DiffServ	RSVP
1.	Queue length, packets	2016	1840	1840
2.	VoIP queue length fraction, packets	210	200	200
3.	LAN queue length fraction, packets	1804	1640	1640
4.	Transaction queue length fraction, packets	2	2	2
5.	Maximal waiting time for VoIP, sec	0,000716	0,0303	0,07508
6.	Maximal waiting time for LAN, sec	0,6	0,27876	0,69
7.	Maximal waiting time for transactions, sec	1	1	1
8.	Priority for VoIP	Highest	Highest	Highest
9.	Priority for LAN	Medium	Medium	Medium
10.	Priority for transactions	Low	Low	Low

Results

Simulation is performed on C++ language. The pseudo exponential pseudo deterministic characteristics of the traffic sources are reached after usage of combination between many random generators [Kleinrock], [Iversen], [Lavenberg]. The queue behavior is complex due to the priorities and limits for waiting times. Many parameters have been derived from the model like time and space loss probabilities, probabilities to wait for different types of traffic, statistical data for probability distribution functions and probability density functions of the packets intervals, queue lengths, waiting times at many interface points of the model like output of the traffic sources, input and output of the queue. Statistical accuracy of the derived results is proven by Student criterion. IntServ, DiffServ and RSVP have different way to gather with packets and this influences the way they drop packets and shape them.

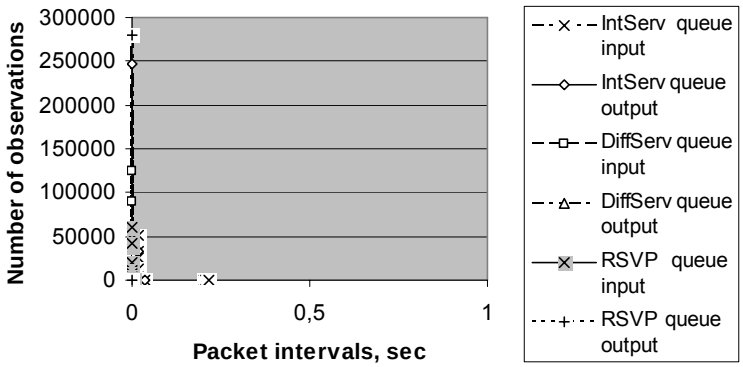


Figure 3. Delay variation reduction in RSVP

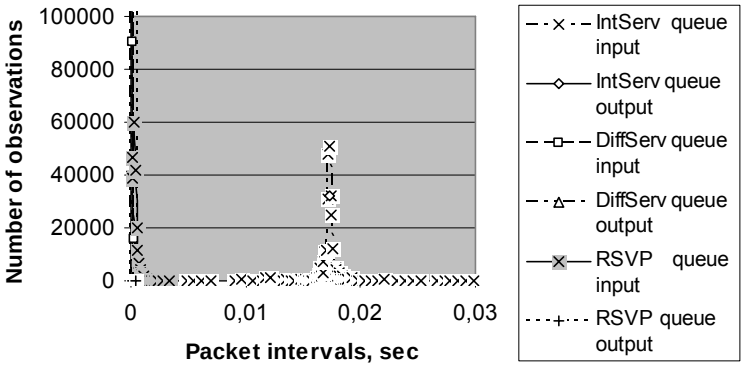


Figure 4. Delay variation reduction in IntServ

On Figure 3, 4 and 5 observations of packet intervals at the input and output of the queue are shown. It is interesting for shaping estimation. The effect of fast servicing in RSVP can be seen from Figure 3. The delay variation of the packet intervals is becoming smoother and tends to constant value. Similar result is visible for IntServ on Figure 4. On Figure 5 shaping of the IntServ and DiffServ is seen again.

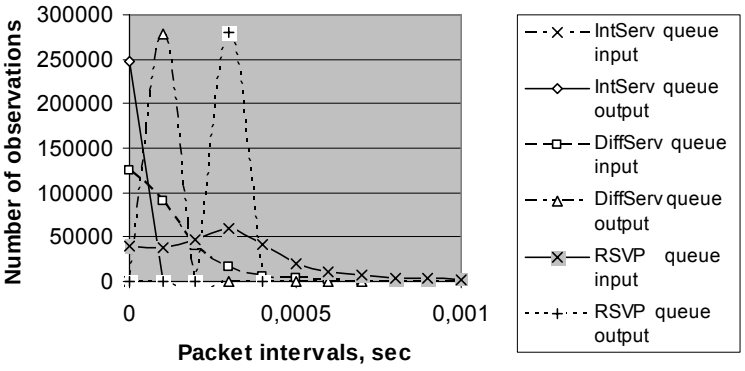


Figure 5. Delay variation reduction in IntServ and DiffServ

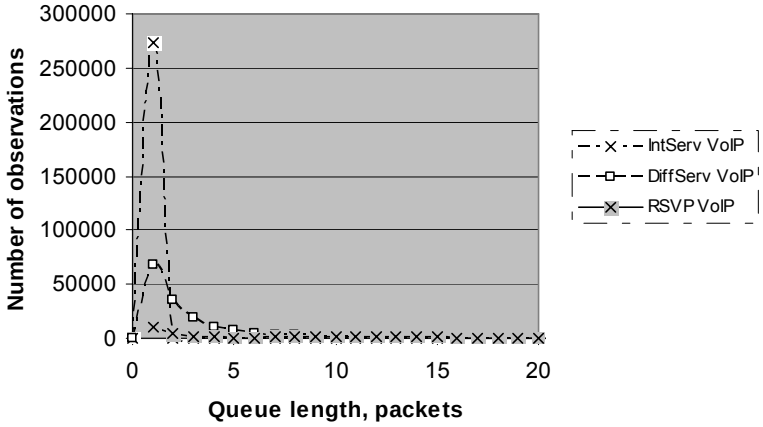


Figure 6. Observations of queue length in VoIP

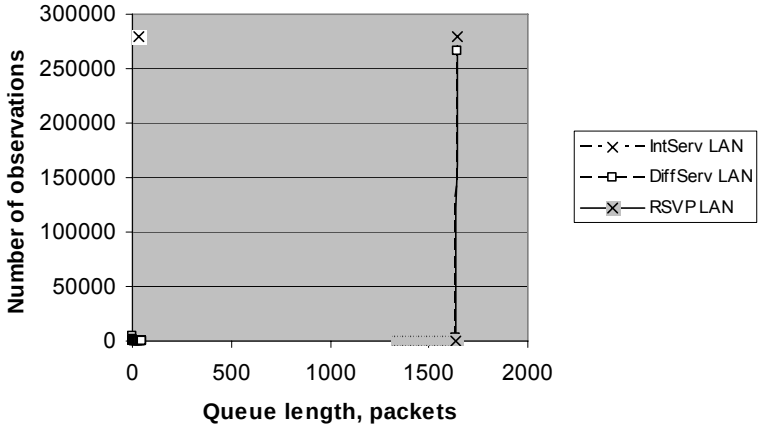


Figure 7. Observations of queue length in LAN

Interesting results that influence directly interfaces and queue management are derived on the basis of queue length per service type. The queue fraction of the three services is observed. It is visible from Figure 6 that for services with highest priority like VoIP IntServ is the most proper mechanism. With some not quite accurate approximation the distribution of the queue length can be considered exponential. Figure 7 represents the observations for LAN service. Because of the less critical waiting times and low priority the distribution tends to be deterministic.

On Fugure 8 and 9 the observations of packet intervals only between voice packets are shown. The statistical multiplexing effect and shaping phenomenon are due to the high priority of the voice traffic in comparison to the priority of the data traffic. On Figure 10 the shaping effect of the three techniques is visible. On Figure 11 and 12 only effect of DiffServ is obvious in different observation scales.

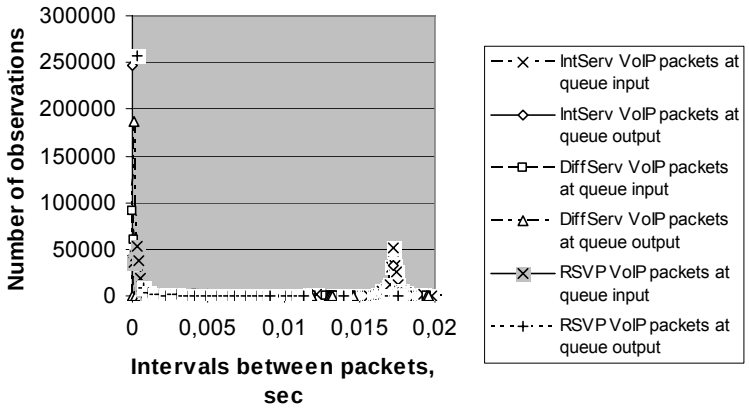


Figure 8. Observations of intervals between VoIP packets

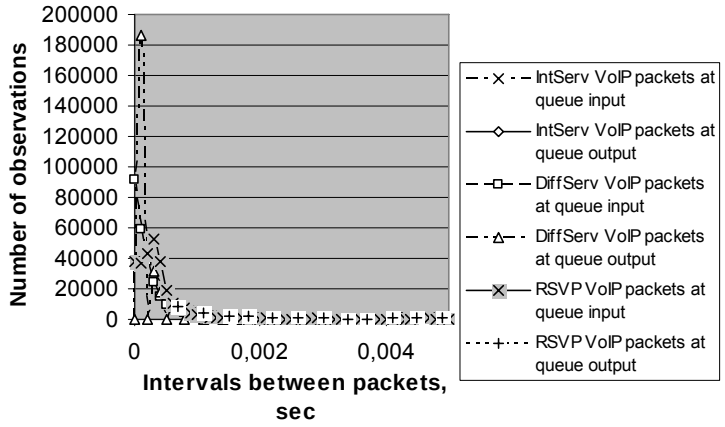


Figure 9. Observations of intervals between VoIP packets for DiffServ

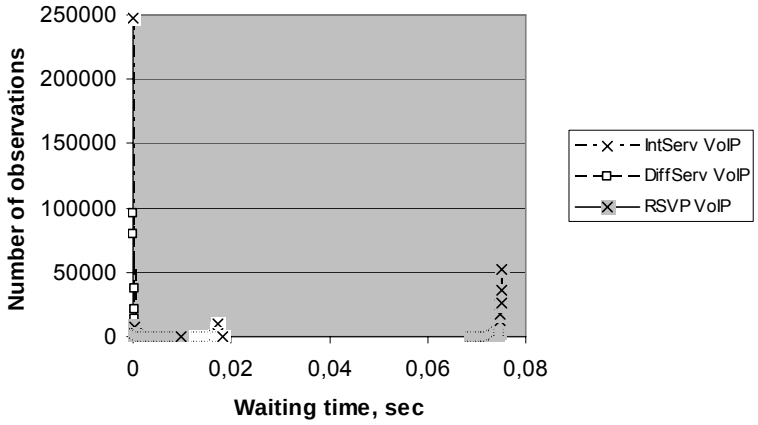


Figure 10. Observations of waiting times for VoIP packets

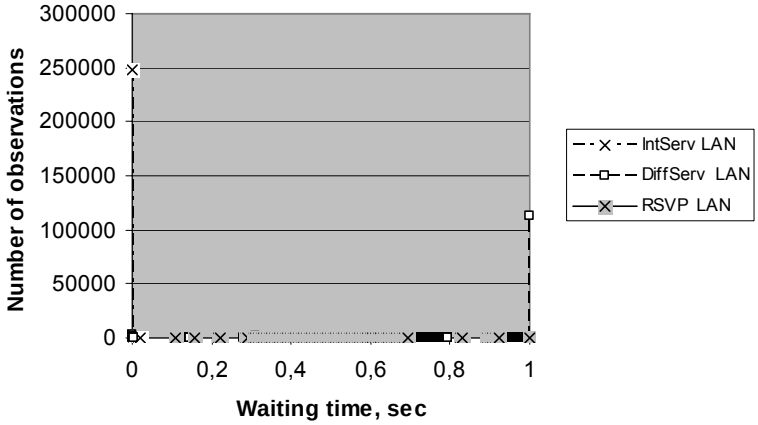


Figure 11. Observations of waiting times for LAN packets

Conclusion

In this paper we show observations of the packet intervals at the queue input and queue output as well as statistical data of queue length, waiting times and loss per service type (Table 3). These results demonstrate the specific characteristics of the queue as a packet shaper in three QoS management algorithms IntServ, DiffServ, RSVP. The shaping effect is possible for priority services types.

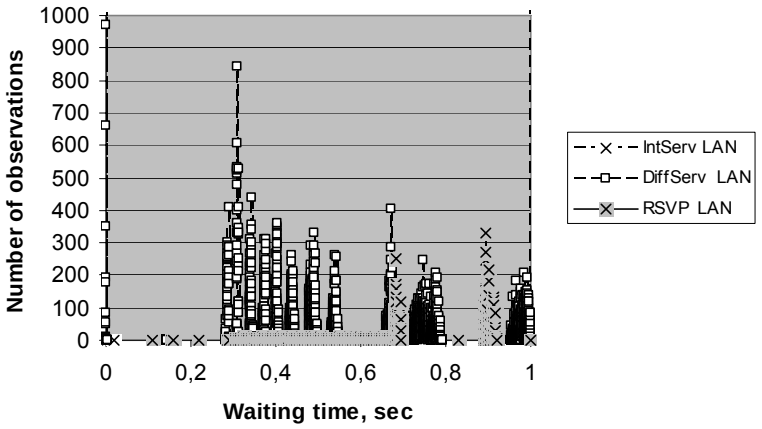


Figure 12. Observations of waiting times for LAN packets

Table 3. Queue length, waiting times, loss probability

Mechanism	IntServ	DiffServ	RSVP
Overall mean queue length, packets	37.97523	1586.961	1798.409
Mean queue length of VoIP fraction, packets	1	2.24207	156.9017
Mean queue length of LAN fraction, packets	35	1583.561	1639.675
Mean queue length of Trans fraction, packets	2	1.98331	2
Overall loss probability due to the lack of space	0.0094	0.77815	0.91222
VoIP packets loss probability due to the lack of space	0	0	0.33540
LAN packets loss probability due to the lack of space	0	0.89373	0.98747
Transaction packets loss probability due to the lack of space	1	0.99574	1
Overall loss probability due to waiting time bound	0.98865	0	0
VoIP packets loss probability due to waiting time bound	0.98109	0	0
LAN packets loss probability due to waiting time bound	1	0	0

Transaction packets loss probability due to waiting time bound	0	0	0
Overall probability to wait	0.00191	0.22074	0.08767
VoIP packet probability to wait	0.0189	0.9996	0.66433
LAN packet probability to wait	0	0.105130	0.01244
Transaction packet probability to wait	0	0.00339	0
Overall interface occupancy, fraction	0.13668	0.13644	0.13704
Interface occupancy due to the VoIP traffic, fraction	0.13621	0.05046	0.08955
Interface occupancy due to the LAN traffic, fraction	0.00048	0.08588	0.04749
Interface occupancy due to the transaction traffic, fraction	0	0.00009	0

The low delays for priority services types are due to the bigger delays for non priority service types. The waiting times are redistributed due to the QoS algorithm and priority. The deterministic nature of the packet streams suppress shaping and increase losses. The statistical multiplexing effect is very limited due to the deterministic streams. Mean values of the queue lengths, probability to wait, loss probability due to the lack of space and waiting time bounds per discipline and per service redistribution are visible from Table 3. They can be used for configuration planning of the time and space limits in the router interfaces.

The results demonstrate the capability of IntServ to define excellent service for its higher priority applications. It is promising in access networks. DiffServ shows excellent resource management and utilization and therefore is better for core services. RSVP is a good counterpart of IntServ in access networks.

The authors refine the simulation model with more traffic sources and more precise generation of the packets from these sources based on the observation of the real traffic. MMPP and geometric/ Weibull distributions are also considered. Limits criteria for queue management are under investigation.

Acknowledgements

This paper is sponsored by the Ministry of Education and Research of the Republic of Bulgaria in the framework of project No 105 “Multimedia Packet Switching Networks Planning with Quality of Service and Traffic Management”.

Bibliography

[Jha] [1] Jha, S., M. Hassan, “Engineering Internet QoS”, Artech Hose, 2002.

[Janevski] [2] Janevski, T., “Traffic Analysis and Design of Wireless IP Networks”, Artech House, 2003.

[Kleinrock] [3] Kleinrock, Leonard, “Queueing Systems”, Volumes I and II, John Wiley and Sons, 1976.

[Iversen] [4] Iversen, V., “Teletraffic Engineering Handbook”, ITU-D, 2005.

[Lavenberg] [5] Lavenberg, Stephen S., Editor, “Computer Performance Modeling Handbook”, Academic Press, 1983, ISBN 0-12-438720-9.

[Pitts] [6] Pitts, J., J. Schormans, “Introduction to IP and ATM Design and Performnce”, John Wiley&Sons, Ltd., 2000.

[Ralsanen] [7] Ralsanen, V., “Implementing Service Quality in IP Networks”, John Wiley & Sons, Ltd., 2003.

[Tanenbaum] [8] Tanenbaum, Andrew S., “Computer Networks, Second Edition”, Prentice-Hall International, Inc., 1989, ISBN 0-13-166836-6.

Authors' Information

Rossitza Iv. Goleva *– Assistant-Professor; Department of Telecommunications, Technical University of Sofia, Bulgaria, Kl. Ohridski blvd. 8, Sofia. 1756, Bulgaria; e-mail: rig@tu-sofia.bg*

Mariya At. Goleva *– Student; Department of Communication Networks, University of Bremen, Germany; e-mail: mgoleva@gmail.com*

Dimitar K. Atamian *- Assistant-Professor; Department of Telecommunications, Technical University of Sofia, Bulgaria, Kl. Ohridski blvd. 8, Sofia. 1756, Bulgaria; e-mail: dka@tu-sofia.bg*

Tashko Nikolov *– Assistant-Professor, Ph.D. Department of Telecommunications, Technical University of Sofia, Bulgaria, Kl. Ohridski blvd. 8, Sofia. 1756, Bulgaria; e-mail: tan@tu-sofia.bg*

Kostadin At. Golev *– Developer; Bianor Ltd., Bulgaria; e-mail: kotseto@gmail.com*

STUDY OF QUEUEING BEHAVIOUR IN IP BUFFERS

Seferin Mirtchev

Abstract*: It is unquestioned that the importance of IP network will further increase and that it will serve as a platform for more and more services, requiring different types and degrees of service quality. Modern architectures and protocols are being standardized, which aims at guaranteeing the quality of service delivered to users. In this paper, we investigate the queueing behaviour found in IP output buffers. This queueing increases because multiple streams of packets with different length are being multiplexed together. We develop balance equations for the state of the system, from which we derive packet loss and delay results. To analyze these types of behaviour, we study the discrete-time version of the "classical" queue model M/M/1/k called Geo/Gx/1/k, where Gx denotes a different packet length distribution defined on a range between a minimum and maximum value.*

Keywords*: delay system, queueing analyses, discrete time queue, IP traffic modelling; packet size distribution.*

ACM Classification Keywords: *G.3 Probability and statistics: queueing theory, I.6.5 Model development*

Introduction

The initial motivation for this paper is the necessity of traffic engineering in IP networks. Many analyses of Internet traffic behaviour require accurate knowledge of the traffic characteristics for purposes ranging from a management of the network quality of service to modelling the effect of new protocols on the existing traffic mix.

Modern architectures and protocols are being standardized, which aims at guaranteeing the quality of service delivered to users. The proper functioning of these protocols requires an increasingly detailed knowledge for statistical characteristics of IP packets. The amount of information flowing through the network also increases, and the challenge is to obtain the accurate information from a huge set of data packets.

The packet queueing in an IP router arises because multiple streams of packets from different input ports are being multiplexed together over the same output port. A key characteristic is that the packets have different length. The minimum header size in IPv4 is 20 octets, and in IPv6, it is 40 octets. The maximum packet size depends on the specific sub-networks technology: 1500 octets in Ethernet and 1000 octets are common in X.25 networks. The packet length distribution measured from the real traces exhibits the well-known multi-mode behaviour, with peaks for very short packets and for the different maximum transfer units in the network, with a

dominating peak at 1500 bytes, due to the size of Ethernet frame. This specific packet length distribution has a direct impact on the service time and we need a different approach to the queueing analysis.

Discrete-time queueing systems have been a research topic for several decades now and there are many reference works on discrete-time queueing theory. Over the years, different methodologies have been developed to assess the performance of queueing systems. The two main analytical approaches are the matrix analytic method and the transform method for discrete and for continuous-time analyses. Many authors have considered the Geo/G/1 queueing system [Pitts, 2000], [Mirtchev, 2006], [Vicari, 1996], [Zang, 2001].

In [Atencia, 2005] is carried out a complete study of a discrete-time single-server queue with geometrical arrivals of both positive and negative customers. Negative arrivals are used as a control mechanism in many telecommunication and computer networks. [Atencia, 2006] is concerned with the study of a discrete-time single-server retrial queue with geometrical inter-arrival times and a phase-type service process. An iterative algorithm to calculate the stationary distribution of Markov chain is given.

[Salvador, 2004] is proposed a traffic model and a parameter fitting procedure that are capable of achieving accurate prediction of the queuing behaviour for IP traffic exhibiting long-range dependence. The modelling process is a discrete-time batch Markovian arrival process (dBMAP) that jointly characterizes the packet arrival process and the packet size distribution. In the proposed dBMAP, packet arrivals occur according to a discrete-time Markov modulated Poisson process (dMMPP) and each arrival is characterized by a packet size with a general distribution that may depend on the phase of the dMMPP.

[Cao, 2004] is presented an introduction to bandwidth estimation and a solution to the problem of the best-effort traffic for the case where the quality criteria specify negligible packet loss. The solution is a simple statistical model, which is built and validated using queueing theory and extensive empirical study.

It has been shown [Dan, 2005] that in the case of real-time communications, for which small buffers are used for delay reasons, short range dependence dominates the loss process and so the Markov-modulated Poisson process (MMPP) might be a reasonable source model. They have presented an exact mathematical model for the loss process of a MMPP+M/Ek/1/K queue and have concluded that the packet size distribution affects the packet loss process and thus the efficiency of forward error correction.

In this paper, we investigate the basic queueing behaviour of packets found in IP output buffers. This queueing is complicated because multiple streams of packets are being multiplexed together. The traffic is being generated from the packets of varying sizes that arrive for transmission on the link. The packets can queue up and loss if their size is bigger than the free positions of the buffer. The quality metrics for the best-effort traffic on the Internet are the packets loss and delay. To analyze these types of behaviour, we study the discrete-time version of the "classical" queue model M/M/1/k called Geo/Gx/1/k, where Gx denotes a different packet length distribution. We developed balance equations for the state of the system, from which we derived packets loss and delay.

Balance equations for the queue model Geo/Gx/1/k

Let us consider a single server finite queue delay system ***Geo/Gx/1/k*** with a geometric distributed inter-arrival time and different distributions of the packet length: truncated geometric, binomial, discrete uniform and discrete triangular. These packet length distributions are defined on a range between a minimum and maximum value.

We consider queueing phenomena in discrete-time queueing systems. That is, we assume a fundamental time unit (time slot), the time to transmit an octet (byte), T_b.

Customers arrive in the queueing system under consideration during the consecutive slots, but they can only start service at the beginning of slots. That is, service of customers is synchronized with respect to slot boundaries. Further, customer service times are integer multiples of the slot length, which implies that customers leave the system at slot boundaries. During the consecutive slots, packets arrive in the system, are stored in a finite capacity queue and are served by a single server on a first in first out (FIFO) basis (fig.1).

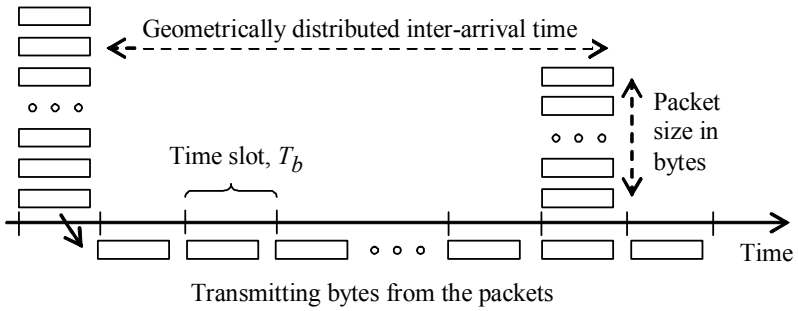


Fig.1. Timing of events in the Geo/Geo/1/k queueing system

We use a Bernoulli process for the packet arrivals, i.e. a geometrically distributed number of slots between arrivals. Let the probability that a packet arrives in an octet slot is ***p***.

In this model, we assume a truncated geometric distribution at variable packet sizes with a minimum value m_1 and a maximum value m_2, as the first kind of distribution.

Let the probability that a packet completes service at the end of an octet slot is ***q***. We define the probability that the packet size is ***n*** octets:

$$b_n = \frac{q(1-q)^{n-m_1}}{q\sum_{r=0}^{m_2-m_1}(1-q)^r}, \qquad m_1 \le n \le m_2 \ . \tag{1}$$

The mean number of bytes in the packet by definition is

$$b = \sum_{i=m_1}^{m_2} i\, b_i \approx 1/q + m_1 \ . \tag{2}$$

The second kind of a packet size distribution is binomial

$$\begin{aligned} b_n &= \binom{m_2-m_1}{n-m_1} q^{n-m_1}(1-q)^{m_2-n}, \qquad m_1 \le n \le m_2 \ , \\ b &= m_1 + (m_2-m_1)q \end{aligned} \tag{3}$$

The third kind of a packet size distribution is discrete uniform

$$\begin{aligned} b_n &= \frac{1}{m_2-m_1+1} \quad for\ all\ values\ of\ n, \qquad m_1 \le n \le m_2 \ , \\ b &= (m_2+m_1)/2 \end{aligned} \tag{4}$$

The next kind of a packet size distribution is discrete triangular. When the mode is equal to the minimum value, we have linear decreasing distribution with the following probabilities that the packet size is ***n*** octets and the mean number of the bytes in the packet

$$\begin{aligned} b_n &= \frac{m_2-n+1}{\sum_{r=m_1}^{m_2} m_2-r+1} \ , \qquad m_1 \le n \le m_2 \ , \\ b &= m_1 + (m_2-m_1)/3 \end{aligned} \tag{5}$$

When the mode is equal to the maximum value, we have linear increasing discrete triangular distribution

$$\begin{aligned} b_n &= \frac{n-m_1+1}{\sum_{r=m_1}^{m_2} r-m_2+1} \ , \qquad m_1 \le n \le m_2 \ , \\ b &= m_1 + 2(m_2-m_1)/3 \end{aligned} \tag{6}$$

Thus we have a batch arrival process with geometrically distributed inter-arrival times. That is, the number of slots that separate consecutive slots where there are customer arrivals, constitute a series of independent and identically geometric distributed random variables. The probability no octets arriving in a time slot is

$$a_0 = 1 - p \quad . \tag{7}$$

The probability that ***n*** octets arriving in a time slot is

$$a_n = p\, b_n, \qquad m_1 \le n \le m_2 \quad . \tag{8}$$

The mean packet service time is the octet transmission time multiplied by the mean number of octets

$$\tau = T_b \sum_{i=m_1}^{m_2} i\, b_i = T_b\, b, \; s \quad . \tag{9}$$

The mean arrival rate is

$$\lambda = p / T_b, \quad packets/s \quad . \tag{10}$$

Therefore, the offered traffic is given by

$$A = \lambda \tau = p \sum_{i=m_1}^{m_2} i\, b_i, \; erl \quad . \tag{11}$$

We define the state probability P_i of being of state ***i***, as the probability that there are ***i*** octets in the system at the end of any time slot. For the system to contain ***i*** bytes at the end of any time slots it could have contained any of *0,1,2,…..,i+1* at the end of the previous slot. State ***i*** can be reached from any of the states ***0*** up to ***i*** by a precise number of arrivals. To move from ***I +1*** to ***i*** requires that there are no arrivals.

We can write the first equation by considering all the ways in which it is possible to reach the empty state

$$P_0 = P_0 a_0 + P_1 a_0 \quad . \tag{12}$$

Similarly, we find a formula for the next state probabilities by writing the balance equations

$$P_i = P_{i+1} a_0, \qquad 1 \le i \le m_1 - 1 \quad . \tag{13}$$

We continue with this process when the packet arrives in a time slot with length between m_1 and m_2.bytes

$$\begin{aligned}
P_{m_1} &= (P_0 + P_1) a_{m_1} + P_{m_1+1} a_0 \\
P_{m_1+1} &= (P_0 + P_1) a_{m_1+1} + P_2 a_{m_1} + P_{m_1+2} a_0 \\
&\quad o \;\; o \;\; o \\
P_{m_2} &= (P_0 + P_1) a_{m_2} + P_2 a_{m_2-1} + \ldots + P_{m_2-m_1+1} a_{m_1} + P_{m_2+1} a_0 \\
P_{m_2+1} &= P_2 a_{m_2} + P_3 a_{m_2-1} + \ldots + P_{m_2-m_1+2} a_{m_1} + P_{m_1+2} a_0 \\
&\quad o \;\; o \;\; o \\
P_{k-1} &= P_{k-m_2} a_{m_2} + P_{k-m_2+1} a_{m_2-1} + \ldots + P_{k-m_1} a_{m_1} + P_k a_0 \\
P_k &= P_{k-m_2+1} a_{m_2} + P_{k-m_2+2} a_{m_2-1} + \ldots + P_{k-m_1+1} a_{m_1} + P_{k+1} a_0
\end{aligned} \tag{14}$$

Then using the fact that all the state probabilities must sum to 1

$$\sum_{i=0}^{k+1} P_i = 1 \quad , \tag{15}$$

We can solve the system equations (12), (13), (14) and 15 and calculate the state probabilities.

Performance Measures

The carried traffic is equivalent to the probability that the system is busy

$$A_o = 1 - P_0 , \ erl \quad . \tag{16}$$

The packet congestion probability is the ratio of lost traffic (offered minus carried traffic) to offered traffic

$$B = \left(A - A_o\right) / A \quad . \tag{17}$$

The mean number of bytes and packets present in the system in steady state by definition is

$$L_b = \sum_{j=1}^{k+1} j P_j , \ bytes; \qquad L_p = L_b / b , \ packets \ . \tag{18}$$

From the Little formula, we have the normalized mean system time of the bytes (time is measured in time slots)

$$\frac{W_b}{T_b} = \frac{L_b}{T_b \lambda b} = \frac{L_b}{A} \quad . \tag{19}$$

Numerical Results

In this section, we give numerical results obtained by a Pascal program on a personal computer. The described methods were tested on a computer over a wide range of arguments.

Figures 2 and 3 show the stationary probability distribution in a single server queue *Geo/Gx/1/k* with 0.8 and 0.7 erl offered traffic respectively, 1000 waiting positions, 30 bytes minimum packet length, 80 bytes maximum packet length and different packet length distributions: discrete uniform, truncated geometric, binomial, discrete triangular decreasing and discrete triangular increasing. We can see that the probability distributions are almost linear decreasing in logarithmic scale and the influence of the packet length distribution kind on the stationary probability is negligible even though in case of discrete triangular increasing packet length distribution.

Figures 4 and 5 illustrate the dependence on the packet congestion probability from the queue length when the offered traffic is 0.7 erl, the range of packet length is from 30 to 80 bytes and different packet length distributions. When the queue length is big the packet congestion probability is almost linear decreasing in logarithmic scale. The packet length distribution in defined rage is not so essential. The main reason for this behaviour is the fact that the packet length is limited.

Figures 6 and 7 compare the packet congestion probability when the offered traffic is 0.8 erl, the range of packet length is from 30 to 80 bytes, truncated geometric and binomial distribution accordingly and different mean packet size. We can see that the influence of the mean packet length on the packet congestion probability is big.

Figures 8 and 9 present the normalized mean system time of the bytes (W/T_b) as function of the traffic intensity when the queue length is 1000 bytes, the range of packet length is from 30 to 80 bytes, truncated geometric and binomial distribution accordingly and different mean packet size. The influence of the mean packet size on the mean system time is significant when the offered traffic is smaller then 1 erl.

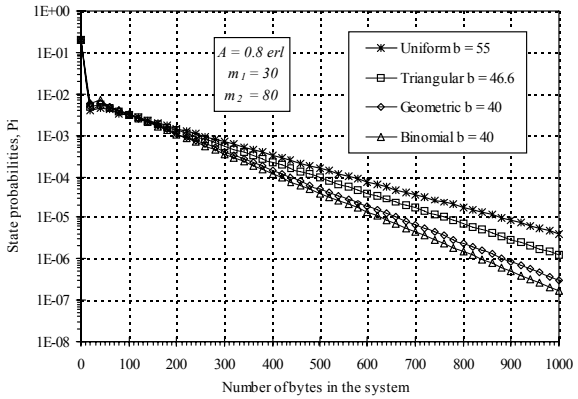


Fig.2. Graph of the state probability distributions for a finite queue with Geometric, Binomial, Uniformly and Triangular decreasing packet length distribution

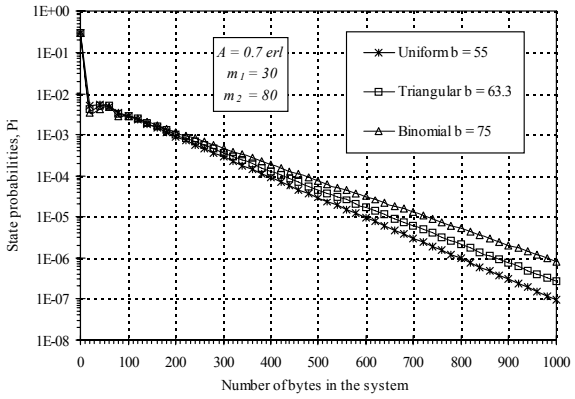


Fig.3. Graph of the state probability distributions for a finite queue with Binomial, Uniformly and Triangular increasing packet length distribution

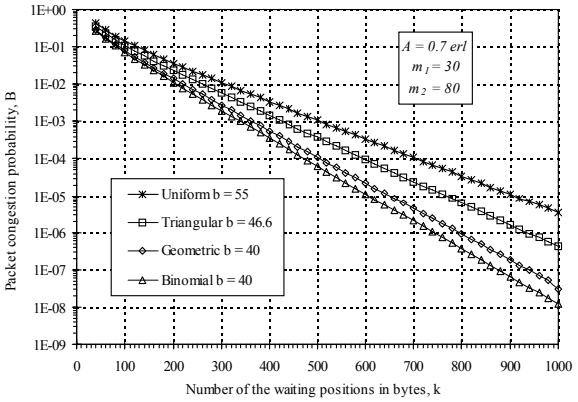


Fig.4. Packet congestion probability in the Geo/Gx/1/k with different packet length distributions and mean packet lengths between the minimum and the average value

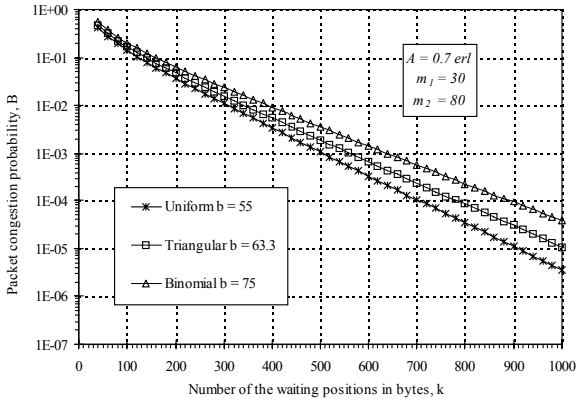


Fig.5. Packet congestion probability in the Geo/Gx/1/k with different packet length distributions and mean packet lengths between the average and the maximum value

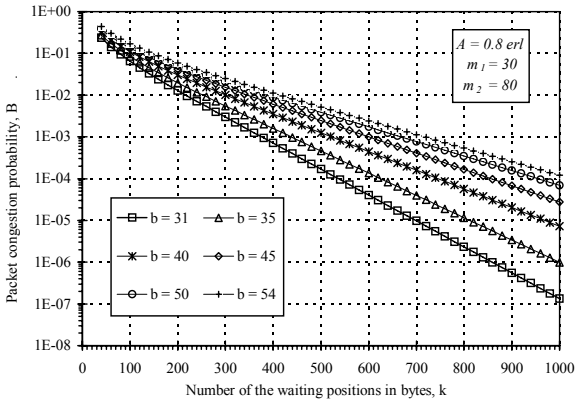


Fig.6. Packet congestion probability in discrete time single server queue with a truncated geometric packet length distribution and different mean packet lengths between the minimum and the average value

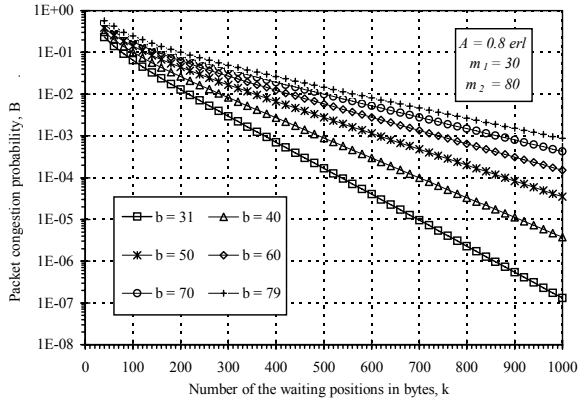


Fig.7. Packet congestion probability in discrete time single server queue with a binomial packet length distribution and different mean packet lengths between the minimum and the maximum value

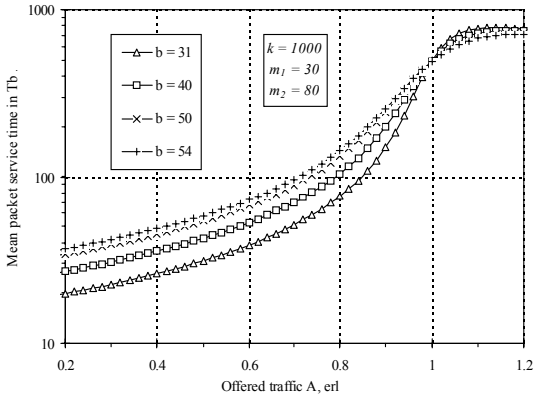


Fig.8. Normalized mean system time of the bytes in discrete time single server queue with a truncated geometric packet length distribution and different mean packet lengths

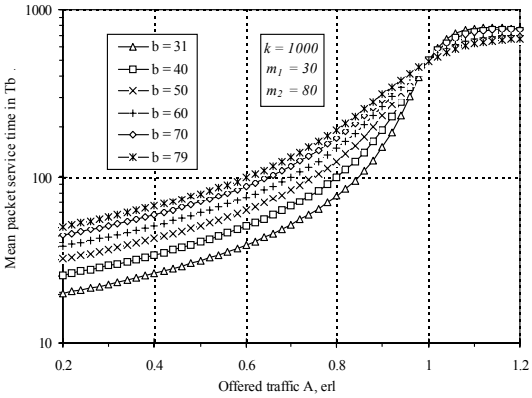


Fig.9. Normalized mean system time of the bytes in discrete time single server queue with a binomial packet length distribution and different mean packet lengths

Conclusion

In this paper, different distributions of the packet length: truncated geometric, binomial, discrete uniform and discrete triangular are used and explained. A basic discrete-time single server teletraffic system Geo/Gx/1/k is examined in detail.

The proposed approach provides a unified framework to model discrete-time single server queue. Numerical results and subsequent experience have shown that this approach is accurate and useful in both analyses and simulations of traffic systems.

The importance of a single server queue in a case of a geometric input stream and different distributions of the packet length comes from its ability to describe behaviour that is to be found in more complex real queueing systems. It is the case in a general traffic system, which is an important feature in designing telecommunication networks and systems.

The results presented here add a new aspect to the evaluation of the discrete-time queueing system, and serve as a basis for future research on guaranteeing the quality of service

In conclusion, we believe that the presented formulas will be useful in practice.

Acknowledgements

This paper is sponsored by the National Science Funds of MES - Bulgaria in the framework of project **BY-TH-105/2005** “Multimedia Telecommunications Networks Planning with Quality of Service and Traffic Management”.

Bibliography

[Atencia, 2005] Atencia I. and P. Moreno. A single-server G-queue in discrete-time with geometrical arrival and service process. Perform. Eval. 59: pp. 85-97 (2005)

[Atencia, 2006] Atencia I, P. Bocharov and P. Moreno. A discrete-time Geo/PH/1 queueing system with repeated attempts. Информационные процессы, Том 6, N: 3, стр. 272-280 (2006).

[Cao, 2004] Cao J., W. Cleveland and D. Sun. Bandwidth estimation for best-effort Internet traffic. Statist. Sci., Volume 19, Number 3 (2004), pp. 518-543.

[Dan, 2005] Dan G., V. Fodor, and G. Karlsson, "Packet size distribution: an aside?" in Proc. of QoS-IP'05, pp. 75–87, February 2005.

[Farber, 2002] Farber J., S. Bodamer and J. Charzinski. Measurement and Modelling of Internet Traffic at Access Networks, Proceedings of the EUNICE'98,1998,196-203.

[Janevski, 2003] Janevski T., D. Temkov, A. Tudjarov: Statistical Analysis and Modelling of the Internet Traffic. ICEST Sofia, 2003, pp. 170-173.

[Mirtchev, 2006] Mirtchev S., G. Balabanov and S. Statev, New Teletraffic Models in the IP Networks. National Conference with Foreign Participation, Telecom'2006, Varna, Bulgaria, 2006 (in Bulgarian).

[Pitts, 2000] Pitts J.and J. Schormans. Introduction to IP and ATM Design and Performance - 2nd Ed., John Wiley & Sons, 2000.

[Salvador, 2004] Salvador P., A. Pacheco and R. Valadas. Modelling IP traffic: joint characterization of packet arrivals and packet sizes using BMAPs. Computer Networks, Volume 44, Issue 3 , 2004, pp. 335-352.

[Vicari, 1996] Vicari N. and P. Tran-Gia. A numerical analysis of the Geo/D/N queueing system. Technical Report 04, COST-257, 1996.

[Zang, 2001] Zhang Z. and N.Tian. Discrete Time Geo/G/1 Queue with Multiple Adaptive Vacations, Queueing Systems, Volume 38, Number 4, August 2001, pp. 419-429.

Authors' Information

Seferin Mirtchev *– Technical University of Sofia, Kliment Ohridski St., N:8, Bl.1, Sofia-1000, Bulgaria; e-mail: stm@tu-sofia.bg*

TOWARDS USEFUL OVERALL NETWORK TELETRAFFIC DEFINITIONS

Stoyan Poryazov

Abstract. *A detailed conceptual and a corresponding analytical traffic models of an overall (virtual) circuit switching telecommunication system are used. The models are relatively close to real-life communication systems with homogeneous terminals. In addition to Normalized and Pie-Models Ensue Model and Denial Traffic concept are proposed, as a parts of a technique for presentation and analysis of overall network traffic models functional structure; The ITU-T definitions for: fully routed, successful and effective attempts, and effective traffic are re-formulated. Definitions for fully routed traffic and successful traffic are proposed, because they are absent in the ITU-T recommendations; A definition of demand traffic (absent in ITU-T Recommendations) is proposed. For each definition are appointed: 1) the correspondent part of the conceptual model graphical presentation; 2) analytical equations, valid for mean values, in a stationary state. This allows real network traffic considered to be classified more precisely and shortly. The proposed definitions are applicable for every telecommunication system.*

Keywords: *Overall Network TrafficTheory, ITU-T Definitions, Virtual Circuits Switching.*

1 Introduction

The first what we need for usable Overall Network and Terminal Traffic Theory, is a complete set of clear, precise and useful definitions, particularly for overall network characteristics.

State of the art: Expressions "offered traffic" and "demand traffic" are not found in "ETSI Publications Download Area" [http://pda.etsi.org/pda/queryform.asp] and in [ANSI 2001]. The ITU-T definition of offered traffic is not valid

for real telecommunication systems [Poryazov 2005] and that one for demand traffic is simply absent, despite usage in ITU-T Recommendations of expression “demand traffic” three times, and of “traffic demand” – 50 times.

Objective of the research: To trigger discussions towards establishing stable fundaments of Overall Network Traffic Theory.

Methods used: Conceptual telecommunication network modeling, influenced by Structural Programming approach. The reasoning is illustrated with circuit switching network models, because:

1) They are relative simple;
2) We have an existing Overall Network Teletraffic Model, consisting conceptual and correspondent analytical models;
3) “…the teletraffic theory of the Internet with dimensioning methods is mainly the topic of the future.” [Molnar 2006];
4) We are discussing base traffic concepts and definitions, which have to be valid in any telecommunication system.

All assumptions, notations and equations, not mentioned here, are explaneted in [Poryazov 2005] and, in more details, in [Poryazov, Saranova 2006].

2. Conceptual and reference models

2.1. Normalized structure of traffic models

In this paper three types of virtual devices are used: base, comprising base devices (enforsing group limitations on the comprised base devices, e.g. maximal sum of capacities) and aggregating base devices (used in the reasoning only).

2.1.1. Base Virtual Devices and Their Parameters

In the normalized models, used in this paper, every base virtual device, except the switches, has no more than one entrance and/or one exit. Switches, as a rule, have one entrance and two exits, but, as exception, may have more. The structural normalization is possible for every computer program [Bohm&Jacopini 1966] and therefore for every model presentable as a computer program (e.g. computer simulation model). We will use base virtual device types with names and graphic notation shown on Fig.1. For every device we propose the following notation for its parameters: Letter *F* stands for calling rate (frequency) of the flow [calls/sec.], *P* = probability for directing the calls of the external flow to the device considered, *T* = mean service time, in the device, of a served call attempt [sec.], *Y* = intensity of the device traffic [Erl].

2.1.2. The Virtual Base Device Names

In the conceptual model each virtual device has a unique name. The names of the devices are constructed according to their position in the model.

The model is partitioned into service stages (**d**ialing, **s**witching, **r**inging and **c**ommunication).

Every service stage has branches (**e**nter, **a**bandoned, **b**locked, **i**nterrupted, **n**ot available, **c**arried), correspondingly to the modeled possible cases of ends of the calls' service in the branch considered.

Every branch has two exits (**r**epeated, **t**erminated) which show what happens with the calls after they leave the telecommunication system. Users may make a new bid (repeated call attempt), or to stop attempts (terminated call attempt).

In virtual device name construction, the corresponding bold first letters of the names of stages, branches end exits above are used in the order shown below:

Virtual Device Name = <BRANCH EXIT><BRANCH><STAGE>

A parameter's name of one virtual device is a concatenation of parameters name letter and virtual device name. For example, "*Yid*" means "traffic intensity in interrupted dialing case"; "*Fid*" means "flow (call attempts') rate in interrupted dialing case"; "*Pid*" means "probability for interrupted dialing"; *Tid* = "mean duration of the interrupted dialing"; "*Frid*" = "repeated flow call attempts' rate, caused by (after) interrupted dialing". All expression "device modeling the service of **r**epeated attempts after **i**nterrupted **d**ialing" is sometimes notated with {rid}.

2.1.3. The Paths of the Call attempts

We consider call attempts generated from terminals and correspondent to content and signaling terminal traffics. In this paper, we ignore the internal network signalization.

Figure 1 shows the paths of the call attempts, generated from (and occupying) the A-terminals in the proposed network traffic model and its environment. *Fo* is the intent rate of call attempts of one idle terminal; *M* is a constant, characterizing the BPP flow of demand attempts (*dem.Fa*). In this paper we assume *M = 0.*

In the model in Fig. 1, some of the blocks are numbered. These are Reference Points, e.g. the input point of call attempts into the model is virtual switch with Reference Point 2 (RP2). Comprising devices ("**a**", "**s**" and "**b**") are notated with graphical blocks.

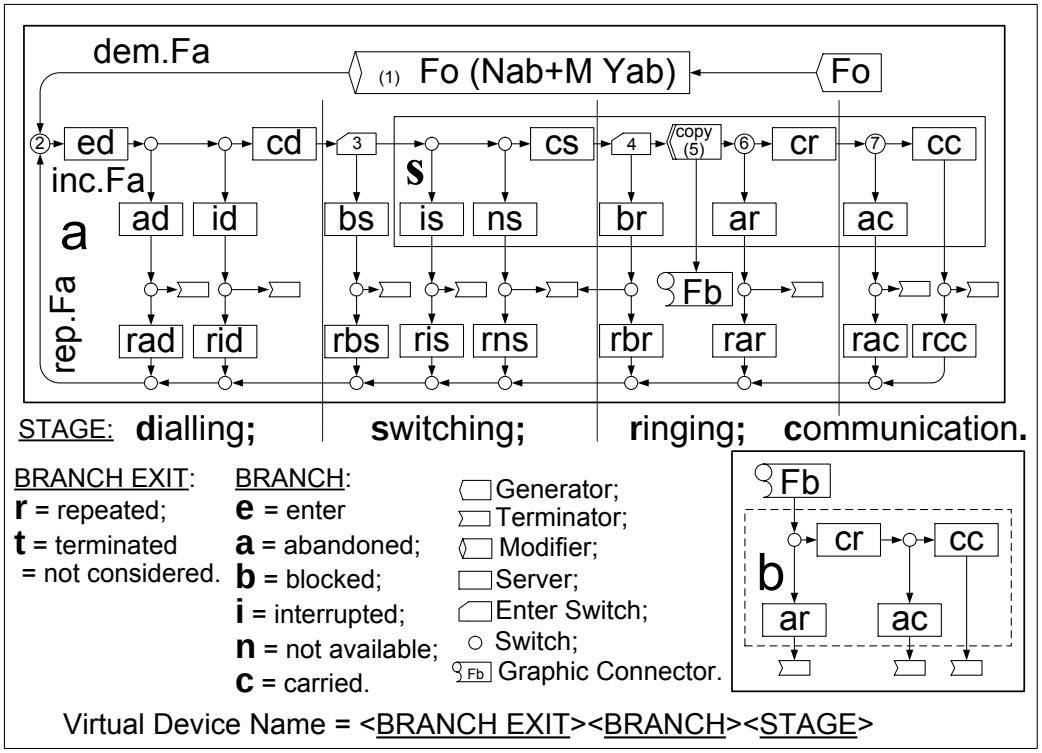


Figure 1. Conceptual model of the telecommunication system and its environment, including: the paths of the call attempts, occupying A-terminals (**a**-device), switching system (**s**-device) and B-terminals (**b**-device); base virtual device types, with their names and graphic notation. Some of switches, on the Carried Communication Branch are numbered as Reference Points.

2.2. Demand and repeated call attempts

2.2.1. The next definitions in [ITU E.600] are connected with demand traffic definition:

E.600, Definition 2.2: **call intent**: The desire to establish a connection to a user;

E.600, Definition 2.3: **call demand**: A call intent that results in a first call attempt;

E.600, Definition 2.4: **call attempt:** An attempt to achieve a connection to one or more devices attached to a telecommunications network;

E.600, Definition 2.5: **first call attempt**: The first attempt of a call demand that reaches a given point of the network;

E.600, Definition 2.6: **repeated call attempt; reattempt**: Any of the call attempts subsequent to a first call attempt related to a given call demand.

Following definitions above, we'll use the following shortenings, as definitions:

2.2.2. Definition **D2.2.1**: "**demand attempts**" for the first call attempts, from all considered call demands, that reach a given point of the network. (The calling rate of demand attempts, generating from calling (A) terminals and incoming in the Reference Point 2 into the network model presented in Fig. 1, we note with $dem.Fa$);

2.2.3. Definition **D2.2.2**: "**repeated attempts**" are call attempts subsequent to first call attempts related to all considered call demands, that reaches a given point of the network.

2.2.4. The calling rate of these repeated attempts, generating from, and occupying calling (A) terminals and incoming in the Reference Point 2 into the network model presented in Figure 1, we note with $rep.Fa$. In this notation, the rate of all, incoming in the network, attempts (Fa) is:

$$Fa = dem.Fa + rep.Fa \,. \tag{1}$$

The traffic of A-terminals, correspondent to Fa is notated with Ya .

2.3. Pie-Model concept

The Pie-Model concept is known through so called "pie-charts". In pie-models all call attempts incoming to the network, e.g. in (RP2) in Figure 1, are distributed into branches with beginning RP2 and with end – the base virtual device considered, inclusively. The all virtual devices in the branch are considered as one device, aggregating them. The name of this aggregative device is the name of the last device in the branch, following from suffix ".p", standing for "pie". For example, the branch "carried communication" has last virtual device "cc" in the normalized model. The corresponding aggregation "pie-device" is named "cc.p" and has main parameters *Pcc.p*, *Tcc.p* and *Ycc.p*. These parameters may be expressed easy by means of normalized devices, e.g. for holding time (Tcc.p) we have (normalized base devices have suffix ".n", standing for "normalized"):

$$\begin{aligned} &Pcc.p = (1 - Pad.n)(1 - Pid.n)(1 - Pbs.n)(1 - Pis.n) \\ &\qquad (1 - Pns.n)(1 - Pbr.n)(1 - Par.n)(1 - Pac.n)\,. \\ &Tcc.p = Ted.n + Tcd.n + Tcs.n + Tcr.n + Tcc.n\,. \\ &Ycc.p = Fcc.p\ Tcc.p = Fa\ Pcc.p\ Tcc.p = (dem.Fa + rep.Fa)Pcc.p\ Tcc.p. \end{aligned} \tag{2}$$

2.4. Ensue-Model concept

Let us consider the call attempts outgoing the device "carried switching" {cs.n} and incoming in RP4 in Figure 1. They have ensured continuation of their way in four possible branches. The set consisting of all base virtual devices, that may be occupied from the call attempts, after their leaving an appointed virtual device, we consider as an aggregative virtual device. The name of this aggregative device is the name of the appointed base device, following from suffix ".e", standing for "ensue"[2]). The parameters of this "ensue device" may be expressed by means of normalized devices, e.g. for ensued traffic intensity ($Ycs.e$) and ensued holding time ($Tcs.e$) we have:

$$\begin{aligned} &Ycs.e = Ybr.n + Yar.n + Ycr.n + Yac.n + Ycc.n\,. \\ &Tcs.e = Pbr.n\ Tbr.n + (1 - Pbr.n)\ Tb\,, \end{aligned} \tag{3}$$

where Tb is the occupation time of the B-terminals (see Figure 1).

[2] ensue = happen afterwards; occur as a result [COD 99].

3. Effective traffic related definitions

Let us consider the following ITU-T definitions, copied from [ITU E.600] and commented by the author:

3.1. E.600, Definition 2.10: **fully routed call attempt; successful call attempt:** A call attempt that receives intelligible information about the state of the called user.

Comment: User's states are, for example, "present" or "absent" and they are different from the terminal's states, e.g. "not available"/"available" (in mobile networks) and "busy"/"free".

3.2. E.600, Definition 2.11: **completed call attempt; effective call attempt:** A successful call attempt that receives an answer signal.

3.3. This definition refers rather the wanted terminal (responding equipment) state, because there are "calls on which an answer signal was received, although the called subscriber did not answer" [ITU E.422].

3.4. E.600, Definition 2.12: **successful call:** A call that has reached the wanted number and allows the conversation to proceed.

3.5. E.600, Definition 5.7: **effective traffic:** The traffic corresponding only to the conversational portion of effective call attempts.

3.6. Answer signal in effective call attempts (see E.600, 2.11 Definition) don't means effective conversation, because: the user may absent (not answering condition); another user may repots that the wanted user is not near by; the quality of connection may be unacceptable, etc., so, an "abandoned conversation" case may occur. The referring to the "successful call" (see 3.4. E.600, 2.12 Definition) is more appropriate.

In view of mentioned and other discrepancies in the ITU-T definitions above, it is necessary to give new editions of the most of them:

3.7. Definition **D3.1**: **fully routed call attempt:** A call attempt that receives intelligible information about the state of the called terminal.

3.8. This means that fully routed attempts are reached RP4 in Fig. 1. In other words, they have created traffic *Ycs.p* and ensure ensued traffic *Ycs.e* with mean holding time Tcs.e (see equations (3)).

3.9. Definition **D3.2**: **fully routed traffic:** The portion of the traffic corresponding to the fully routed attempts, from the moment they reach the called terminal.

3.10. Definition D3.2 is in conformity with definition D3.1 and the value of the fully routed traffic is *Ycs.e*.

3.11. **D3.3**: **successful call attempt:** A fully routed call attempt that receives intelligible information about the state of the called user.

3.12. The successful attempt is reached RP6 in Fig. 1.

3.13. **D3.4**: **successful traffic:** The portion of the traffic corresponding to the successful attempts, from the moment they occupy the called terminal.

3.14. The B-terminal occupation happens in RP6. Following the reasoning in definitions D3.1 and D3.2, the value of successful traffic is $Yar.n + Ycr.n + Yac.n + Ycc.n$.

3.14. Definition **D3.5**: **effective call attempt:** A call attempt that has reached the called terminal and allows the communication with a user to proceed.

3.15. The effective attempt is reached the RP7 on Fig. 1 and a conversation (abandoned or carried) is occurring.

3.16. Definition **D3.6**: **effective traffic:** The portion of the traffic corresponding to the effective call attempts, from the moment of beginning the communication with a user.

3.17. Following the reasoning in definitions D3.1 and D3.2, the value of effective traffic is $Ycc.n$, because the abandoned communication is difficult to be accepted as "effective". This is a strict definition. Some administrations might prefer a broad definition - to include the abandoned communications in definition also,

because effective traffic is known as "cost effective traffic". In this case, the effective traffic is $Ycr.e = Yac.n + Ycc.n$.

3.18. Definition D3.6 doesn't contradict to E.600, definition 5.7. It's only reformulated in order to reflect packet switching reality.

3.19. The (strict) effective attempts are moving only along the branch, corresponds to the {cc.p} (from RP2 to {cc.n}, see Figure 1) we call it "the *Carried Communication Branch*". It's parameters are: $Pcc.p, Tcc.p, Ycc.p$ (see equations (2)).

3.20. The "*ineffective attempts*" are all call attempts which are not effective.

4. Denial traffic concept

4.1. Every call attempt is generated with a will for success and it is moving in the Carried Communication Branch. In the normalized model every virtual switch, in the Carried Communication Branch, has two exits, so there are only two possibilities for a call attempt: 1) to continue its way towards {cc}; 2) to become ineffective and deflects of the effective way (to be failed in that switch point).

4.2. Definition **D4.1**: **denial traffic**: The portion of the traffic corresponding to the ineffective call attempts, created after call attempt deflection from the Carried Communication Branch.

4.3. In other words, denial traffic is served after the call attempt's failure in a point of Carried Communication Branch.

4.4. The denial traffic is real traffic, a part of ineffective traffic, corresponding to the ineffective call attempts. In the model in Figure 1, denial traffic is served in 8 devices: {ad.n}, {id.n}, {bs.n}, {is.n}, {ns.n}, {br.n}, {ar.n}, and {ac.n} (the including of {ac.n} in the list depends on the accepted effective traffic definition, see 3.17 above). The blocking is only a cause for denial traffic appearance.

4.5. The denial traffic concept is a next generalization step, following a generalization tendency in ITU-T: "End-to-end connection, party, and multi-party set-up failure, …, can occur from a lack of resources due to insufficient dimensioning or failure from other errors. End-to-end failure from a lack of resources due to insufficient dimensioning can be considered as a special case of the set-up failure probability." [ITU I.358].

5. Carried traffic concept

Let us consider a portion of the network on Figure 1, named "switching stage" (between the two vertical dotted lines). That portion of the network consists of four virtual devices: blocked switching {bs.n}, with traffic intensity $Ybs.n$; interrupted switching {is.n}, with traffic $Yis.n$; not switching (incorrect number, etc.), {ns.n} with traffic $Yns.n$; carried switching {cs.n} with traffic $Ycs.n$.

5.1. ITU concept for equivalent offered traffic [ITU E.501] is based on the carried traffic, defined in [ITU E.600]: E.600, Definition **5.1**: **traffic carried:** The traffic served by a pool of resources.

5.2. E.600, Definition 5.5 doesn't reflect the difference between the parts of the served traffic: carried and denial traffics. The distinguishing is necessary for service assessment and optimization. The ratio carried/served traffic is a good efficiency indicator.

5.3. Obviously $Yis.n$ and $Yns.n$ are denial traffics, following of attempt's termination and possible repeated attempts. These traffics are real, they load switching system and must be taken into considerations in dimensioning, but it is a little forcedly to name them "carried"[3], better leave for them the name "denial".

5.4. There is a big distinction between carried traffic ($crr.Ys$) in the cases of circuit switching and packet switching networks. In the circuit switching, the carried traffic coincides with the traffic corresponded to the carried

[3] carry: 1. support or hold up, esp. while moving; 2. convey with one from one place to another; 3. have on one's person (carry a watch); 4. conduct or transmit (pipe carries water; wire carries electric current)…[COD 99]

in the switching system attempts and the denial traffic ($Ycs.n + Ycs.e$, see equations (3)). In the packet switching networks, carried packets occupy switching system for a relative short time ($Tcs.n$ and correspondent $Ycs.n$).

The presented above gives grounds for a common, for circuit and packet switching networks, definitions, with illustrations based on the system presented on Fiure 1. These definitions are in force not only for switches.

6. Target traffic related definitions

In traffic engineering, many parameters have target values, which are interpreted as design objectives, see [ITU E.726].

6.1. The usual target value of blocking probability is zero (in our example, $trg.Pbs.n = 0$).

The traffic corresponding to this target value, in ITU-T recommendations is named “offered traffic”: ITU E.600

6.2. The natural generalization of the target traffics is demand traffic concept. Since nobody demands unproductive attempt’s occupation (and correspondent repeated attempts), the next definition is proposing:

Definition **D6.2.1: demand traffic**: The traffic that would be carried, in the overall network, from the demand attempts, if they all are served as current effective attempts. Consequently: $Pcc.p = 1$ and $rep.Fa = 0$.

6.2.1. Following definitions D2.2.1, D2.2.2, D6.2.1 and equations (2), putting $Pcc.p = 1$ and $rep.Fa = 0$ for the demand traffic ($dem.Ya$) of A-terminals, we receive:

$$dem.Ya = dem.Fa \;\; Tcc.p \; . \tag{4}$$

6.2.2. The only difficulty, in evaluation of the demand traffic, through measurements in the real systems, is the estimation of $dem.Fa$, because it is connected with determination of repeated attempts flow.

6.2.3. Demand traffic is a dream target value for users and in the traffic management. Together with effective, carried and served traffics, it is useful for overall network performance evaluation.

6.2.4. The phrase “demand traffic” is used three times, without any definition, in the ITU-T Recommendations; “traffic demand” is used 50 times.

7. Conclusions

7.1. In addition to normalize and pie-models [Poryazov 2001], ensue model and denial traffic concept are proposed, as a parts of a technique for presentation and analysis of overall network traffic models functional structure. This allows real network traffic considered to be classified more precisely and shortly.

7.2. The ITU-T definitions for: fully routed, successful and effective attempts and effective traffic are re-formulated in order to avoid some discrepancies and to reflect packet switching reality. Definitions for fully routed traffic and successful traffic are proposed, because they are absent in the ITU-T recommendations.

7.3. A definition of demand traffic (absent in ITU-T Recommendations) is proposed. Together with effective, carried and served traffics, it is useful for overall network performance estimation.

7.4. For each definition are appointed:

1) the correspondent part of the conceptual model graphical presentation;

2) analytical equations, valid for mean values, in a stationary state.

7.5. The ITU-T definitions are needed a careful over-thinking for accuracy and completeness, because ITU is the base body for common fundamental concepts acceptation. Most of discussed in this paper terms are absent in [ANSI 2001] and ETSI definitions.

Bibliography

[ANSI 2001] ATIS Committee T1A1 Performance and Signal Processing. ATIS Telecom Glossary 2000. T1.523-2001. Approved February 28, 2001, American National Standards Institute, Inc. (http://www.its.bldrdoc.gov/projects/devglossary/)

[Bohm&Jacopini, 1966] Bohm, C., G. Jacopini. Flow diagrams, Turing machines and languages with only two formation rules. Comm. ACM, 9 (1966), pp. 366-371.

[COD 99] Concise Oxford Dictionary 9th Edition, Oxford, 1999.

[Engset 1918] Engset, T., 1918. The Probability Calculation to Determine the Number of Switches in Automatic Telephone Exchanges. English translation by Mr. Eliot Jensen, Telektronikk, juni 1991, pp 1-5, ISSN 0085-7130. (Thore Olaus Engset (1865-1943). "Die Wahrscheinlichkeitsrechnung zur Bestimmung der Wählerzahl in automatischen Fernsprechämtern", Elektrotechnische zeitschrift, 1918, Heft 31.)

[ITU E.501] ITU-T Recommendation E.501: Estimation of Traffic Offered in The Network. (26th of May 1997).

[ITU E.526] ITU-T Recommendation E.526. Dimensioning a circuit group with multi-slot bearer services and no overflow inputs. (approved: 1993).

[ITU E.600] ITU-T Recommendation E.600: Terms and Definitions of Traffic Engineering. (Melbourne, 1988; revised at Helsinki, 1993).

[ITU E.726] ITU-T Recommendation E.726. Network grade of service parameters and target values for B-ISDN (13 March 2000).

[Iversen 2006] Iversen Villy B. Teletraffic Engineering and Network Planning. Lyngby, Denmark, June 20, 2006, pp. 354. http://oldwww.com.dtu.dk/education/34340/material/telenookpdf.pdf (Access 26.11.2006).

[Molnar 2006] Molnár, Sándor. Traffic models and teletraffic dimensioning. Chapter in: "Scientific Association for Infocommunications. Telecommunication Networks and Informatics Services". On-line book, Budapest, Hungary, 03.01.2006. (http://www.hte.hu/index.php?option=com_content&task=view&id=69&Itemid=102&lang=en)

[Poryazov 2001] Poryazov, S. A., 2001. On the Two Basic Structures of the Teletraffic Models. Conference "Telecom'2001" - Varna, Bulgaria, 10-12 October 2001 – pp. 435-450).

[Poryazov 2005] Poryazov, S. A. What is Offered Traffic in a Real Telecommunication Network? 19th International Teletraffic Congress, Beijing, China, August 29- September 2, 2005, Volume 6a, Liang X.J., Xin Z. H., V.B. Iversen and Kuo G. S.(Editors), Beijing University of Posts and Telecommunications Press, pp. 707-718.

[Poryazov, Saranova 2006] S. A. Poryazov, E. T. Saranova. Some General Terminal and Network Teletraffic Equations in Virtual Circuit Switching Systems. Chapter in: A. Nejat Ince, Ercan Topuz (Editors). "Modeling and Simulation Tools for Emerging Telecommunications Networks: Needs, Trends, Challenges, Solutions", Springer Sciences+Business Media, LLC 2006, pp. 471-505. Printed in USA, Library of Congress Control Number: 2006924687.

Author's Information

Stoyan Poryazov *– Institute of Mathematics and Informatics, Bulgarian Academy of Sciences, Acad. G. Bonchev Str., Block 8, 1113 Sofia, Bulgaria, phone: (+359 2) 979 28 46; fax: (+359 2) 971 36 49, e-mail: stoyan@cc.bas.bg*

Cyber Security

ICT SECURITY MANAGEMENT

Jeanne Schreurs, Rachel Moreau

***Abstract**: Security becomes more and more important and companies are aware that it has become a management problem. It's critical to know what are the critical resources and processes of the company and their weaknesses. A security audit can be a handy solution. We have developed BEVA, a method to critically analyse the company and to uncover the weak spots in the security system. BEVA results in security scores for each security factor and also in a general security score. The goal is to increase the security score Ss to a postulated level by focusing on the critical security factors, those with a low security score.*

***Keywords**: Security, Scan, Audit*

ACM Classification Keywords: Security

Introduction

As a consequence of the fast integration of technologies as Internet, Intranet, Extranet, Voice over IP and e-commerce, a companies ICT-infrastructure will move to more openness to the outside world and as a consequence will become more vulnerable for security threats. This offers lots of new opportunities but also creates new threats. That's why focus and responsibility concerning security become even more and more important. The Computer Crime and Security Survey 2005 shows that these are the 10 most frequent attacks or misuses: Virus, insider abuse of net access, laptop/mobile theft, unauthorized access to information, denial of service, abuse of wireless network, system penetration, theft of proprietary info, telecom fraud and financial fraud. Figures show that attacks come from inside as well as from outside the organisation and bring along large costs. Especially unauthorized access and laptop and mobile theft becomes a enormous expense for the companies during the last years. Because of these large costs, companies became more and more aware that they not only deal with a technical problem but also with a management problem. To tackle this management problem, it is quite important to know the ICT-security state your company is in.

ICT security management

Spending each year a certain amount on security measures is not enough. A company needs a total security approach. It is a must to know what are the critical resources and processes of the company and their weaknesses so the can be protected in the right way.

A solution to this is a security audit. A security audit is ideal to detect the weak spots in the ICT security state of the company. Based on the results of the audit, a security policy can be developed, adjusted to the company situation. A security audit can be used to analyse and describe the security level.

1. Security audit checklist

We have developed a security audit, called BEVA. BEVA is a method to analyse critically the company and to uncover the weak spots of the security system. It positions the company on point of the security aspects in the

different areas of business functions. We have developed a standard list that covers all aspects of security, structured in 10 domains being:

- Security policy
- Organization of information security
- Asset management
- Human resources security
- Physical and environmental security
- Communications and operations management
- Access control
- Information systems acquisition, development and maintenance
- Information secuirity incident management
- Business continuity management

Each of these areas consists of different security factors. The factors are in their turn tested on the basis of several subcriteia. Our list for the security factors is based on the standard ISO 17799. The 38 security factors are spread over the 10 domains, as set forward in the standard ISO17799 model.

For example you have the domain “access control” and in this domain you have the factors: requirements for access, management of user access, user responsibility, control of network access, control access to OS, control of access to applications and information and use of mobile infrastructure.

For each of the 38 factors, a number of subcriteria are formulated. We developed a list of questions, covering the subcriteria we created. The questions are partly based on the “checklists in information management” SDU publishers. (www.riskworld.net/7799-2.htm).

Security Factor Sfi	Importance	Sub Factor	Relevance/weight 1 to 4	Code question	Question	evaluation 1 to 4
Domain: Access control						
Sf20. Business requirements for access controlPremise	B	access control policymanagement	3	20.1	Is the access control policymanagement based on the business security requirements?	3
				20.2	Are aspects of logical and physical access control included?	3
				20.3	Is it clear for users and service providers which rules are applicable?	2
Sf21. User access management	C	registration of users	2	21.1	Is there any formal user registration and de-registration procedure for granting access to multi-user IS and services?	1
		privilege management	1	21.2	are privileges and allocated on need-to-use basis?	3
				21.3	are privileges only allocated after formal authorisation process?	1
		user password management	4	21.4	should the allocation and the reallocation of passwords be controlled through a formal management process?	3
				21.5	are the users asked to sign a statement to keep the password confidential?	1
		review of user access rights	3	21.6	does there exist a process to review user access rights at regular intervals?	4
					Does there exist a procedure to block the	

Figure 1: Questions audit checklist

2. The audit process and the calculation of security factor scores Sfi's and the security score Ss

To collect the information about the current security situation of the company, we start with the questioning of the key persons in the company using the audit checklist questionnaire.

The company determines which systems or processes are critical for them and connected with it, which security factors are important or relevant. An importance rate is given to the security factors from A (low importance) to E (high importance) (see figure 1).

In BEVA, we express the state of security into scores of the security factor (Sfi's). We do this for all the factors and in the end we give a general security score (Ss) over all security factors. We based our security analysis partly on the Marion-AP method.

To evolve to a security factor score, the key persons is asked to allocate a weight from 0 to 4 to the subcriteria of the security factors to indicate the relevance. Subsequently the evaluation starts and the list of questions is asked. Each question is given a score between 1 and 4. (see figure 2). The management team evaluates the company for all aspects on a one to four scale and at the same time measures the importance or relevance of all subfactors.

Security factor Sfi	Security Subfactor Ssfij	Relevance /weight 1 to 4 w(i,j)	Code question	evaluation 1 to 4	mean evaluation 1 to 4 eval(i,j)	Security factor score Sfis
Domain: Access control						
Sf20. Business requirements for access controlPremise	access control policymanagement	3	20.1	3	2,67	2,67
			20.2	3		
			20.3	2		
		3				
Sf21. User access management	registration of users	2	21.1	1	1	2,25
	privilege management	1	21.2	3	2	
			21.3	1		
	user password management	4	21.4	3	2	
			21.5	1		
	review of user access rights	3	21.6	4	3,5	
			21.7	3		
		10				
Sfis = sum [(w(i,j) * eval(i,j)] / sum w(i,j)						

Figure 2: Calculation of the Sf i's

When the questionnaire is completed, BEVA now calculates the security factor scores (Sf) being:

Sfi s = sum [eval (i,j) * w(i,j)] / sum w(i,k)

If all the factor scores are calculated also a general security score Ss is given:

Ss= sum [eval (1,38) * w(1,38)] / sum w(1, 38)

For example see factor 21 in the example: Sf21:[2*1 + 1*2 + 4*2 + 3*3,5]/10 = 2.25

Ss= in this example 2.66

Based on the evaluated questionnaire and the allocated weights, a realistic picture of the security situation of the company can be created as well general as by factor. The system BEVA creates a graphical output of the correlation diagram between these two variables measured for all aspects. Figure 3 shows the scores of all the security factors.

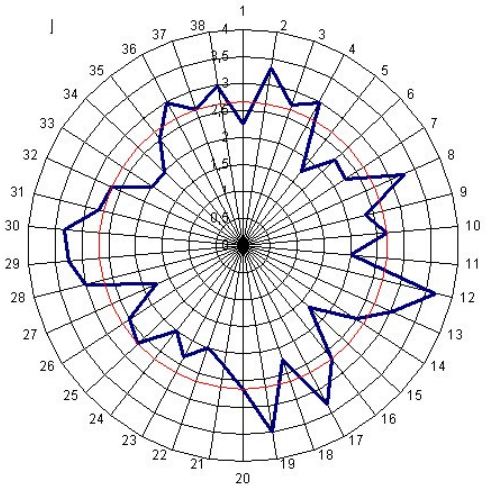


Figure 3: Graph of the security scores

The red line states Ss the general security score. The blue line connects the individual scores of the security factors. Security factors 1, 5, 6, 7, 9, 11, 14, 15, 18, 21, 22, 24, 26, 27, 33 and 34 score beneath the general security score.

Figure 4 combines the scores of the security factor with its importance. For example factor 33 scores low namely 2 but has importance A, low importance. Factor 34 scores also 2 but had importance E, high importance.

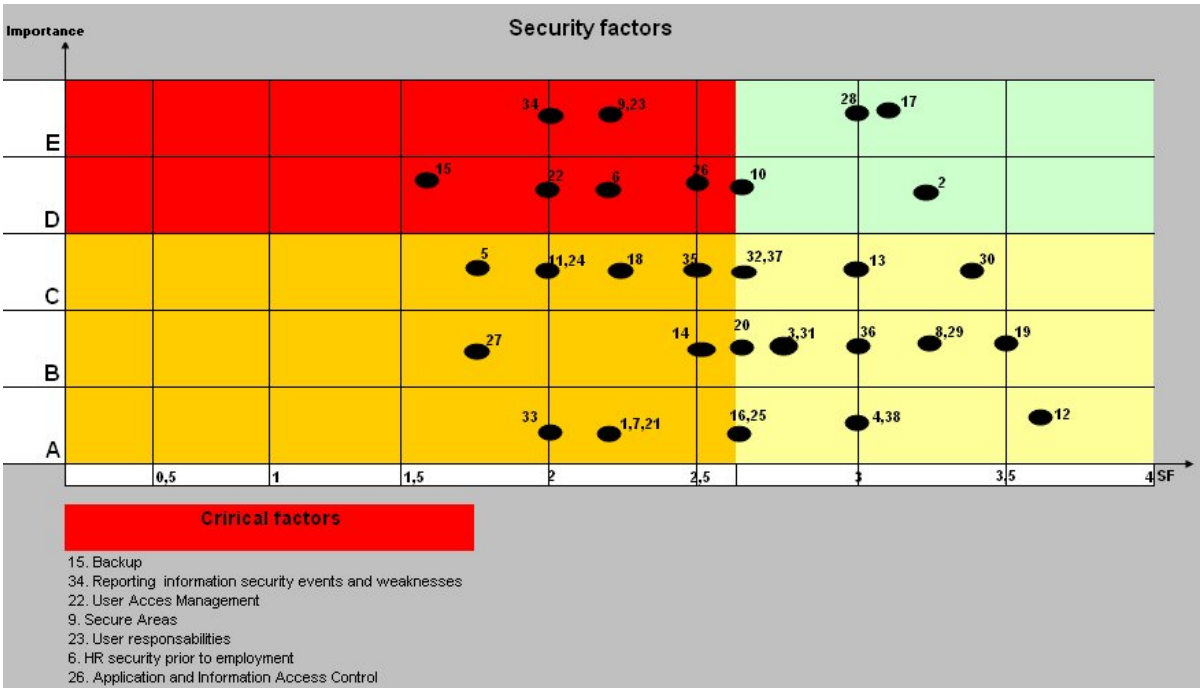
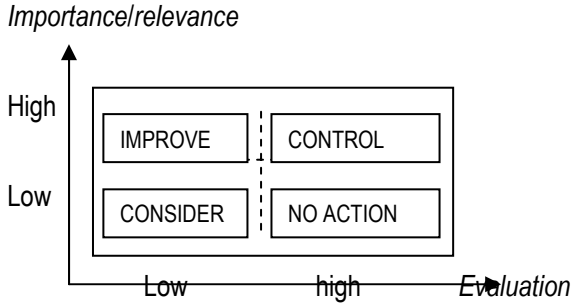


Figure 4: Graph of security factors and their importance

These differences are well stressed in this graphic. As you can see the ***red*** area highlights the security factors that score low and have a high importance. The factors lying in this area are critical and need immediate attention.

The ***green*** area is important and good secured. It is important to continue these actions and follow up these factors well. The *yellow* zone scores good but isn't that important, no action needs to be taken here. The less important factors that don't score well are situated in the *orange* zone. These factors need to be considered but probably with a small piece of the budget.



Now a clear view of the security situation is obtained. Feedback is given to the company and the evaluation states immediate points of action.

3. The occurrence of threats

The yearly organised CSI/FBI-study delivers the following probabilities for the threats:

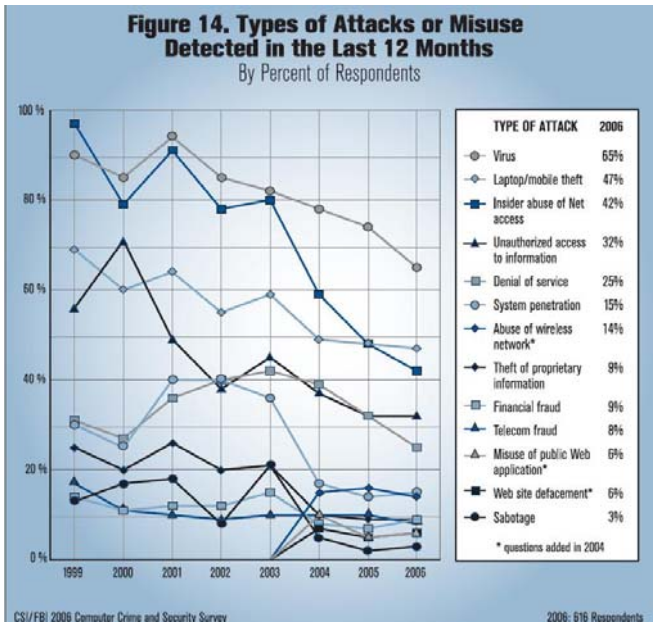


Figure 5: Threats and their occurance

Our final goal is to influence the occurrence of the threats, or the probability of the occurrence of them, by implementing selective security measures in the company. This will impact in the long run the security situation.

We must concentrate on the critical security factors, following the results of the audit. If the security factor is critical, than the threats linked with it have a critical risk too.

In figure 6 we figured out the relations between the threats and the security factors

Threat	1.1 Information security aspects of continuity mgt	2.1 Business requirements for access control	2.2 User access management	2.3 User responsabilities	2.4 Network access control	2.5 OS access control	2.6. Application en information access control	2.7 Mobile computing and telenetworking	3.1 Security requirements of IS	3.2 Correct processing in applications	3.3 Crypto- graphic controls	3.4 Security of systemfiles	3.5 Security in development and support processes	3.6 Technical vulnerability management	4.1 Secure areas	4.2 Equipment security	5.1 Compliance with legal requirements	5.2 Compliance with sequirity policies and standards	5.3 Information Systems audit considerations
Virus																			
Laptop/Mobile theft								x							x	x			
Insider abuse of net access					x						x								x
Unauthorised access to information		x	x	x					x	x	x	x	x	x				x	
Denial of Service - aanval	x																		
System penetration		x	x	x					x	x	x	x		x				x	
Abuse of wireless network		x	x		x			x			x								
Theft of proprietary information		x	x								x				x	x			
Financial fraude			x		x												x		
Telecom fraude								x											
Misuse of public web application							x												
Website defacement		x	x																
Sabotage	x					x							x	x	x	x			
Illigal software applications on the system (bots, trojan horses,…)													x						
Phishing																			
Misuse of the chat					x		x												
Pasword sniffing		x	x	x															
Exploiting the DNS server of the organisation																			

Threat	6.1 Prior to employment	6.2 During employment	6.3 Termination of change of employment	7.1 Information security policy	7.2 Internal organisation	7.3 External parties	8.1 Operational Procedures and Operations mgt	8.2 Third party service delivery management	8.3 System planning and acceptance	8.4 Protection against malicious and mobile code	8.5 Back-Up management	8.6 Network security management	8.7 Media Handling	8.8 Exchange of information	8.9 E-commerce	8.10 Monitoring	9.1 responsibility for assets	9.2 Information classification	10.1 Information incident management
Virus										x									
Laptop/Mobile theft			x	x										x			x		
Insider abuse of net access		x	x	x	x							x				x			
Unauthorised access to information			x	x		x	x	x				x		x		x			
Denial of Service - aanval												x			x				
System penetration			x			x	x					x			x	x			
Abuse of wireless network		x	x									x				x			
Theft of proprietary information			x			x	x	x			x		x	x	x			x	
Financial fraude	x		x			x	x	x							x	x			x
Telecom fraude	x	x	x			x	x	x							x	x			x
Misuse of public web application		x		x											x				
Website defacement			x			x				x		x							
Sabotage				x		x								x			x	x	
Illigal software applications on the system (bots, trojan horses,…)						x				x									
Phishing							x			x					x				
Misuse of the chat	x	x		x												x			
Pasword sniffing						x	x			x		x			x				
Exploiting the DNS server of the organisation										x		x			x				

Figure 6: Relation between threats and security factors

4. Security measures and follow up

A next step is to create a list of action points. Taking into account the stated security budget and the factors and their importance, an action plan is suggested. In the CSI study we can find the most used measures. A table is created were the most used measures are related with the threats they prevent.

Measures	Virus	Laptop/Mobile theft	Insider abuse of net access	Unauthorised access to information	Denial of Service - aanval	System penetration	Abuse of wireless network	Theft of proprietary information	Financial fraude	Telecom fraude	Misuse of public web application	Website defacement	Sabotage	Illigal software applications on the system (bots, trojan horses,…)	Phishing	Misuse of the chat	Pasword sniffing	Exploiting the DNS server of the organisation
Firewall				x	x	x					x	x		x				
AntiVirus Software	x													x			x	
AtiSpyware Software														x			x	
Server Based Acces control list				x		x		x										
Intrusion detection system				x		x		x			x	x						
Ecryption for data				x				x									x	
Reusable account system																	x	
Intrusion prevention system				x		x		x			x	x						
Log management software			x						x	x						x		
Application level firewall	x				x									x				
Smart card/ one time password token				x		x		x										
Specialized wireless security							x											
Training personeel		x					x						x					
Endpoint security client software	x													x				
Update server	x			x		x		x				x		x				

Figure 7: Relation between measures and threats

The action plan concerning security will be implemented, taking into account the weakest security factors and of course considering the budget.

After a period of approximately 3 months after implementing the security measures, a new security audit should be taken. The new security score Ss is calculated and compared to the stated aimed Security score using the security measures. If there are security factors that score too low, these should be investigated and adjusted.

Conclusion

The awareness that security is a management problem is everywhere present. It's critical to know what are the critical resources and processes of the company and their weaknesses. Our security audit is a handy solution. We have developed BEVA, a method to critically analyse the company and to uncover the weak spots in the security system. BEVA results in security scores for each security factor and also in a general security score. The goal is to increase the security score Ss to a postulated level by focusing on the critical security factors, those with a low security score. The results of the audit are an ideal start to do risk analysis.

Bibliography

[Shannon, 1949] C.E.Shannon. The Mathematical theory of communication. In: The Mathematical Theory of Communication. Ed. C.E.Shannon and W.Weaver. University of Illinois Press, Urbana, 1949.

[Jean-Marc Lamère] la sécurité informatique; Dunod : La méthode MARION (Méthodologie d'Analyse de Risques Informatiques Orientée par Niveaux) www.eisti.fr/~bg/COURSITACT/TXT/m_marion.txt

[Val Thiagarajan B.E,2005] Information Security Management ; BS ISO/ IEC 17799:2005 ; SANS Audit Check List: author:., M.Comp, CCSE, MCSE, SFS, ITS 2319, IT Security Specialist.

Security Management: A New Model to Align Security With Business Needs; Sumner Blount, CA Security Solutions ;August 2006

[Schreurs J, Moreau R.] ICT security management- ECEC 2007

(www.riskworld.net/7799-2.htm)

[Lawrence A. Gordon, Martin P. Loeb, William Lucyshyn and Robert Richardson] 2006 CSI/FBI-study about cybercrime: COMPUTER CRIME AND SECURITY SURVEY

https://event.on24.com/eventRegistration/EventLobbyServlet?target=registration.jsp&eventid=27372&sessionid=1&key=42F39B89EE0B30BA951711A5E7A98EDD&sourcepage=register

http://mediaproducts.gartner.com/gc/webletter/computerassociates/vol3issue3_risk/index.html

Authors' Information

Jeanne Schreurs *– prof. Business informatics, Universiteit Hasselt; gebouw D, Agoralaan, 3590 Diepenbeek, Belgium; e-mail: jeanne.schreurs@uhasselt.be*

Rachel Moreau *- Universiteit Hasselt; gebouw D, Agoralaan, 3590 Diepenbeek, Belgium; e-mail: Rachel.moreau@uhasselt.be*

ANALYSIS OF INFORMATION SECURITY OF OBJECTS UNDER ATTACKS AND PROCESSED BY METHODS OF COMPRESSION

Dimitrina Polimirova-Nickolova, Eugene Nickolov

Abstract*: In this paper a methodology for evaluation of information security of objects under attacks, processed by methods of compression, is represented. Two basic parameters for evaluation of information security of objects – TIME and SIZE – are chosen and the characteristics, which reflect on their evaluation, are analyzed and estimated. A co-efficient of information security of object is proposed as a combination of the coefficients of the parameter TIME and SIZE. From the simulation experiments which were carried out methods with the highest co-efficient of information security had been determined. Assessments and conclusions for future investigations are proposed.*

Keywords*: Information Security, File Objects, Information Attacks, Methods of Compression, Information Flows, Coefficient of Information Security*

ACM Classification Keywords*: D.4.6 Security and Protection: information flow controls*

Introduction

The development of information systems and technologies extends the necessity of processing, transferring and saving of volume sizable information flows, which are in network TCP/IP environment. These information flows,

in the form of file objects, are an object of non-stop attacks according to their information security, which determines the significant necessity for investigation of methods and means for their protection.

A general strategy for protecting file objects could include applying compression methods to objects to achieve decrease in volume size of information flow.

For the purposes of this paper the following reservation can be made: it is enough to investigate only the influence of compression methods on objects exposed to one or more attacks, as the difference in their behavior before and after the attacks when standard and not corporate (government) requirements are used is taken into consideration.

The Problem

The main aim of this paper is to make analysis of the information security of the file objects, found in TCP/IP environment, under information attacks, noting the influence of the compression methods.

The following tasks are set in reaching the aim:

1) to offer a methodology for evaluation of the information security of objects under attack and processed with a method of compression;

2) to set a co-efficient of information security of an object;

3) to find the methods of compression those reach the highest values of the co-efficient of information security.

For the aim of this paper the following work definitions are proposed [1], [2], [3]: 1) as information security we will note the protection of the information in an object from a random or purposeful access aimed at reading, transferring (coping), modifying or destroying the information in it; 2) as file object we will note the whole interconnected data or program records, saved under one name; 3) as information attack we will note an attack in connection with the content of the current information stream; 4) as method of compression we will note the procedure for data encoding aimed at shrinking their volume during the processes of transfer and storage.

1. METHODOLOGY OF EVALUATION OF THE INFORMATION SECURITY.

The methodology for evaluation of the information security of an object supposed to attack and processed with a method of compression will meet the following limitations:

- only the potential sets of attacks, methods and objects will be analyzed. These sets are made by stagely reduction of the known at the moment of study information attacks, methods of compression and file objects by using of matrix transformations. The stages of reduction of the multitudes are described in [4];
- the experiments are conducted at standard users', non-corporations' (governments') requirements;
- in order to simplify the computations the lossy methods of compression are exempt;
- in conducting the experiments for determining the co-efficient of information security, the objects used have equal or similar starting size.

Upon determining [4] the real relationships between attacks, methods and objects, studies and analysis can be made in the following three directions:

- ✓ evaluation of the success of the attack, made on an object processed with a method of compression;
- ✓ evaluation of the method of compression made on an object aimed at its defense from attacks it could be supposed to;
- ✓ evaluation of the security of an object supposed to an attack and processed with a method of compression.

This paper is aimed at the possibility to evaluate the information security of objects supposed to information attacks noting the influence of the methods of compression.

1.1 Setting the basic parameters for evaluating the information security.

The information security of an object can be determined as a quantitative value, which depends on several fundamental parameters, which can be represented as ratios of separate values before and after certain impact.

For the purposes of this paper considering the usage of standard users' requirements, not corporations' (governments') requirements it is enough to study and evaluate only the parameters *TIME* and *SIZE*, by marking the difference in the objects behavior before and after applying the method of compression.

The parameter *TIME* (*T*) reflects the evaluation of time for attack at an object before and after the influence of the method of compression. The parameter *SIZE* (*S*) reflects the evaluation of the size of an object before and after its processing with a method of compression.

1.2. Determining the characteristics which influence over chosen parameters.

After determining the main parameters, which will be analyzed and evaluated with regard to the information security of an object, is necessary to determine the basic characteristics, which have influence on the evaluation of the main parameters.

The basic characteristics, which have influence on the evaluation of the parameters BEFORE applying a method of compression to the object, are described below.

- For the evaluation of the parameter *TIME* the following characteristics can be taken into consideration: *preliminarily time of the attack* and *the time of the attack to process the object*;
- For the evaluation of the parameter *SIZE* will pointed characteristics depending of the category to which file objects belong to. Two basic categories are: DIRECTLY USED (these are objects, which have to be used directly) and NON-DIRECTLY USED (these are objects, requiring secondary processing to become directly used):
  - the characteristics, which have influence on the evaluation of the parameter *SIZE* for objects belonging to DIRECTLY USED category, are: *bits of information*, *entropy of the message*, *information redundancy and message's length*.
  - the characteristics, which have influence on the evaluation of the parameter *SIZE* for objects belonging to NON-DIRECTLY USED category, are: *image size*, *resolution*, *bit depth* (for image objects); *duration*, *bits-per-second*, *frame rate*, *frame size* (for video objects); *sample size*, *sample rate*, *number of channel* (for audio objects).

The basic characteristics, which have influence on the evaluation of the parameters AFTER applying a method of compression to the object, are described below.

- For the evaluation of the parameter *TIME* the following characteristics can be taken into consideration: *preliminarily time of the attack*, *time needed to break the defense mechanism* and *the time of the attack to process the object*.
- For the evaluation of the parameter *SIZE* will be specified characteristics, depending of the method of compression applied over the object:
  - when statistical methods of compression are applied, the characteristics, which have influence on the evaluation, are: *entropy of the message*, *information redundancy*, *level of compression*, *bits of information*, *size of the model for decompression*;
  - when dictionary methods of compression are applied, the characteristics, which have influence on the evaluation, are: *size of the dictionary*, *entropy of the message*, *coincidence of words in the dictionary*, *level of compression*;
  - when image methods of compression are applied the characteristics, which have influence on the evaluation, are: *bit depth*, *image size*, *number of pixel repetitions*, *level of compression*, *number of consecutive pixels*, *correlation of neighborhood pixels*;
  - when audio methods of compression are applied the characteristics, which have influence on the evaluation, are: *sample size*, *sample rate*, *number of blocks of consecutive samples*.

1.3. Determining the evaluations of the characteristics, which have influence on the general valuation of the respective parameter.

Each characteristic is necessary to be evaluated with respect to the information security of an object under attack before and after applying a method of compression. To determine these evaluations is taken into consideration additional factors, which have influence on the evaluation of the respective characteristic. After that is necessary to examine each characteristic by providing simulation experiments, which will determine the relationship between the obtained after the examination result and the evaluation of the characteristic with respect to the information security of an object (for example: the increasing of the time of an attack to manipulate an object, increases object's information security, which leads to higher valuation of this characteristic; the increasing of the size of the model for decompression decreases the possibility for better compression of the object, which leads to faster restoration in its original state, respectively to faster braking the protection mechanism of the object as a mean of method of compression, that means lower valuation of this characteristic with respect to the information security of this object). At the end the valuation (*V*) of the respective characteristic is determined.

1.4. Setting a weighted co-efficient for each characteristic.

The weighted co-efficient (*W*) determine the level of influence which each valuation of the respective characteristic have influence on the general evaluation of the parameter to which it belongs to. For determining the weighted co-efficient of the characteristic is used the AHP (Analytic Hierarchy Process) method [5], which consists of four basic stages: 1) determining the characteristics which have to be evaluated; 2) arranging the chosen characteristics in a matrix; 3) comparing each couple of characteristics by preliminarily selected measurement scales for evaluation; 4) determining the respective weights of the characteristics by consecution of mathematical operations.

1.5. Estimating the general evaluation of the respective parameter.

The estimating of the general evaluation of the parameter consists of the following stages: 1) determining the evaluation of the characteristics, which have influence on the basic evaluation of the selected parameter $V_{(\text{charact}_{\cdot n})} = [0 \div 1]$, where *n* is the number of the characteristics; 2) setting the weighted co-efficient of each characteristic $W_{(\text{charact}_{\cdot n})}$, like $\sum_{i=1}^{n} W_i = 1$; 3) determining the evaluation of the parameter as $V_{(\text{parameter}_1)} = \sum_{i=1}^{n} \left(V_{(\text{charact}_{\cdot i})} . W_i \right)$ (Fig. 1).

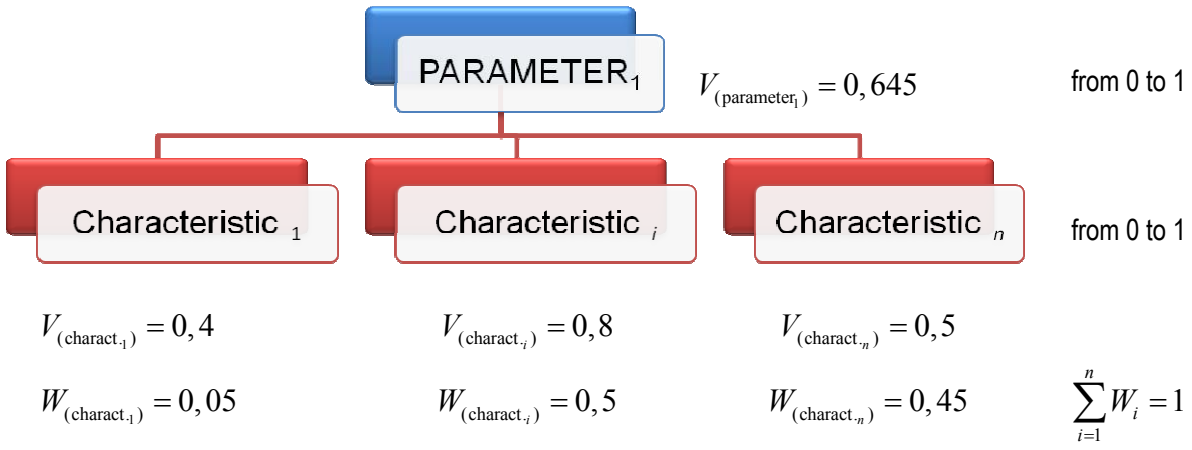


Fig. 1 Determining the evaluation of the parameter

2. DETERMINING THE CO-EFFICIENT OF INFORMATION SECURITY.

A co-efficient of information security is compounded to analyze the information security of the objects. It is presented as a variable, formed from the examined above parameters *TIME* and *SIZE*, reflecting the condition of the object before and after applying methods of compression.

2.1. Determining the co-efficient of information security of an object in regard to the evaluation of the parameter *TIME* ($K^{IS(T)}$).

The parameter TIME can be determined as a co-efficient of information security for evaluating the time ($K^{IS(T)}$):

$$K^{IS(T)} = \frac{\Delta V_{(T)}}{V'_{(T)}}$$

where $\Delta V_{(T)} = V''_{(T)} - V'_{(T)}$ like $V'_{(T)}$ is the determined informational security of an object in regard to the time before applying the method of compression and $V''_{(T)}$ is the determined information security of an object in regard to the time after applying the method of compression.

In each relationship attack—method—object is determined ($K^{IS(T)}$):

$$K_z^{IS(T)} = f\left(a_i, m_j\right) \quad \text{for each } o_f$$

where $a_i \in A_{pot}\left\{a_1, a_2, ..., a_i, ..., a_p\right\}$, $m_j \in M_{pot}\left\{m_1, m_2, ..., m_j, ..., m_q\right\}$, $o_f \in O_{pot}\left\{o_1, o_2, ..., o_f, ..., o_r\right\}$, and the index ***z*** is changing within the bounds of the formula a_p, m_q and o_r.

2.2. Determining the co-efficient of information security of an object in regard to the evaluation of the parameter *SIZE* ($K^{IS(S)}$).

The parameter *SIZE* can be determined as a co-efficient of information security for evaluating the size ($K^{IS(S)}$):

$$K^{IS(S)} = \frac{\Delta V_{(S)}}{V'_{(S)}}$$

where $\Delta V_{(S)} = V''_{(S)} - V'_{(S)}$ like $V'_{(S)}$ is the determined informational security of an object in regard to the size before applying the method of compression and $V''_{(S)}$ is the determined information security of an object in regard to the size after applying the method of compression.

In each relationship attack—method—object is determined ($K^{IS(S)}$):

$$K_z^{IS(S)} = f\left(a_i, m_j\right) \quad \text{for each } o_f$$

where $a_i \in A_{pot}\left\{a_1, a_2, ..., a_i, ..., a_p\right\}$, $m_j \in M_{pot}\left\{m_1, m_2, ..., m_j, ..., m_q\right\}$, $o_f \in O_{pot}\left\{o_1, o_2, ..., o_f, ..., f_r\right\}$, and the index ***z*** is changing within the bounds of the formula a_p, m_q and o_r.

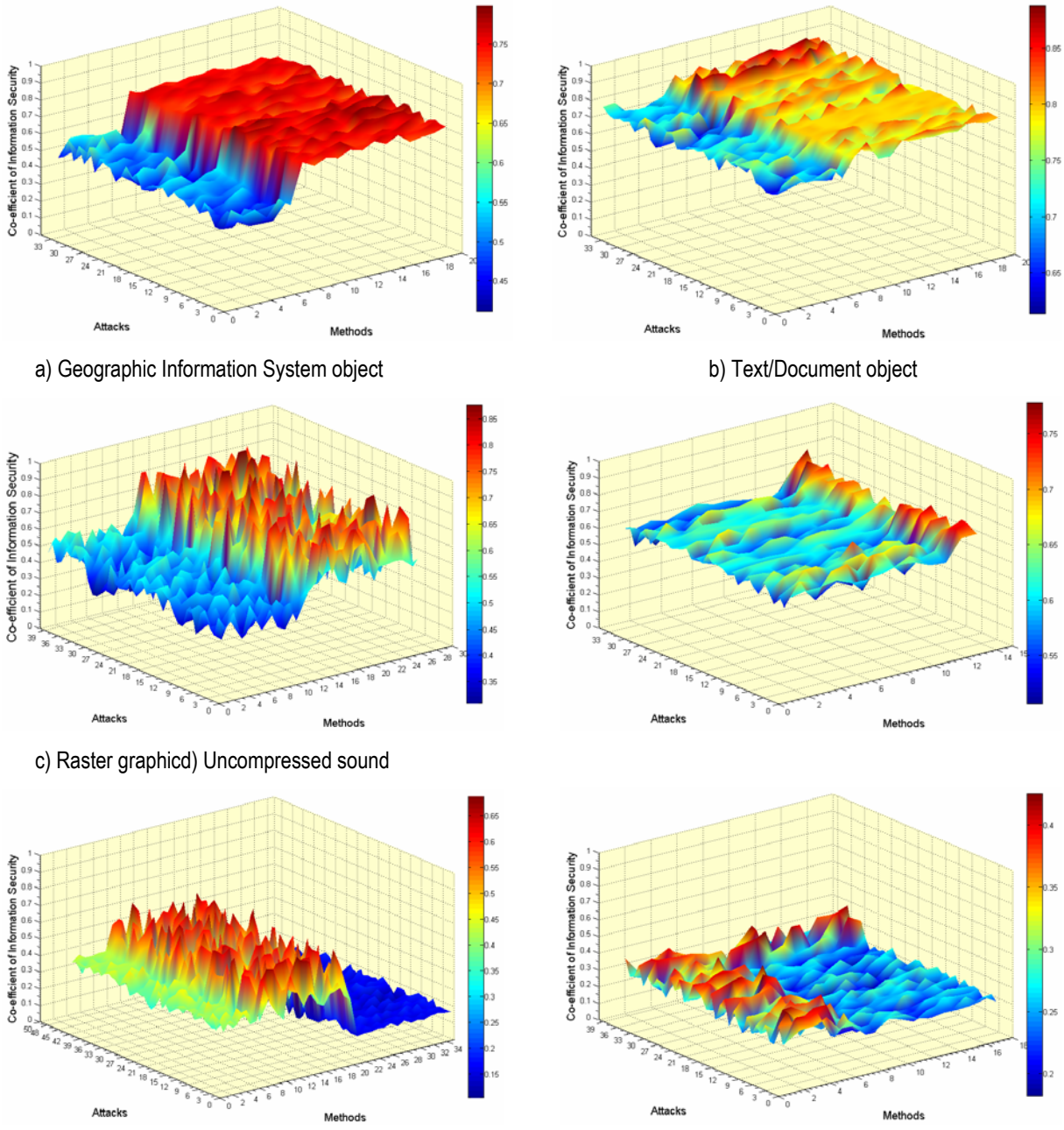
2.3 Determining the co-efficient of information security of an object as the interconnection between the co-efficients for evaluating the two parameters (K^{IS}).

The coefficient of information security (K^{IS}) can be determined as a combination of the coefficients for evaluation of the parameters TIME and SIZE:

$$K_z^{IS} = K^{IS(T)} + K^{IS(S)}$$

where ***z*** is changing within the bounds of the formula a_p, m_q and o_r.

Graphic interpretation for determined sizes of the K^{IS} for most frequently used file objects is shown on Figure 2a), b), c), d), e), f).



a) Geographic Information System object

b) Text/Document object

c) Raster graphicd) Uncompressed sound

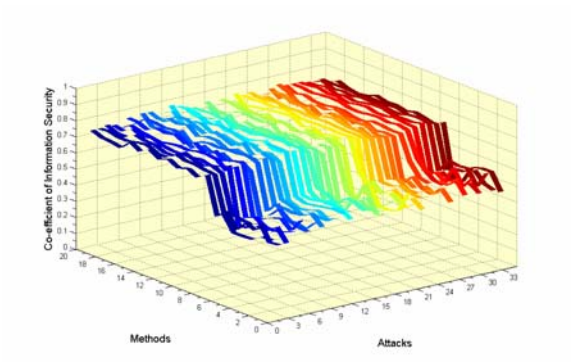
e) Dynamic web pagef) Source code

Fig. 2 Graphic interpretation for determined sizes of the co-efficient of information security for different file objects

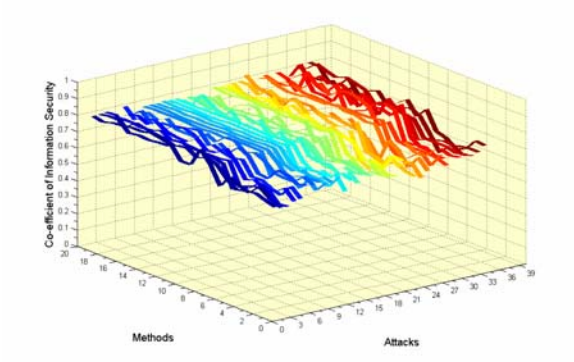
3. METHODS WITH THE HIGHEST VALUES OF THE CO-EFFICIENT OF INFORMATION SECURITY.

3.1. Determining the methods with the highest values of the co-efficient of information security for each object for the given attack.

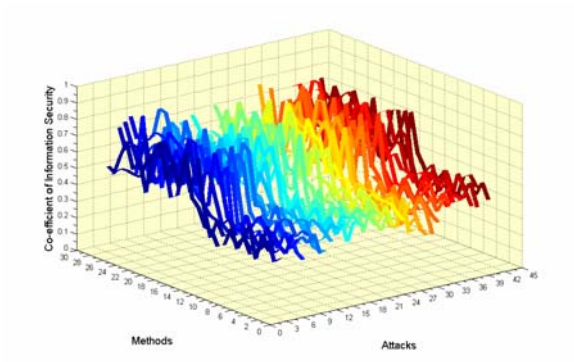
After determining K^{IS} for each object we can determine which is the method with the highest value of K^{IS} for the given object and attack. On fig 3a), b), c), d), e), f) we can see a graphical presentation of the change in the co-efficient of information security for given objects in regard to given attacks, determined after applying the given methods for compression.



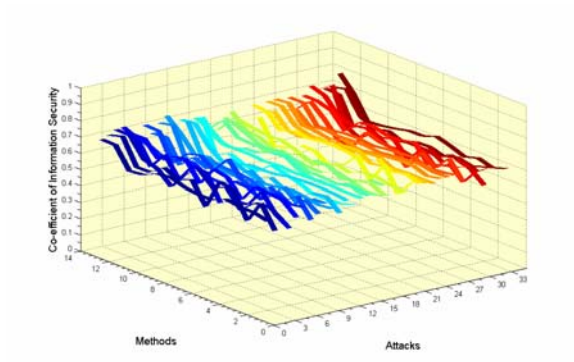
a) Geographic Information System object

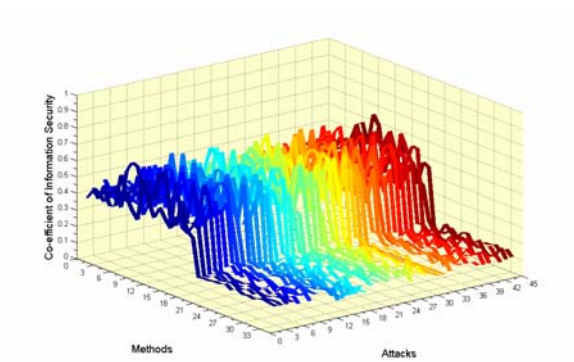


b) Text/Document object



c) Raster graphicd) Uncompressed sound





e) Dynamic web pagef) Source code

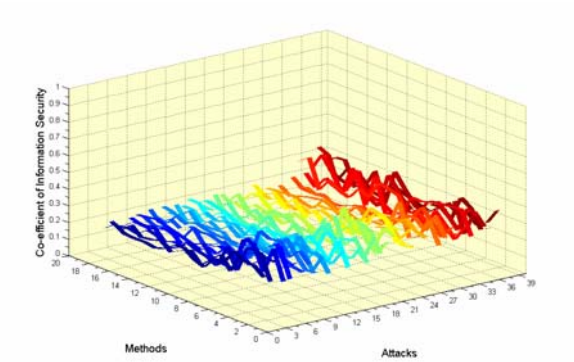


Fig. 3 Distribution of the co-efficient for informational security for given object and attack, when a given method of compression is applied

3.2. Determining the methods with the highest values of the co-efficient of information security for all objects in regard to given attacks.

Thus for each object can be set up a group of methods of compression, reaching the highest values of K^{IS} with respect to all attacks on which the object can be exposed.

Derived from this particular scientific work the results are the basis for further research in connection with the opportunity to determine the method of compression, which will have the lowest risk in regard to the information security for the given object and attacks, for which it can be applied.

Assessments

1) Parameters used for determining *TIME* and *SIZE* are sufficient for researching information security of objects and computer systems and networks for consumer, not governmental (corporate) needs.

2) Evaluation in regard to the selected objects, which were processed with methods of compression, is positive and the allowances do not affect the derived result.

3) In regard to the methods of compression we used the assessment is positive and the above mentioned experiments can be used and tailored to other methods of compression.

4) We can conclude, looking at the experiments, that with the decreasing size of an object after compression, time needed for an attack to complete its work over the object will increase.

5) As with the co-efficient of information security the best results were obtained from data objects, processed with dictionary methods of compression, and the worst results were obtained with the graphics objects processed with statistical methods of compression.

6) From all 59 methods of compression, 13 of them gave us the highest value of the co-efficient of information security of the object. They are from the group of dictionary methods and image methods of compression.

Bibliography

[1] Elena Ferrari, Bhavani M. Thuraisingham, *Web and Information Security*, IRM Press, 2006, ISBN: 1-59140-589-0, p. 215

[2] http://www.answers.com/file

[3] David Salomon, *Data Compression: The Complete Reference*, Springer, 2006, ISBN: 1846286026, p.1-9

[4] Dimitrina Polimirova, Eugene Nickolov, Cecko Nikolov, *Investigating The Relations Of Attacks, Methods And Objects In Regard To Information Security In Network TCP/IP Environment*, Proceedings of the First Workshop "Cyber Security", 28-29 June 2006, Varna, Bulgaria, ISBN-10: 9541600417, p. 20-27

[5] Hubert Hasenauer, *Sustainable Forest Management: Growth Models for Europe*, Springer 2006, ISBN: 9783540260981 p.267-269

Authors' Information

PhD Student, ***Dimitrina Polimirova,*** *Research Associate, National Laboratory of Computer Virology, Bulgarian Academy of Sciences, Phone: +359-2-9733398, E-mail: polimira@nlcv.bas.bg.*

Prof. ***Eugene Nickolov,*** *DSc, PhD, Eng, National Laboratory of Computer Virology, Bulgarian Academy of Sciences, Phone: +359-2-9733398, E-mail: eugene@nlcv.bas.bg.*

КОМПЛЕКСНАЯ СИСТЕМА ЗАЩИТЫ РАСПРЕДЕЛЕННЫХ ИНФОРМАЦИОННЫХ СИСТЕМ, УПРАВЛЯЕМЫХ МЕТАДАННЫМИ

Денис Курилов, Людмила Лядова

***Abstract**: Предлагается описание архитектуры и подходов к реализации многоуровневой комплексной системы защиты динамически настраиваемых распределенных информационных систем, основанных на использовании метаданных. Исследуются возможности различных механизмов защиты. Описанная в работе система защиты представляет собой многоуровневый комплекс, построенный на базе мультиагентной системы, объединяющей функциональные возможности современных систем обнаружения вторжений (IDS – intrusion detection systems) с механизмами защиты структуры и программной логики информационных систем.*

***Keywords**: адаптируемые информационные системы, механизмы защиты, метаданные, мультиагентная система.*

***ACM Classification Keywords**: D.2 Software Engineering: D.2.0 General – Protection mechanisms; K.6 Management of Computing and Information Systems: K.6.5 Security and Protection – Authentication, Insurance, Invasive software (e.g., viruses, worms, Trojan horses), Unauthorized access (e.g., hacking, phreaking); I.2 Artificial Intelligence: I.2.11 Distributed Artificial Intelligence – Multiagent systems.*

Введение

Характерными особенностями современных информационных систем (ИС), разрабатываемых для различных предметных областей, влияющими на степень их защищенности, надежности функционирования, являются [1]:

- *Сложность*. С возрастанием сложности растёт количество уязвимостей, обнаружение и устранение которых затруднено.
- *Открытость и интегрируемость*. Открытость информационных систем и их интегрируемость, реализация различных механизмов взаимодействия с внешними ИС является потенциальным источником уязвимостей.
- *Адаптируемость и расширяемость*. Возможность гибкой настройки ИС на конкретные условия работы и потребности пользователей, расширения функциональности сторонними разработчиками также создаёт опасность внедрения вредоносного кода.
- *Распределённость*. Возможность взаимодействия подсистем ИС через сеть создаёт дополнительные угрозы безопасности, такие как атака на серверные компоненты ИС с использованием клиентских компонентов.

Существующие на сегодняшний день методы защиты не позволяют защитить динамически настраиваемые информационные системы, функционирующие в распределенной среде, в комплексе: они либо защищают программный код, либо данные ИС. Возможности настройки ИС основаны на использовании средств динамической реструктуризации баз данных (БД); автоматической генерации и настройки пользовательского интерфейса; средств генерации запросов и отчетов; средств управления бизнес-процессами; средств подключения программных компонентов, созданных сторонними разработчиками. Все это делает еще более важной проблему защиты ИС, их ресурсов и программного обеспечения (ПО) от несанкционированного доступа и распространения, так как, используя предоставленные в их распоряжение средства, недобросовестные пользователи, обладающие достаточной квалификацией, фактически могут использовать возможности технологий создания адаптируемых систем в своих целях.

Предлагаемый в данной статье подход рассматривает программное обеспечение ИС, функционирующее в распределенной среде, как цельный, неделимый программный продукт, который необходимо защищать именно в комплексе с данными и метаданными, описывающими ИС. Под комплексной защитой понимается защита данных и программного кода ИС и выбор наилучшей схемы лицензирования.

Устоявшийся взгляд на ИС как на сложные программные комплексы, компоненты которых установлены в узлах сети и взаимодействуют посредством передачи информации через линии связи, породил соответствующий подход к организации защиты ИС. Основа этого подхода – в реализации *различных механизмов защиты узлов сети от несанкционированного доступа и использования их ресурсов* (в частности, от вредоносного программного обеспечения, включающего различные вирусы и троянские программы), а также *защиты каналов взаимодействия в сети* при помощи различных технических и программных средств (например, экранирования, анализа сетевого трафика и т.п.). Такой подход к организации защиты вполне применим и оправдан, если речь идёт о защите программных систем, исполняемых на отдельных компьютерах или в рамках небольшой сети. В случае же распределенных ИС, масштаб которых выходит за рамки отдельного компьютера или небольшой локальной сети, проявляются существенные недостатки традиционного подхода, среди которых можно выделить следующие:

- *Сложность поддержания защиты ИС на должном уровне*. В силу ориентации подавляющего большинства коммерческих средств поддержки информационной безопасности (таких как Symantec Intruder Alert и NetProwler, семейство систем ISS RealSecure и др.) на сигнатурный метод выявления атак, требуется их постоянное обновление на всех узлах ИС. Предлагаемый в данной работе подход к организации системы защиты не требует отказа от средств такого рода, но значительно снижает зависимость от них.
- *Существенная уязвимость к новым типам атак*, например, основанных на выявлении и использовании ранее не применявшихся для этих целей уязвимостей самой ИС, операционной системы и сетевого программного обеспечения. Причина этого – опять же в преобладании сигнатурных методов анализа на современном рынке систем поддержки информационной безопасности.
- *Практически полное отсутствие защиты от атак, разработанных специально для взлома данной ИС*, основанных, в частности, на уязвимостях и ошибках в программной реализации её модулей. Как следствие – необходимость в дополнительных средствах защиты от атак, проводимых «изнутри» ИС, с использованием предварительно взломанных модулей.
- *Необходимость в дополнительных средствах контроля входящего и исходящего потоков информации*. Под контролем входящего потока подразумевается защита от спама, фишинга, вредоносных ad-ware и других аналогичных внешних угроз. Под контролем исходящего потока понимается сканирование всей исходящей информации, передаваемой во внешние системы, на предмет нахождения в ней защищенной корпоративной информации.
- *Сложность обеспечения достаточного уровня аутентификации пользователей*. В крупных ИС, содержащих защищённые данные и сервисы, стандартный механизм аутентификация пользователей, основанный на проверке знания некоторого секретного ключа, является малоэффективным в виду возможности утечки информации.

Причина многих недостатков традиционного подхода к обеспечению безопасности при применении его к защите распределённых ИС заключается в том, что подавляющее большинство средств защиты *не использует информацию о структуре и семантике защищаемой ИС*.

Архитектура системы защиты

В данной работе предлагается принципиально иной взгляд на построение систем защиты. Информационная система рассматривается не как совокупность взаимодействующих при выполнении

своих функций вычислительных узлов, а как *совокупность сервисов*, предоставляемых компонентами ИС, реализуемых на базе нескольких взаимосвязанных узлов сети, требующая защиты в целом, а не на уровне отдельных структурных единиц и каналов связи. Оправданность такого подхода особенно проявляется на фоне тенденции расширения функциональных возможностей крупных ИС, имеющей целью обеспечить всю необходимую для повседневной работы пользователей функциональность в рамках одной интегрированной ИС (например, выполнение типовых операций по вводу и редактированию данных в рамках бизнес-процессов, автоматизируемых компонентами ИС, отправка и получение писем, обработка данных и генерация отчетов и др.).

Авторами предлагается подход к проектированию *комплексной системы защиты динамически адаптируемых распределенных информационных систем*, *основанных на использовании метаданных*, описывающих все стороны функционирования ИС. Архитектура комплексной системы защиты включает несколько уровней и объединяет различные механизмы защиты, интегрированные с защищаемой информационной системой на основе использования метаданных, описывающих эту ИС.

Предлагаемая система защиты представляет собой многоуровневый комплекс, построенный на базе *мультиагентной системы* (МАС). Комплекс сочетает в себе функциональные возможности современных систем *обнаружения вторжения* (IDS – *intrusion detection systems*) с механизмами *защиты структуры и программной логики* ИС.

Основой системы защиты является *распределённая МАС*. Сообщество агентов системы является *закрытым* и *защищенным* от злоумышленного влияния извне как при помощи механизмов, реализуемых на уровне собственной безопасности, так и за счёт самого способа организации работы агентов. Каждый *агент является независимой сущностью, скрыто функционирующей в рамках защищаемой системы*. Информация об агентах не содержится нигде в системе за пределами сообщества агентов, оперирующих в системе и потоках, в которых они исполняются. Сокрытие осуществляется двумя основными способами: исполнение агентов в потоках, *скрытых на уровне ядра ОС при помощи драйвера защиты*, и исполнение агентов в потоках защищаемой системы при помощи *механизма переключения контекста потока*, активно используемого самой ОС [2]. Все агенты МАС делятся на два класса: *агенты-аналитики* и *агенты-сенсоры*. Агенты-аналитики – это интеллектуальные агенты, построенные на базе архитектуры InteRRaP, относящейся к классу многослойных архитектур с вертикальным делением на слои [4]. Каждый слой реализует определённый тип взаимодействия агента со средой (областью системы, в которой оперирует данный агент). Информация о текущем состоянии системы передаётся с низших слоёв на высшие, управление – с высших на низшие.

Структура агентов представлена тремя слоями:

- *Слой поведения* отвечает за реализацию реактивности, поведения в режиме реального времени. Поле ответственности данного слоя – *принятие решений в условиях шаблонных ситуаций*, примерами которых могут служить регистрация пользователя, установка соединения с удалённым узлом сети или попытка взлома известного типа, сигнатура (т.е. сценарное описание) которого уже содержится в системе.
- *Слой планирования (локального планирования)* – это реализация *когнитивной парадигмы* построения МАС. При передаче информации со слоя поведения на слой планирования производится вывод по базе знаний агента, целью которого является определение класса текущей ситуации и выбор адекватного шаблона поведения (набора реакций на изменения состояний среды) с целью дальнейшего его применения на слое поведения.
- *Слой коммуникации (коллективного планирования)* – отвечает за реализацию *механизмов коммуникации агентов*. Данный слой представляет реализацию возможности принятия решения на основе данных, подготовленных другими агентами системы, отвечает за организацию командной работы.

Представление знаний, используемых агентами для определения ситуации, осуществляется в рамках *фреймовой* парадигмы [3], правила принятия агентами решений представляются *продукционно*.

Агенты-сенсоры служат для сбора данных о текущем состоянии системы защиты, самой ИС и её модулей, а так же сети, в рамках которой функционирует ИС. Данные агенты реализуются на основе реактивного типа архитектуры (Reactive architecture [4]), и служат для сбора статистических данных, регистрации событий и выявление аномалий на базовом уровне.

Взаимодействие агентов основано на модели «заказчик-подрядчик» (Contract Net [4]), предполагающей решение различных задач посредством направления их на выполнение наиболее подходящим для этого агентам. Такой выбор обусловлен наличием у данной модели ряда преимуществ, обеспечивающих наиболее полное её соответствие требованиям решаемой задачи:

- Наличие у каждого агента системы функциональности, позволяющей выполнять некоторые задачи без привлечения других агентов системы (высокая степень самостоятельности агентов).
- Малый промежуток времени между возникновением задачи и началом процесса её решения.
- Малая вероятность неверного решения задач в связи с их назначением наиболее компетентным агентам, содержащим всю необходимую функциональность для их решения.
- Низкие накладные расходы в связи с отсутствием необходимости постоянного анализа каждым агентом текущего состояния среды.
- Высокая эффективность организации контроля системы, следующая из возможности организации агентов в иерархические структуры.

Уровни и механизмы защиты

Защита структуры и программной логики ИС является необходимым элементом защиты, неоправданно игнорируемым современными системами поддержки информационной безопасности по уже указанной выше причине – отсутствия в них информации о семантике защищаемой ИС. *Защита структуры ИС* необходима для пресечения возможности подмены серверных и клиентских компонентов ИС специально подготовленными для проведения дальнейшего взлома программными модулями, а также для более эффективной организации защиты от несанкционированных подключений к сервисам ИС. *Защита программной логики* обеспечивает пресечение попыток несанкционированной модификации программного кода ИС, целью которых может быть внесение ошибок и создание потайных каналов (*back doors* [5]) для организации проведения атак.

Информация о семантике защищаемой ИС представляется при помощи *иерархической трехслойной модели*, в полной мере описывающей все важные с точки зрения организации защиты аспекты ИС.

Вся информация о функционировании ИС, ее предметной области [6] распределяется по трём слоям модели защиты системы $S = (Str, Ev, Msg)$, где:

- *Слой структур* Str содержит описание структуры распределенной ИС, включает информацию об узлах сети и доменах приложений (подсистемах ИС), каналах связи, по которым осуществляется взаимодействие подсистем. Слой структур в модели представляется *P-графом* (*графом с полюсами* [7]) $Str = (N, A)$, где N – множество вершин с полюсами, представляющих домены приложений, узлы сети; A – множество связывающих их дуг, представляющих каналы связи.
- *Слой событий* $Ev = \{T, E, Q, Init(Q), Init(E), Ch, Sch\}$, где T – множество моментов времени; E – конечное множество событий; Q – конечное множество состояний; $Init(Q)$: $T \to Q$ – отображение, задающее начальное состояние; $Init(E)$: $T \to E \times T$ – отображение, задающее начальное планирование событий; Ch: $E \times Q \times T \to Q$ – отображение, определяющее новое состояние, в которое система переходит в результате совершения события; Sch: $E \times Q \times T \to E \times T$ отношение планирования, представляющее причинно-следственные связи между событиями. Слой событий

представляет описание работы ИС во времени. Данный слой включает информацию о различных состояниях, в которых может находиться система, и событиях, вызывающих смену состояний. Представлением данного слоя в модели ИС является *ориентированный граф*, вершинам которого ставятся в соответствие *состояния* ИС, в которых может находиться система в различные моменты времени, а дуги представляют *события* (в том числе связанные с получением сообщений), вызывающие смену состояний. Такой набор свяжем с каждой вершиной структуры Str.

– *Слой сообщений* Msg содержит описание данных, которыми подсистемы ИС могут обмениваться между собой, и правила преобразования этих данных. Слой задается как доопределение слоя событий.

Система защиты также является многоуровневой и включает в себя следующие уровни:

– уровень основной защитной логики,

– уровень контроля привилегий,

– уровень собственной безопасности,

– уровень системной безопасности.

Многоуровневый подход к организации защиты позволяет, кроме всего прочего, осуществлять *независимое проектирование различных механизмов защиты*. В частности, появляется возможность реализовывать «высокоуровневую» защитную логику (например, функции проверки вводимых активационных данных), основываясь на предположении невозможности изменения злоумышленником программного кода, реализующего эти функции, т.к. защита этого кода осуществляется на другом уровне.

На уровне *основной защитной логики* реализуется основная функциональность, требуемая от систем обнаружения вторжений согласно стандарту ISO 15408: контроль сетевого трафика, мониторинг работы сервисов ИС, выявление аномальных активностей.

Основным механизмом данного уровня является *подсистема активного аудита*, реализующая статистический и сигнатурный подходы к выявлению и анализу активности, описанные в требованиях FAU_SAA «Анализ данных аудита безопасности» (*Security audit analysis*). Основная функция подсистемы активного аудита – выявление аномалий в работе ИС. В общем случае любая попытка взлома является аномалией, выявляемой на основе статистического анализа работы ИС за продолжительный промежуток времени. В первую очередь имеется в виду анализ работы различных сервисов, расположенных на серверных модулях ИС. В целях компенсации недостатков статистического подхода к анализу активностей, таких как сложность принятия решений в условиях отсутствия устоявшейся эмпирической базы фактов и сложность обнаружения атаки в случае постепенного планомерного изменения параметров активности в сторону характерных для атаки, в рамках подсистемы активного аудита применяется *сигнатурный метод* выявления злоумышленной активности, соответствующий требованиям FAU_SAA.4 «Сложная эвристика атаки» (*Complex attack heuristics*). Под *сигнатурой* в данном случае понимается определённая последовательность событий, характерная для попытки взлома системы. Эффективная реализация механизма активного аудита достигается за счёт использования возможностей *распределённой мультиагентной системы*, лежащей в основе системы защиты. Информация от *агентов-сенсоров* с различных узлов сети, служащей инфраструктурой ИС, стекается к *агентам-аналитикам*, отвечающим за её обработку и формирование вывода о текущем состоянии системы и потенциальных угрозах её безопасности. Анализ производится на основе описанной выше иерархической трёхслойной модели ИС, содержащей также и информацию о самой системе защиты.

Уровень *контроля привилегий* содержит функциональность, обеспечивающую поддержку контроля прав пользователей системы на основе хранимых профилей активности в соответствии с требованиями FAU_SAA.2 «Выявление аномальной активности, основанное на применении профилей» (*Profile based anomaly detection*). Как правило, целью любой атаки на крупные корпоративные ИС является получение доступа к конфиденциальным данным или к защищённым сервисам, что, в конечном счете,

подразумевает необходимость получения высокого уровня привилегий в атакуемой системе [5]. Основным механизмом данного уровня является *подсистема анализа активностей пользователей*.

В ИС выделяются *группы пользователей*, каждой из которых соответствует определённый *набор привилегий*. В ходе этапа настройки и тестирования ИС производится сбор статистических данных о *типах активностей*, присущих данным группам пользователей, и формируются *групповые модели*, представляемые *графами активностей*. В групповой модели содержится информация, характеризующая поведение состоящих в группе пользователей при входе в систему, при работе в системе и при выходе из неё. После завершения построения групповых моделей для каждого пользователя строится *индивидуальная модель*.

Модель поведения представляет собой ориентированный граф $G = \{V, A\}$, где $V = \{v_i\}$ – множество вершин, на котором определено отношение порядка по следующему правилу: элемент, включённый в множество V последним, имеет старший номер в нём; $A = \{a_{ij}\}$ – множество дуг графа G. Каждому элементу $a_{ij} \in A$ ставится в соответствие некоторый вес $w_{ij} \in W$, где W – множество допустимых весов дуг. Вершины $v_i \in V$ представляют значения контролируемых параметров. Дуги $a_{ij} \in A$ представляют семантические связи между значениями контролируемых параметров, характеризующие очерёдность добавления вершин, соответствующих значениям параметров, т.е. элементов $v_i \in V$, в граф G. Веса $w_{ij} \in W$, назначенные дугам $a_{ij} \in A$, задают семантические расстояния между значениями контролируемых параметров, соответствующими инцидентным этим дугам вершинам $v_i, v_J \in V$. Семантическое расстояние характеризует различие между значениями контролируемого параметра активности. Построенная модель позволяет контролировать соответствие параметров некоторым эталонным значениям, «накапливая» происходящие изменения для последующего анализа.

В основе моделей лежит анализ различных типов *параметров активности пользователей*:

- *Категориальные параметры*. Примерами категориальных параметров могут служить измененные файлы, записи в БД, используемые сервисы ИС, инициированные команды, типы ошибок и т.п. Анализ категориальных параметров активности носит *событийно-ориентированный* характер.
- *Числовые параметры*. К данному типу относятся любые параметры активности, значения которых можно оценить количественно, например, объём переданной и запрошенной информации, количество сервисов, используемых единовременно, а так же количество вершин и дуг в модели.
- *Параметры интенсивности,* например, количество входов пользователя в систему за фиксированный промежуток времени, интенсивность запросов к БД, и т.п.
- *Параметры распределения событий*. К этому типу можно отнести, например, соотношения частоты таких событий как запрос на просмотр и запрос на изменение, обращёний к определённым сервисам ИС.

Основное применение моделей заключается в реализации *механизма аутентификации*, основанного на сопоставлении текущего поведения пользователя со статистическими сведениями об обычных параметрах его активности. Данный механизм является *дополнением стандартных механизмов аутентификации* и служит для защиты от несанкционированного получения привилегий путём кражи идентификационной информации привилегированных легальных пользователей.

Индивидуальные модели пользователей и групповые модели в динамически настраиваемых ИС могут быть использованы и для целей, не связанных с защитой, таких, например, как автоматическая генерация и настройка пользовательского интерфейса на основе статистической информации о применяемой данным пользователем или группой пользователей функциональности ИС.

Рассмотренные выше уровни защиты проектируются на основании предположения о невозможности модификации злоумышленником лежащего в его основе программного кода. На уровне *собственной безопасности* реализуются механизмы, при помощи которых осуществляется защита программного кода ИС от анализа и изменения.

Основные механизмы данного уровня:

- Механизм *явного контроля целостности программного кода*, инициирует мгновенную реакцию системы защиты. В рамках данного механизма реализуются проверки программного кода приложения на предмет присутствия в нем несанкционированных изменений, а так же криптографические средства защиты программных модулей.
- Механизм *неявного контроля* используется для организации отложенной реакции системы на факт взлома с целью предотвращения возможности использования взломанного приложения. При обнаружении факта внесения злоумышленником изменений в программный код или деактивации им механизмов защиты первых уровней и/или механизма явного контроля система защиты переводится в имитирующий режим. При этом отсутствуют какие-либо внешние проявления обнаружения попытки взлома, но модули ИС, подвергшиеся злоумышленному воздействию, фактически изолируются в том смысле, что предотвращается возможность обращения с их помощью к ключевым данным и сервисам ИС.
- Механизм *сокрытия местонахождения функций системы защиты*. Данный механизм направлен в первую очередь на функции, отвечающие за обратную связь с пользователями ИС и, в частности, вывод сообщений об ограничении доступа блокировке защищённых сервисов в случае обнаружении попыток взлома (так называемые nag screens). Функции обратной связи, генерирующие такого вида сообщения, в большинстве случаев являются наиболее удобной отправной точкой для взлома системы [8]. Сокрытие производится путём вынесения всех потенциально опасных с точки зрения угрозы их обнаружения функций в динамически генерируемые программные модули. При этом функции, генерирующие «опасные» сообщения, ни в каком виде не хранятся в файлах приложения и их выявление и изменение становятся достаточно сложной задачей.

Уровень системной безопасности. Функционирование подавляющего большинства вредоносных программ невозможно без получения определённых привилегий, дающих возможность доступа к защищённым системным функциям. Доступ к системным функциям необходим для таких задач как открытие сетевых портов (например, для взаимодействия с внедрённым в атакуемую систему троянским модулем), исполнение программ в режиме отладки (в целях выявления брешей в защите), получение доступа к защищённым разделам внешней памяти и к адресным пространствам исполняемых программ, а так же к контроллерам ввода/вывода. Идеальным вариантом является получение возможности исполнения кода на нулевом уровне привилегий – это даёт возможность получения прямого доступа к любым ресурсам атакуемой системы, в том числе и к функциям ядра операционной системы и физическим устройствам. Защита уровня ядра служит для предотвращения возможности получения злоумышленником доступа к защищённым функциям операционной системы и, в частности, к ядру операционной системы.

Основными механизмами данного уровня являются: механизм выявления скрытых процессов, механизм контроля сетевого взаимодействия, механизм защиты нулевого уровня привилегий.

Механизм *контроля сетевого взаимодействия* служит для анализа состояния сетевых портов с целью выявления несанкционированных попыток открытия новых и изменения режима работы активных портов.

Данный механизм реализуется путём отслеживания вызовов соответствующих функций ядра ОС (в случае ОС Windows это Native API [8]), осуществляемого путём установки на данные функции оболочек, реализующих интерфейсы обратного вызова. Отслеживание вызовов функций ядра ОС является достаточным условием обнаружения несанкционированных попыток получения доступа к потенциально опасным с точки зрения организации безопасной работы функциям ОС, т.к. вызов любой функции прикладных программных интерфейсов, в конечном счёте, приводит к вызову некоторой функции ядра ОС. Кроме того, в большинстве случаев одной функции ядра ОС соответствует несколько различных функций прикладных программных интерфейсов, являющихся, по сути, оболочками данной функции, осуществляющими её вызов с некоторым конкретным набором параметров [8], и только контроль на

уровне ядра может гарантировать защиту, не зависящую от возможности появления новых программных интерфейсов и новых способов получения доступа к потенциально опасным функциям ОС.

Механизм защиты нулевого уровня привилегий обеспечивает защиту функций, выполняемых на нулевом уровне привилегий системы. Наибольшую угрозу безопасности представляют так называемые наборы средств для взлома [5] – руткиты (от англ. *rootkit*), работающие на нулевом уровне привилегий. Это вредоносное программное обеспечение, которое предоставляет злоумышленнику практически полный контроль над инфицированной системой и практически не поддающееся обнаружению и ликвидации. Возможны реализации руткита в виде отдельного драйвера или в виде оболочки некоторой функции ядра ОС. Предотвращение внедрения в защищаемую систему руткита осуществляется путём *контроля вызовов функций ядра ОС*, отвечающих за загрузку драйверов и образов в систему. Обнаружение руткитов, устанавливающих оболочки на функции ядра ОС, является технически несложной задачей, т.к. для обнаружения факта несанкционированной модификации достаточно иметь информацию об исходной структуре функций ядра. В рамках данного механизма осуществляются *периодические проверки соответствия значений хеш-функций, вычисленных от программного кода функций ядра ОС*, эталонным значениям, полученным при развертывании системы защиты и санкционированном внесении изменений в данные функции. Дополнительной мерой может служить отслеживание попыток получения доступа к памяти по адресам, соответствующим функциям ядра ОС, но это неминуемо приведёт к заметному снижению производительности защищаемой системы, и поэтому данная мера может применяться только в ситуациях, требующих обеспечения максимального уровня безопасности.

Механизм выявления скрытых процессов (невидимых на прикладном уровне) служит для обнаружения вредоносного ПО, работающего на прикладном уровне. Выявление скрытых процессов осуществляется при помощи работающего на системном уровне *драйвера защиты*.

Заключение

Основными преимуществами предлагаемой системы защиты являются:

- *Гибкость и возможность быстрой динамической адаптации* системы защиты к новым угрозам путём модификации баз знаний.
- *Универсальность*. Предлагаемая система защиты основывается на детализированной многоуровневой модели защищаемой ИС, что даёт потенциальную возможность интеграции в любую ИС.
- *Простота изменения и расширения функциональности*. Предлагаемая система защиты основана на работе со знаниями, и в большинстве случаев её функциональность можно расширить без внесения изменений в программный код.
- *Высокая производительность*. Метазнания, заложенные в системе защиты, дают возможность максимизировать эффективность её работы на основе данных анализа работы в рамках защищаемой ИС.
- *Возможность интеграции* системы защиты на поздних этапах разработки ИС за счет отсутствия необходимости внедрения защитной логики непосредственно в модули защищаемой ИС.

В настоящее время ведется разработка программных компонентов комплексной системы защиты на основе использования метаданных, управляющих функционированием информационных систем, построенных на базе технологии METAS, созданной сотрудниками АНО «Институт компьютинга».

Библиографический список

[1] Лядова Л.Н. Архитектура информационной системы «Образование Пермской области» // Математика программных систем: Межвузовский сборник научных трудов / Перм. ун-т. Пермь, 2002. С. 25-35.

[2] Кастер Х. Основы Windows NT и NTFS / Пер. с англ.— М.: Издательский отдел «Русская редакция» ТОО «Channel Trading Ltd.», 1996.

[3] Минский М. Фреймы для представления знаний. М.: Энергия, 1979.

[4] Huhns M., Stephens L. Multiagent Systems and Societies of Agents // Weiss G. Multiagent systems: a modern approach to a distributed artificial intelligence / Massachusetts Institute of Technology.

[5] Хогланд Г., Мак-Гроу Г. Взлом программного обеспечения: анализ и использование кода. М.: Вильямс, 2005.

[6] Лядова Л.Н., Мороз А.А. Модель защиты программного обеспечения от несанкционированного распространения // В кн.: Сборник трудов Второй международной научно-технической конференции «Инфокоммуникационные технологии в науке, производстве и образовании» (Инфоком 2) / Кисловодск, 2006. С. 120-124

[7] Миков А.И. Автоматизация синтеза микропроцессорных управляющих систем. Иркутск: Изд-во Иркут. ун-та, 1987.

[8] Касперски К. Техника и философия хакерских атак. М.: СОЛОН-Пресс, 2004.

Сведения об авторах

Денис Курилов *– студент кафедры математического обеспечения вычислительных систем Пермского государственного университета; Россия, г. Пермь, 614990, ул. Букирева, 15; e-mail: Denis.Kurilov@mail.ru*

Людмила Лядова *– заведующий кафедрой математического обеспечения вычислительных систем Пермского государственного университета; Россия, 614990, ул. Букирева, 15; e-mail: LNLyadova@mail.ru*

ГЕНЕТИЧЕСКИЙ АЛГОРИТМ ДЛЯ ОПРЕДЕЛЕНИЯ ДЛИНЫ КЛЮЧА И ДЕШИФРОВАНИЯ ПЕРЕСТАНОВОЧНОГО ШИФРА

Алексей Городилов, Владимир Морозенко

Аннотация*: В статье рассматривается возможность применения генетических алгоритмов к задачам криптоанализа. Разработан и описан генетический алгоритм для поиска секретного ключа блочного перестановочного шифра. Ключом в данном случае является перестановка начального фрагмента натурального ряда. Разработанный алгоритм точно определяет длину секретного ключа и с регулируемой «точностью» находит саму секретную перестановку. Анализ результатов вычислительного эксперимента свидетельствует о возможности почти полного автоматического дешифрования текста.*

Keywords*: криптография, криптоанализ, блочный перестановочный шифр, генетический алгоритм, защита информации.*

ACM Classification Keywords*: I.2 Artificial Intelligence: I.2.8 Problem Solving, Control Methods, and Search - Heuristic methods.*

Введение

Основными задачами криптологии являются разработка надежных схем шифрования (задача криптографии) и нахождение эффективных методов дешифрования существующих схем (задача криптоанализа). Криптографический способ защиты информации предусматривает такое её преобразование, при котором она становится доступной для прочтения лишь обладателю секретного ключа. Надежность этого способа защиты определяется стойкостью используемой схемы шифрования к криптоанализу. При криптоанализе конкретного шифра предполагается, что сама схема шифрования известна, а неизвестным является только секретный ключ и/или его длина. Иными словами, задача

вскрытия шифра состоит в нахождении единственного настоящего секретного ключа среди множества всех возможных ключей, т.е. является задачей поиска. При этом пространство поиска велико, а критерий «качества» найденного решения, как правило, не поддается строгой формализации.

В данной работе рассматривается задача криптоанализа блочного перестановочного шифра. Секретным ключом здесь является перестановка начального фрагмента натурального ряда, длина которой также неизвестна. Для решения поставленной задачи в данной работе применяются генетические алгоритмы. Поскольку одной из областей успешного применения генетических алгоритмов являются именно задачи оптимизации и поиска, то их использование в данном случае представляется вполне объяснимым. Вопросам применения генетических алгоритмов в криптологии посвящен ряд статей [Delman, 2005, Лебедев, 2005, Jakobsen, 1995], но в них длина ключа считается известной, используются стандартные подходы, не учитывающие специфику задачи, либо не исследуется влияние параметров на скорость сходимости алгоритмов.

Разработке всякого генетического алгоритма должны предшествовать выбор подходящего способа кодирования допустимого решения в виде символьной строки, а также описание основных операторов – селекции, кроссовера и мутации. Качество алгоритма должно регулироваться за счет вариации его параметров – численности популяции, числа поколений, вероятностных характеристик основных операторов.

Блочный перестановочный шифр

В данной работе рассматривается конкретная схема шифрования – симметричный блочный перестановочный шифр. Она состоит в том, что входной текст разбивается на блоки, т.е. строки фиксированной длины N, а затем внутри каждого блока символы переставляются в соответствии с заданной перестановкой

$$P = \begin{pmatrix} 1 & 2 & 3 & \dots & i & \dots & N \\ p_1 & p_2 & p_3 & \dots & p_i & \dots & p_N \end{pmatrix}.$$

где $p_i \in \{1,2,3,\dots,N\}$. Иными словами, символ, стоящий на i-й позиции, перемещается на позицию p_i. Секретным ключом в таком шифре является перестановка P. Расшифровывание происходит с использованием обратной перестановки P^{-1}.

Генетический алгоритм

Поскольку решением задачи криптоанализа в данном случае является перестановка, то и особью в генетическом алгоритме будем считать перестановку. Для начала предположим, что длина ключа нам известна, и нам остается лишь найти сам ключ, т.е. перестановку фиксированной длины N.

Важный вопрос, который необходимо решить – какой смысл будет вкладываться в отдельные гены особи. Простейший вариант, который является на первый взгляд наиболее очевидным, – считать отдельными генами элементы перестановки P, то есть i-ым геном особи считать число p_i. Очевидно, что при таком подходе гены получаются зависимыми друг от друга. Если какой-то ген равен j, то никакой другой ген этой особи уже не должен принимать значение j, так как в перестановке P все числа от 1 до N встречаются ровно по одному разу. Зависимость генов накладывает значительные ограничения на операторы мутации и скрещивания. Стандартные операторы в данной ситуации неприменимы, так как все они работают с представлением набора генов в виде строки независимых бит. Такие зависимости генов не свойственны живой природе, что делает аналогию неточной и ставит под сомнение эффективность применения генетического алгоритма. Тем не менее, выбранная интерпретация генов как элементов перестановки интуитивно понятна и не требует дополнительных затрат на их формирование.

Заметим, что альтернативным подходом могло бы стать использование промежуточного представления особей в виде некоторого объекта, легко трансформируемого в перестановку. В то же время такой объект должен задаваться при помощи битовой строки, так чтобы были применимы стандартные операторы скрещивания и мутации. При этом подходе задача выбора подходящего промежуточного представления особи может оказаться достаточно трудной. В данной работе выбран первый – интуитивно понятный – из указанных подходов, т.е. везде в дальнейшем в качестве отдельных генов будут рассматриваться элементы перестановки.

Следующий вопрос – как вычислять приспособленность особей. Мы будем исходить из предположения, что шифруемый текст представляет собой осмысленный текст на русском языке. Фитнесс-функция (целевая функция) будет заимствована из работ Якобсена [Jakobsen, 1995, Аграновский, 2002]. Томас Якобсен в 1995 году предложил автоматический метод раскрытия ключа шифра простой замены. В своем методе вскрытия шифров замены Якобсен использовал информацию о распределении частот встречаемости биграмм в осмысленных текстах. Биграмма – это две подряд идущие буквы в тексте. Целевую функцию Якобсен предложил вычислять как сумму модулей разностей между заранее известным среднестатистическим количеством биграмм в осмысленных текстах и их реальным количеством в шифртексте. Пусть T_{ij} – это относительные частоты встретившихся в тексте T биграмм (i j). Тогда целевая (фитнесс-) функция будет иметь вид

$$W(T) = \sum_{ij} \left| T_{ij} - E_{ij} \right|,$$

где E_{ij} – относительные частоты биграмм, заранее известные и зафиксированные в алгоритме в качестве эталонных значений. Матрица частот E высчитывается заранее на осмысленных текстах большой длины, т.е. отражает среднестатистическое распределение биграмм. Нетрудно видеть, что чем ближе текст к осмысленному, тем меньше значение целевой функции и тем «ближе» найденный ключ к настоящему секретному ключу. Это означает, что меньшему значению целевой функции соответствует большее значение приспособленности особи, и наоборот. Важно подчеркнуть, что указанная фитнесс-функция W, вообще говоря, не обращается в ноль, даже если расшифрованный текст T получен из шифр-текста при использовании настоящего секретного ключа.

Таким образом, если нам дан зашифрованный текст S, то для вычисления приспособленности особи P, необходимо выполнить следующие шаги.

1. Расшифровать зашифрованный текст S с использованием выбранного ключа P, в результате чего получим текст $T = Decrypt_P(S)$.
2. Подсчитать частоты T_{ij} всевозможных биграмм (i j) в тексте T.
3. Найти значение целевой функции $W(T)$ по указанной выше формуле.

Более подробно рассмотрим свойства фитнесс-функции $W(T)$. Пусть две особи P_1 и P_2 различаются только двумя первыми генами (очевидно, что только одним геном они отличаться не могут). Положим для определенности

$$P_1 = \begin{pmatrix} 1 & 2 & 3 & \dots & N \\ p_1 & p_2 & p_3 & \cdots & p_N \end{pmatrix},$$

$$P_2 = \begin{pmatrix} 1 & 2 & 3 & \dots & N \\ p_2 & p_1 & p_3 & \cdots & p_N \end{pmatrix}.$$

Пусть имеется зашифрованное сообщение S. Тогда расшифрованные при помощи ключей P_1 и P_2 тексты $Decrypt_{P_1}(S)$ и $Decrypt_{P_2}(S)$ отличаются только символами, стоящими на позициях с номерами p_1, p_2, $p_1 + N$, $p_2 + N$, $p_1 + 2N$, $p_2 + 2N$ и т.д. Номера указанных позиций образуют две арифметических

последовательности с разностью *N*. Поскольку тексты $Decrypt_{P_1}(S)$ и $Decrypt_{P_2}(S)$ могут отличаться только в указанных позициях, то в каждом блоке длины *N* этих текстов не более трех биграмм будут отличаться своими частотами. Таким образом, при достаточно большой длине *N* ключа, что справедливо для реальных систем, небольшому изменению особи будет соответствовать небольшое изменение целевой функции.

Заметим также, что в достаточно длинных осмысленных текстах частоты встречаемости биграмм T_{ij} близки к соответствующим среднестатистическим вероятностям E_{ij}. Поэтому в результате дешифровки текста *S* с помощью настоящего секретного ключа значение целевой функции должно быть близким к нулю. Значит, если *K* – настоящий секретный ключ, то величина $W(Decrypt_K(S))$ должна быть минимально возможной.

Как было сказано выше, стандартные операторы скрещивания применимы только тогда, когда особь представляется битовой строкой, состоящей из независимых бит. Поскольку в нашем случае это не так, то стандартные операторы неприменимы и требуется разработка оператора кроссовера (скрещивания) специального вида. Основное требование, накладываемое на данный оператор, заключается в том, чтобы в результате его применения всегда получалась допустимая перестановка. В данной работе предлагается следующий оператор скрещивания.

1. Гены пронумерованы числами 1, 2, 3, …, *N* и просматриваются в порядке возрастания номеров.
2. Ген с очередным номером берется от одного из предков, если это возможно.
3. Если гены от обоих родителей недопустимы, берется произвольное допустимое число.

Будем использовать обозначение $P(i)$ для числа, на которое перестановка *P* заменяет число *i*. Таким образом, получаем следующий алгоритм скрещивания двух родителей P_1 и P_2 и получением потомка *P**.

1. Множеству *Used* уже использованных значений генов присвоить начальное значение Ø, установить номер текущего вычисляемого гена $i = 1$.
2. Выяснить, принадлежат ли числа $P_1(i)$ и $P_2(i)$ множеству *Used*.
3. Если множеству *Used* не принадлежит только одно из чисел $P_1(i)$ и $P_2(i)$, то присвоить его *i*-му гену потомка *P**. Если множеству *Used* не принадлежат оба числа $P_1(i)$ и $P_2(i)$, то либо с вероятностью p_c выбрать число $P_1(i)$, либо с вероятностью $(1 – p_c)$ выбрать $P_2(i)$. Выбранное число присвоить *i*-му гену потомка *P**. Наконец, если оба числа лежат во множестве *Used*, то *i*-му гену потомка *P** присвоить произвольное число из множества {1, 2, 3, …, N}\ *Used.*
4. Число, которое было присвоено *i*-му гену потомка *P**, включить во множество *Used*.
5. Перейти к рассмотрению следующего гена, т.е увеличить *i* на единицу и перейти к шагу 2.

В качестве оператора мутации можно использовать стандартный оператор обмена, при котором с заданной вероятностью в особи меняются местами два гена. При селекции особей использовалось две основные классические идеи: «принцип рулетки» и «принцип элитизма» [Mitchell, 1999]. Размер популяции оставался постоянным за счет того, что каждый раз при появлении двух потомков, из популяции удалялись две наименее приспособленные особи. Наконец, условием окончания работы предлагаемого генетического алгоритма было выбрано превышение количества эпох заранее фиксированной величины *M*. Это позволяет заранее предсказать время, которое потребуется для работы алгоритма или, наоборот, задать параметр *M* так, чтобы алгоритм завершил свою работу за указанное ограниченное время.

Влияние параметров генетического алгоритма

Итак, выше был предложен генетический алгоритм для решения задачи криптоанализа блочного перестановочного шифра в предположении, что длина *N* ключа известна. Исследуем вопрос об оптимальных значениях параметров этого алгоритма. Выделим те параметры алгоритма, которые

упоминались в тексте и которые оказывают влияние на скорость сходимости и качество получаемого решения:

- численность начальной популяции k;
- вероятность унаследования генов от одного из родителей p_c;
- вероятность мутации p_m;
- количество поколений M.

Каких-либо общих теоретических рекомендаций для выбора этих параметров не существует. Можно лишь сказать, что вероятность мутации должна быть незначительной, а вероятность унаследования генов от одного из родителей, наоборот, должна быть величиной, близкой к 0,5. Количество поколений может быть выбрано исходя из имеющихся временных ресурсов. Вообще говоря, если единственным условием окончания работы алгоритма является количество поколений, то оно должно быть достаточно большим. Также важна численность начальной популяции, поскольку в предлагаемом генетическом алгоритме численность популяции остается неизменной и всегда равна k. Следует отметить, что численность популяции существенным образом влияет на объем вычислений, и, следовательно, на затрачиваемое время.

Проведенные численные эксперименты показали, что малая численность популяции дает плохие результаты: алгоритм быстро находит точку локального минимума фитнесс-функции, и все дальнейшие популяции формируются в его окрестности. Оказалось, что при длине ключа, равной 10, оптимальным значением численности популяции является 15-17 особей. Изменения показателя вероятности мутации в разумных пределах не оказали заметного влияния на работу алгоритма. Частично это можно объяснить тем, что специфически выбранный оператор скрещивания уже подразумевает некоторую мутацию. Действительно, когда ген не может быть унаследован ни от одного из предков, он просто выбирается случайным образом, то есть в этом гене потомок получается «непохожим» ни на одного из своих предков. Оптимальное число поколений сильно зависит от длины N ключа. При малых значениях параметра N слишком большое число поколений неэффективно, так как после некоторого числа поколений алгоритм находит локальный минимум, и дальнейшие действия не приводят к улучшению решения. С ростом длины ключа быстро растет пространство возможных решений и, естественно, требуется большее число поколений для нахождения наилучшего решения.

Получаемый в результате работы алгоритма текст далеко не всегда совпадает с исходным открытым текстом, однако не сильно от него отличается. Во многих случаях, восстановить отдельные символы можно с помощью контекста. Таким образом, алгоритм хотя и не решает задачу криптоанализа полностью автоматически, но значительно помогает дешифровать зашифрованный текст.

Определение длины ключа

Описанный выше генетический алгоритм работает при известной заранее длине ключа N. Результатом его работы является перестановка, соответствующая наиболее приспособленной особи из финальной популяции. Поскольку эта перестановка доставляет целевой функции $W(T)$ наименьшее из её найденных значений, то побочным эффектом в работе описанного алгоритма является вычисление по заданной длине ключа N числа $F(N)$, указывающего на минимальное найденное значение целевой функции.

Рассмотрим теперь задачу нахождения длины ключа по имеющемуся зашифрованному тексту. Пусть N – предполагаемое значение длины настоящего секретного ключа, а L – его настоящая длина. Если числа N и L совпадают, то можно ожидать, что найденная в результате работы генетического алгоритма перестановка будет близка к искомой, а целевая функция при использовании этой перестановки примет относительно небольшое значение.

Пусть настоящим секретным ключом является перестановка

$$K = \begin{pmatrix} 1 & 2 & \cdots & L \\ p_1 & p_2 & \cdots & p_L \end{pmatrix}.$$

Рассмотрим ключ длины 2·*L* вида

$$K' = \begin{pmatrix} 1 & 2 & \cdots & L & L+1 & \cdots & 2 \cdot L \\ p_1 & p_2 & \cdots & p_L & L+p_1 & \cdots & L+p_L \end{pmatrix},$$

в котором первые *L* столбцов в точности совпадают с первыми *L* столбцами перестановки *K*, а остальные получены из первых *L* столбцов перестановки *K* увеличением всех чисел на *L*. Поскольку оба ключа K и K' превращают любой исходный текст в идентичные шифр-тексты, то шифр-текст будет правильно расшифрован также и при помощи ключа K'. Это означает, что при поиске ключа длины 2·*L* генетический алгоритм найдет перестановку, близкую к ключу K'. При этом найденной перестановке будет соответствовать относительно небольшое значение целевой функции, поэтому полученная на выходе алгоритма величина *F*(2·*L*) также примет небольшое значение. То же самое справедливо для всех ключей с длинами, кратными числу *L*. Иными словами, если на вход генетическому алгоритму подать шифр-текст и длину ключа *N* = *m*·*L*, где *m* – натуральное число, то на выходе алгоритма следует ожидать перестановку с небольшим значением фитнесс-функции. Если же *N* не кратно *L*, то перестановок длины *N*, близких к ключу *K*, вообще не существует. В этом случае полученное значение фитнесс-функции и, соответственно, величина *F*(*N*), должны быть относительно велики. Таким образом, для всех *N*, не кратных *L*, величина *F*(*m*·*L*) должна быть существенно меньше, чем *F*(*N*).

Исходя из вышесказанного, можно предложить следующий алгоритм определения длины ключа. Выбирается некоторое начальное значение длины ключа N_0 и для него с помощью имеющегося генетического алгоритма вычисляется значение $F(N_0)$. На следующем шаге предполагаемую длину ключа увеличиваем на единицу и, вновь применив генетический алгоритм, вычислим $F(N_1)$, где $N_1 = N_0 + 1$, и т.д. На *i*-ом шаге вычисляем значение $F(N_i)$, где $N_i = N_0 + i$. Увеличение длины ключа производятся до достижения некоторой установленной заранее верхней границы *N**. В результате такого итерационного процесса, на каждом шаге которого запускается описанный генетический алгоритм, образуется последовательность чисел

$$F(N_0),\ F(N_0+1),\ F(N_0+2),\ \ldots,\ F(N_0+i),\ \ldots,\ F(N^*).$$

В ней своими малыми значениями должны выделяться элементы, чьи номера образуют арифметическую прогрессию с шагом *L*. Величина шага *L* и будет являться искомой длиной настоящего секретного ключа.

Отметим, что для ускорения работы алгоритма при нахождении длины ключа можно использовать малые численности популяции и небольшое число поколений, а после нахождения истинной длины ключа *L* можно повторно применить генетический алгоритм с большими значениями параметров для отыскания самого секретного ключа.

Пример работы алгоритма

В качестве открытого текста был взят фрагмент литературного текста на русском языке – отрывок из рассказа И.С. Тургенева «Рудин»:

> *Было тихое летнее утро. Солнце уже довольно высоко стояло на чистом небе, но поля еще блестели росой, из недавно проснувшихся долин веяло душистой свежестью, и в лесу еще сыром и не шумном, весело распевали ранние птички. На вершине пологого холма, сверху донизу покрытого…*

В вычислительном эксперименте этот текст был зашифрован при помощи блочного перестановочного шифра с длиной ключа, равной 10. В результате был получен следующий набор символов:

т иолоеыхБенеуеттл оСлц. еонрд оожелув ьвско оноьляонтоас отм ис ч,н бепеонщеебя лл оил отессрези е, дйноп рсноновоасхядшиов уевяон иллтсо шисуйдтсь,же еювелс в е уирыо сиемщушмое мнн есл вер о,авл пе сиаеи тнниапрН ави.ек ч еплиношорх омгоаолгрех свд у, упкизрноо о ог т ы…

Исходя из такого бессмысленного набора символов, трудно воссоздать исходный текст «вручную». Далее этот зашифрованный текст был подан на вход разработанного генетического алгоритма с целью определения длины секретного ключа. Диапазон предполагаемых значений длины ключа составил [4; 34], т.е. были выбраны параметры $N_0 = 4$, $N^* = 34$. Для каждой предполагаемой длины ключа от 4 до 34 были найдены соответствующие значения $F(4)$, $F(5)$, …, $F(34)$. На рис. 1 для удобства представлены обратные значения $1/F(4)$, $1/F(5)$, …, $1/F(34)$. Чем выше точка на графике, тем ближе её абсцисса к истинной длине L секретного ключа или к кратной ей величине $m \cdot L$.

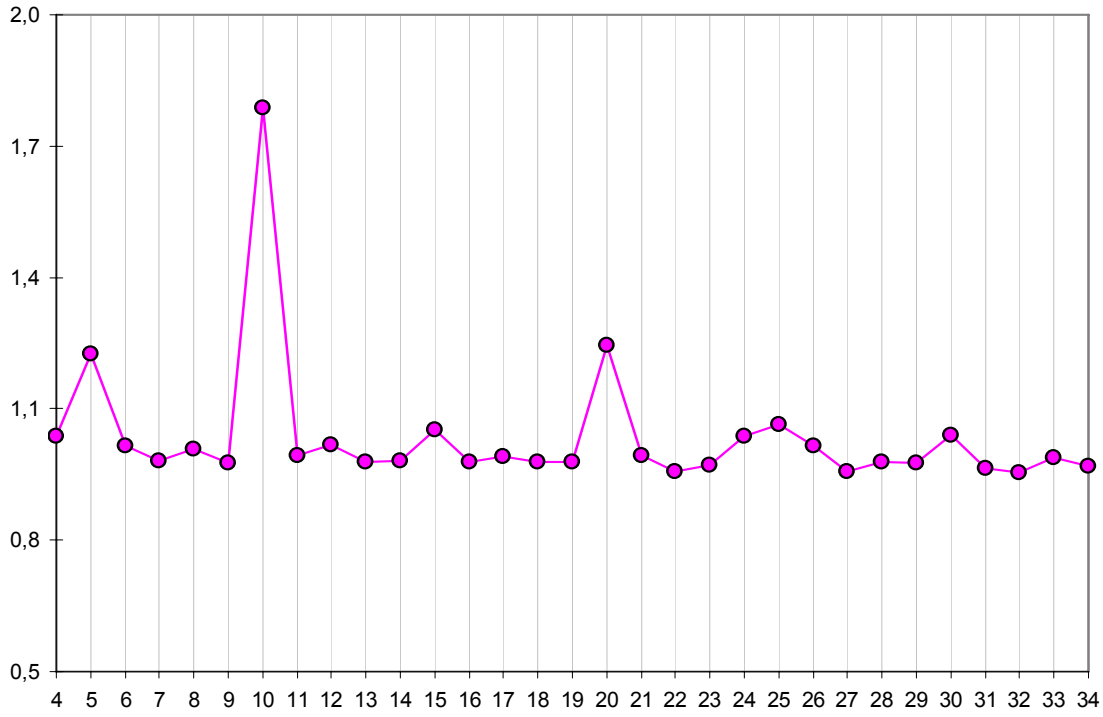


Рис. 1. График зависимости 1/F(N), где N – предполагаемая длина секретного ключа

На рисунке виден максимальный «всплеск» функции приспособленности, который приходится на значение длины, равное 10, что соответствует длине L настоящего секретного ключа. Следующий заметный локальный максимум наблюдается при $N = 20$, что равно удвоенному значению $2 \cdot L$. Далее заметных «всплесков» не наблюдается, хотя при $N = 30$ (т.е. при утроенном значении $3 \cdot L$) можно заметить локальный максимум.

Результаты вычислительного эксперимента хорошо согласуется с теоретическими предсказаниями. Постепенное затухание «всплесков» функции приспособленности по мере увеличения параметра N связано с тем, что для всех значений N генетический алгоритм работал с одним и тем же набором параметров. В соответствии же с вышесказанным, для больших значений предполагаемой длины ключа требуется увеличивать количество поколений. В противном случае, при малом числе поколений генетический алгоритм просто «не успевает» найти хорошее решение.

Таким образом, в данном эксперименте на основе анализа полученных данных можно утверждать, что длина секретного ключа равна $L = 10$. Чтобы теперь найти секретный ключ, повторно был применен генетический алгоритм со следующим набором параметров: численность популяции – 15, количество поколений – 30, вероятность мутации – 0,2. В результате работы алгоритма был получен следующий текст:

> *Былохти ое лет еенутро. нолСце ужевдо ольно оысвко сто лояна чис омт небе он, поля щееблестери лосой, нз иедавноопр снувши сяхдолин леяво душийтос свежеютьс, и в уесл еще смроы и не нумшом, веоелс распеиалв раннпе итички. На вершино пелогоголхо ма, свурхе донизо пукрытог о…*

Как видно, полностью дешифровать зашифрованный текст не удалось, т.к. полученный текст не совпадает с исходным. Однако можно заметить, что он «близок» к исходному. Например, в нем можно заметить осмысленные слова, которые совпадают с соответствующими словами исходного текста: *было*, *уже*, *поля*, *небе и др.* На основе полученного текста уже можно «вручную» довести процесс дешифрования до конца.

Кроме того, секретный ключ, которым в данном случае является перестановка

$$\begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 \\ 10 & 8 & 5 & 6 & 2 & 1 & 3 & 9 & 4 & 7 \end{pmatrix},$$

и ключ, полученный в результате работы генетического алгоритма

$$\begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 \\ 10 & 8 & 5 & 6 & 9 & 1 & 3 & 2 & 4 & 7 \end{pmatrix},$$

отличаются только в одной паре значений, что в генетическом алгоритме соответствует одной паре хромосом. Учитывая, что исходный открытый текст был осмысленным, нетрудно «вручную» перебором небольшого числа вариантов найти настоящий секретный ключ.

Безусловно, предложенный в данной работе генетический алгоритм поиска секретного ключа хотя и не автоматизирует полностью процесс дешифровки, но довольно существенно ускоряет его, делая несложным доведение «вручную» процесс дешифровки до получения осмысленного текста.

Библиографический список

[Delman, 2005] Delman B. GeneticAlgorithms in Cryptography // A Thesis Submitted in Partial Fulfillment of the Requirements for the Degree of Master of Science in Computer Engineering. New York, 2004.

[Лебедев, 2005] Лебедев А. В криптографии мы способны конкурировать / ЛАН Крипто, 2005.

[Jakobsen, 1995] Jakobsen T. A Fast Method for the Cryptanalysis of Substitution Ciphers, 1995.

[Аграновский, 2002] Аграновский А. В., Хади Р.А. Практическая криптография: алгоритмы и их программирование. М.:СОЛОН-Пресс, 2002.

[Mitchell, 1999] Mitchell M. An Introduction to Genetic Algorithms. Fifth printing. Cambridge, MA: The MIT Press, 1999.

Сведения об авторах

Алексей Городилов *– Пермский государственный университет, студент магистратуры кафедры математического обеспечения вычислительных систем; Россия, г. Пермь, 614990, ул. Букирева, д. 15; e-mail: gora830@yandex.ru*

Владимир Морозенко – *Пермский государственный университет, доцент кафедры математического обеспечения вычислительных систем; Россия, г. Пермь, 614990, ул. Букирева, д. 15; e-mail: v.morozenko@mail.ru*

60th Anniversary of

INSTITUTE OF MATHEMATICS AND INFORMATICS of Bulgarian Academy of Sciences

Acad. G. Bonchev Str., block 8, Sofia 1113, Bulgaria
Tel. (+359-2) 979-3824, Fax (+359-2) 971-3649
http://www.math.bas.bg

The Institute of Mathematics and Informatics (IMI) at BAS was founded in 1947 as Institute of Mathematics. At the beginning about ten research fellows were working at the Institute. In 1961 a computational centre was established as part of the Institute. Later specialist in Mechanics also worked at the Institute, hence and it was named Institute of Mathematics and Mechanics. Its present name dates from 1995

The Institute has considerable achievements in the field of Mathematics that are not discussed here.

The development of the Informatics in Bulgaria started at the Institute. Many researchers have built the career of Informatics specialists.

The Institute was the first in Bulgaria to buy an universal analog computing machine MH-7. The first Bulgarian computer was created at the Institute. Soon after that came into exploitation the first imported into Bulgaria computer “MINSK-2”. An original software for this computer – auto code “MIKOD”, operation systems “MID” and “MID-2”, a system for symbol programming “MIKS” and a rich library of programs were created here as well.

The fellows of the Institute also carried out the first Informatics researches in Bulgaria. The Institute has a wide range of activities in Applied Informatics and it continues to produce original software for the solving important problems. Researchers from the Institute organized and taught the first courses in Informatics at the Sofia University “St. Kliment Ohridski” for students in Mathematics. In a short time a major in Informatics was launched with the help of the Institute and later on it became a specialty at the Sofia University. Researchers of the Institute prepared the first syllabus, textbooks, and manuals. The staff of the Institute is also involved in training teachers in Informatics for the secondary school.

In the course of the years the informaticians at focused upon the research activities and many of them are still lecturing Informatics at a number of Bulgarian universities.

Departments of IMI : *Algebra; Artificial Intelligence; Biomathematics; Complex Analysis; Differential Equations; Education in Mathematics and Informatics; Geometry and Topology; Information Research; Laboratory of Mathematical Linguistic; Logic; Mathematical Foundations of Informatics; Mathematical Linguistics; Mathematical Physics; Computational Mathematics; Operation Research; Probability and Statistics; Real and Functional Analysis; Software Engineering; Telecommunications Department.*

15th Anniversary of

ADUIS

ASSOCIATION OF DEVELOPERS AND USERS OF INTELLIGENT SYSTEMS

ADUIS consists of about one hundred members including ten collective members. The Association was founded in Ukraine in 1992. The main aim of ***ADUIS*** is to contribute to the development and application of the artificial intelligence methods and techniques. The efforts of scientists engaged in ***ADUIS*** are concentrated on the following problems: expert system design; knowledge engineering; knowledge discovery; planning and decision making systems; cognitive models designing; human-computer interaction; natural language processing; methodological and philosophical foundations of AI.

Association has long-term experience in collaboration with teams, working in different fields of ***research and development***. Methods and programs created in Association were used for revealing regularities, which characterize chemical compounds and materials with desired properties. Some thousands of high precise prognoses have been done in collaboration with chemists and material scientists of Russia and USA.

Association can help ***businessmen*** to find out conditions for successful investment taking into account region or field peculiarities as well as to reveal user's requirements on technical characteristics of products being sold or manufactured.

Physicians can be equipped with systems, which help in diagnosing or choosing treatment methods, in forming multi-parametric models that characterize health state of population in different regions or social groups.

Sociologists, politicians, managers can obtain the Association's help in creating generalized multi-parametric "portraits" of social groups, regions, enterprise groups. Such "portraits" can be used for prognostication of voting results, progress trends, and different consequences of decision making as well.

Association provides a useful guide in technical diagnostics, ecology, geology, and genetics.

ADUIS has at hand a broad range of high-efficiency original methods and program tools for solving analytical problems, such as knowledge discovery, classification, diagnostics, prognostication.

ADUIS unites the creative potential of highly skilled scientists and engineers

Since 1992 ***ADUIS*** holds regular conferences and workshops with wide participation of specialists in AI and users of intelligent systems. The proceedings of the conferences and workshops are published in scientific journals. ***ADUIS*** cooperates through its foreign members with organizations that work on AI problems in Russia, Byelarus, Moldova, Georgia, Bulgaria, Czechia, Germany, Great Britain, Hungary, Poland, etc. ***ADUIS*** is the collective member of the European Coordinating Committee for Artificial Intelligence (ECCAI).

Products developed by ADUIS: **Confor**: Tools for Knowledge Discovery, Classification, Diagnostics and Prediction; **Analogy**: Tools for Solving Problems on the Basis of Analogy; **Manager**: Tools for Decision Support Systems Design; **Discret**: Tool for Discretization of Numerical Data; **Gobsec**: The System for Investment Scheduling.

For contacts: V.M.Glushkov Institute of Cybernetics; National Academy of Science of Ukraine;

Prospect Akademika Glushkova, 40, 03680 GSP Kiev-187, Ukraine;

Phone: (380+44) 5262260; Fax: (380+44) 5263348; E-mail: glad@aduis.kiev.ua

10th Anniversary of

Association for the Development of the Information Society

Acad. G. Bonchev St., block 8, Sofia 1113, Bulgaria
Tel. (+359-2) 979-3813, -3808, Fax (+359-2) 739-808
e-mail: ario@math.bas.bg, adis@einet.bg
http://www.adis.org

The Association for the Development of the Information Society (ADIS) was established in April 1997 and is an independent, non-government, non-profit organization with the non-commercial objective to support the development of the information society in Bulgaria. This objective is extensively defined in the Association's statute and includes:

- Interaction with individuals and organizations working for the development of the information society in Bulgaria and in the world.
- Support of the comprehensive utilization of the capacity of the information infrastructure and information technologies by all layers of society and all ages and professions, as well as by unemployed, ethnic minorities, people with disabilities, etc.
- Development and implementation of national and international projects whose goal is establishing, developing, and governing the information society.
- Participation in the elaboration and implementation of educational, promotional, and demonstration programs dedicated to information society issues.
- Participation in international activities on issues of the development of the information society, and maintenance of ties to and interaction with foreign and international organizations.
- Organization of conferences, forums, workshops dedicated to the information society.
- Publishing of a newsletter distributed among the individual and collective members of the Association.

Besides individual persons, the Association has as collective members from various regions of Bulgaria: Plovdiv University 'Paisii Hilendarski'. Technical University—Gabrovo, the Police Academy, the Institute of Mathematics and Informatics, the Institute of Information Technologies, the Central Laboratory of Computer Security of the Bulgarian Academy of Sciences (Sofia), and other organizations. Societies in the cities of Plovdiv, Shoumen, and Bourgas have been formed as autonomous subsidiaries of the Association. Its membership and associated structures are growing quickly and already include foreign members. The Association has existed since recently but it unites people and organizations with several decades of experience in the field of computer science and information technologies. Since 1999, the Association has organized monthly national seminars in the framework of the Forum Global Information Society. The seminars are devoted to the development of the information society in all fields of the human activities and aspects. Other activities include implementing a project for training disabled (deaf) people to use computers and the Internet, a project for training secondary school teachers in a broad range of computer technologies, participation in the drafting of the Bulgarian national strategy for the Information Society, drafting of models and principals for creating, management and development of public centers for access to Internet, information and communication services and public e-information and e-services for the Bulgarian citizens as well as delivering of talks on Information Society issues at various national and regional events by members of the Association.

The Association gladly welcomes contacts with organizations from abroad whose activities are related to the development of the global information society.

15th Volume of

International Journal "Information Theories and Applications"

Verba volant, scripta manent !

International Journal "Information Theory and Applications" (IJ ITA) has been established in 1993 as independent scientific printed and electronic media. IJ ITA is edited by the Institute of Information Theories and Applications FOI ITHEA in collaboration with the Institute of Cybernetics "V.M.Glushkov", NASU (Ukraine) and Institute of Mathematics and Informatics, BAS (Bulgaria).

During the years, IJ ITA became as well-known international journal. Till now more than 600 papers have been published. IJ ITA authors are widespread in 39 countries all over the world: Armenia, Belarus, Brazil, Belgium, Bulgaria, Canada, Czech Republic, Denmark, Egypt, Estonia, Finland, France, Germany, Greece, Hungary, Ireland, Israel, Italy, Japan, Kirghizia, Latvia, Lithuania, Malta, Mexico, Moldavia, Netherlands, Poland, Portugal, Romania, Russia, Scotland, Senegal, Serbia and Montenegro, Spain, Sultanate of Oman, Turkey, UK, Ukraine, and USA.

IJ ITA major topics of interest include, but are not limited to:

INFORMATION THEORIES

Artificial Intelligence
- *Computer Intellectualisation*
- *Intelligent Networks and Agents*
- *Intelligent Technologies*
- *Knowledge Discovery and Engineering*
- *Knowledge Acquisition and Formation*
- *Distributed Artificial Intelligence*
- *Models of Plausible Reasoning*
- *AI Planning and Scheduling*

Bioinformatics
Cognitive Science
Decision Making
General Information Theory
Hyper Technologies
Information Models
Intellectualisation of Data Processing
Knowledge-based Society
Logical Inference
Natural language Processing
Neuroinformatics
Philosophy and Methodology of Informatics
Quality of the Programs
Software Engineering
Theory of Computation

APPLICATIONS

Communication Systems
Computer Art and Computer Music
Hyper Technologies
Intelligent Information Systems
Multimedia Systems
Programming Technologies
Program Systems with Artificial Intelligence
Very Large Information Spaces

More information about the IJ ITA rules for preparing and submitting the papers as well as how to take out a subscription to the Journal may be obtained from www.foibg.com/ijita.

Second Volume of

International Journal "Information Technologies and Knowledge"

Intelligo ut credam !

International Journal "Information Technologies and Knowledge" (IJ ITK) has been established in 2007 as independent scientific printed and electronic media. IJ ITK is edited by the Institute of Information Theories and Applications FOI ITHEA in collaboration with the Institute of Cybernetics "V.M.Glushkov", NASU (Ukraine); Institute of Mathematics and Informatics, BAS (Bulgaria); Institute of Information Technologies, BAS (Bulgaria); University of Hasselt (Belgium); Natural Computing Group (NCG) of the Technical University of Madrid (Spain); Astrakhan State Technical University (Russia); Taras Shevchenko National University of Kiev (Ukraine); University of Calgary (Canada); VLSI Systems Centre of Ben-Gurion University (Israel).

The main scope of the IJ ITK covers but is not limited to the theoretical research, applications and education in the area of the Information Technologies for:

- Knowledge Collecting and Accumulation
- Knowledge Discovery and Acquisition
- Knowledge Level Modeling
- Knowledge Management -Transfer and Distributing
- Knowledge Market
- Knowledge Representation and Processing
- Knowledge Utilization
- Knowledge-based Society
- Knowledge-based Systems

Many scientific and practical areas are connected to the topics of interest of IJ ITK:

- Business Informatics: e-Management, e-Finance, e-Commerce, e-Banking,
- Business Intelligence: Methodology, Tools and Technologies, Analytics and Statistics;
- Cognitive science
- Competitive Intelligence;
- Data Mining
- Decision Making
- e-Management in Governmental and Municipal Structures: Models, Systems, e-Government, etc.
- Information Technologies in Biomedicine
- Intelligent Communication Technologies and Mobile Systems
- Intelligent Robots
- Intelligent Systems
- Intelligent Technologies in Control and Design
- Modern (e-) Learning Information Technologies
- Multimedia Semantic Systems
- P2P e-Learning Applications
- Planning and Scheduling
- Socio-cognitive engineering
- Technology and Human Resource Issues
- Technology-based Blended, Distance and Open Learning
- Web-based Technologies and Systems, AI/Semantic Web
- etc.

More information about the IJ ITK rules for preparing and submitting the papers as well as how to take out a subscription to the Journal may be obtained from http://www.foibg.com/ijitk .