

Krassimir Markov, Vitalii Velychko,
Lius Fernando de Mingo Lopez, Juan Casellanos
(editors)

**New Trends
in
Information Technologies**

I T H E A

SOFIA

2010

Krassimir Markov, Vitalii Velychko, Lius Fernando de Mingo Lopez, Juan Casellanos (ed.)
New Trends in Information Technologies

ITHEA®

Sofia, Bulgaria, 2010

ISBN 978-954-16-0044-9

First edition

Recommended for publication by The Scientific Concil of the Institute of Information Theories and Applications FOI ITHEA

This book maintains articles on actual problems of research and application of information technologies, especially the new approaches, models, algorithms and methods of membrane computing and transition P systems; decision support systems; discrete mathematics; problems of the interdisciplinary knowledge domain including informatics, computer science, control theory, and IT applications; information security; disaster risk assessment, based on heterogeneous information (from satellites and in-situ data, and modelling data); timely and reliable detection, estimation, and forecast of risk factors and, on this basis, on timely elimination of the causes of abnormal situations before failures and other undesirable consequences occur; models of mind, cognizers; computer virtual reality; virtual laboratories for computer-aided design; open social info-educational platforms; multimedia digital libraries and digital collections representing the European cultural and historical heritage; recognition of the similarities in architectures and power profiles of different types of arrays, adaptation of methods developed for one on others and component sharing when several arrays are embedded in the same system and mutually operated.

It is represented that book articles will be interesting for experts in the field of information technologies as well as for practical users.

General Sponsor: Consortium FOI Bulgaria (www.foibg.com).

Printed in Bulgaria

Copyright © 2010 All rights reserved

© 2010 ITHEA® – Publisher; Sofia, 1000, P.O.B. 775, Bulgaria. www.ithea.org ; e-mail: info@foibg.com

© 2010 Krassimir Markov, Vitalii Velychko, Lius Fernando de Mingo Lopez, Juan Casellanos – Editors

© 2010 Ina Markova – Technical editor

© 2010 For all authors in the book.

® ITHEA is a registered trade mark of FOI-COMMERCE Co.

ISBN 978-954-16-0044-9

C\o Jusautor, Sofia, 2010

PREFACE

ITHEA International Scientific Society (**ITHEA ISS**) is aimed to support growing collaboration between scientists from all over the world.

The scope of the books of the ITHEA ISS covers the area of Informatics and Computer Science.

ITHEA ISS welcomes scientific papers and books connected with any information theory or its application.

ITHEA ISS rules for preparing the manuscripts are compulsory.

ITHEA Publishing House is the official publisher of the works of the members of the ITHEA ISS.

Responsibility for papers and books published by ITHEA belongs to authors.

This book maintains articles on actual problems of research and application of information technologies, especially the new approaches, models, algorithms and methods of:

- membrane computing and transition P systems;
- decision support systems;
- discrete mathematics;
- problems of the interdisciplinary knowledge domain including informatics, computer science, control theory, and IT applications;
- information security;
- disaster risk assessment, based on heterogeneous information (from satellites and in-situ data, and modelling data);
- timely and reliable detection, estimation, and forecast of risk factors and, on this basis, on timely elimination of the causes of abnormal situations before failures and other undesirable consequences occur;
- models of mind, cognizers;
- computer virtual reality;
- virtual laboratories for computer-aided design;
- open social info-educational platforms;
- multimedia digital libraries and digital collections representing the European cultural and historical heritage;
- recognition of the similarities in architectures and power profiles of different types of arrays, adaptation of methods developed for one on others and component sharing when several arrays are embedded in the same system and mutually operated.

It is represented that book articles will be interesting for experts in the field of information technologies as well as for practical users.

The book is recommended for publication by the Scientific Council of the Institute of Information Theories and Applications FOI ITHEA.

Papers in this book are selected from the ITHEA ISS Joint International Events of Informatics "ITA 2010":

CFDM	Second International Conference "Classification, Forecasting, Data Mining"
i-i-i	International Conference "Information - Interaction – Intellect"
i.Tech	Eight International Conference "Information Research and Applications"
ISK	V-th International Conference "Informatics in the Scientific Knowledge"
MeL	V-th International Conference "Modern (e-) Learning"
KDS	XVI-th International Conference "Knowledge - Dialogue – Solution"
CML	XII-th International Conference "Cognitive Modeling in Linguistics"
INFOS	Thirth International Conference "Intelligent Information and Engineering Systems"
NIT	International Conference "Natural Information Technologies"
GIT	Eight International Workshop on General Information Theory
ISSI	Forth International Summer School on Informatics

ITA 2010 took place in Bulgaria, Croatia, Poland, Spain and Ukraine. It has been organized by
ITHEA International Scientific Society

in collaboration with:

- ITHEA International Journal "Information Theories and Applications"
- ITHEA International Journal "Information Technologies and Knowledge"
- Institute of Information Theories and Applications FOI ITHEA
- Universidad Politecnica de Madrid (Spain)
- V.M.Glushkov Institute of Cybernetics of National Academy of Sciences of Ukraine
- Taras Shevchenko National University of Kiev (Ukraine)
- University of Calgary (Canada)
- BenGurion University (Israel)
- University of Hasselt (Belgium)
- Dorodnicyn Computing Centre of the Russian Academy of Sciences (Russia)
- Institute of Linguistics, Russian Academy of Sciences (Russia)
- Association of Developers and Users of Intelligent Systems (Ukraine)
- Institute of Mathematics and Informatics, BAS (Bulgaria)
- Institute of Mathematics of SD RAN (Russia)
- New Bulgarian University (Bulgaria)
- The University of Zadar (Croatia)
- Rzeszow University of Technology (Poland)
- Kharkiv National University of Radio Electronics (Ukraine)
- Kazan State University (Russia)
- Alexandru Ioan Cuza University (Romania)
- Moscow State Linguistic University (Russia)
- Astrakhan State University (Russia)

as well as many other scientific organizations. For more information: www.ithea.org .

We express our thanks to all authors, editors and collaborators as well as to the General Sponsor.

The great success of ITHEA International Journals, International Books and International Conferences belongs to the whole of the ITHEA International Scientific Society.

Sofia – Kiev - Madrid

June 2010

K. Markov, V. Velychko, L. F. de Mingo Lopez, J. Casellanos

TABLE OF CONTENTS

Preface	3
Table of Contents	5
Index of Authors	7
Benchmark of Pso-DE Using Bbob 2010	
<i>Nuria Gómez Blas, Luis F. de Mingo</i>	9
Intelligent P-systems	
<i>Alberto Arteta , Angel Luis Castellanos, Jose Luis Sanchez</i>	15
Communication Lateness in Software Membranes	
<i>Miguel Angel Peña, Jorge Tejedor, Juan B. Castellanos, Ginés Bravo</i>	23
A Bounded Algorithm Based on Applicability Domains for the Application of Active Rules in Transition P-systems	
<i>F. Javier Gil, Jorge A. Tejedor, Luis Fernández</i>	29
Collision Detection and Treatment Using 2D Reconfigurable Hardware	
<i>Alejandro Figueroa, Gustavo Méndez, Francisco J. Cisneros, Adriana Toni</i>	39
Bacterial Technology to Build Computers: a Survey	
<i>Paula Cordero, Sandra Gómez, Rafael Gonzalo</i>	48
Chain Split of Partially Ordered Set of k-Subsets	
<i>Hasmik Sahakyan, Levon Aslanyan</i>	55
Upper Bound on Rate-Reliability-Distortion Function for Source with Two-Sided State Information	
<i>Mariam Haroutunian, Arthur Muradyan</i>	66
Using the Group Multichoice Decision Support System for Solving Sustainable Building Problems	
<i>Filip Andonov, Mariana Vassileva</i>	74
Influence Analysis of Information Technologies on Progress in Control Systems for Complex OBJECTS	
<i>Boris Sokolov, Rafael Yusupov, Michael Okhtilev, Oleg Maydanovich</i>	78
Flood Risk Assessment Based on Geospatial Data	
<i>Nataliia Kussul, Sergii Skakun, Andrii Shelestov, Yarema Zyelyk</i>	92
Large VLSI Arrays – Power and Architectural Perspectives	
<i>Adam Teman, Orly Yadid-Pecht and Alexander Fish</i>	102

System Approach to Estimation of Guaranteed Safe Operation of Complex Engineering Systems	
<i>Nataliya Pankratova</i>	115
Soa Protocol with Multiresulting	
<i>Michał Plewka, Roman Podraza</i>	129
Calculating of Reliability Parameters of Microelectronic Components and Devices by Means of Virtual Laboratory	
<i>Oleksandr Palagin, Peter Stanchev, Volodymyr Romanov, Krassimir Markov, Igor Galelyuka, Vitalii Velychko, Oleksandra Kovyriova, Oksana Galelyuka, Iliya Mitov, Krassimira Ivanova</i>	134
Optimizing Routing Process with a Kinetic Method	
<i>Olexandr Kuzomin, Ievgen Kozlov</i>	144
Methods of Analysis for the Information Security Audit	
<i>Natalia Ivanova, Olga Korobulina, Pavel Burak</i>	152
On Measurable Models of Promotion of Negentropic Strategies by Cognition	
<i>Pogossian Edward</i>	161
Systemological Classification Analysis in Conceptual Knowledge Modeling	
<i>Mikhail Bondarenko, Nikolay Slipchenko, Kateryna Solovyova, Viktoriia Bobrovska, Andrey Danilov</i>	169
Search and Administrative Services in Iconographical Digital Library	
<i>Desislava Paneva-Marinova, Radoslav Pavlov, Maxim Goynov, Lilia Pavlova-Draganova, Lubomil Draganov</i>	177
The Influence of the Computer Game's Virtual Reality upon the Human Psychology	
<i>Helen Shynkarenko, Viktoriya Tretiyachenko</i>	188
EDUKIT: Info-Educational Platform Enabling to Create Websites for Secondary Schools	
<i>O.Y. Stepanovsky, S.O. Ryzhikova, D.P. Nechiporenko, B.S. Elkin, O.B. Elkin, I.V. Garyachevska, O.M. Dyachenko, D.V. Fastova</i>	197

INDEX OF AUTHORS

Adam Teman	102	Mariam Haroutunian	66
Adriana Toni	39	Mariana Vassileva	74
Alberto Arteta	15	Maxim Goynov	177
Alejandro Figueroa	39	Michael Okhtilev	78
Alexander Fish	102	Michał Plewka	129
Andrey Danilov	169	Miguel Angel Peña	23
Andrii Shelestov	92	Mikhail Bondarenko	169
Angel Luis Castellanos	15	Natalia Ivanova	152
Arthur Muradyan	66	Nataliia Kussul	92
Boris Sokolov	78	Nataliya Pankratova	115
Borys S. Elkin	197	Nikolay Slipchenko	169
Darya V. Fastova	197	Nuria Gómez Blas	9
Desislava Paneva-Marinova	177	Oksana Galelyuka	134
Dmitriy P. Nechiporenko	197	Oleg Maydanovich	78
Edward Pogossian	161	Oleksandr Elkin	197
F. Javier Gil,	29	Oleksandr Palagin	134
Filip Andonov	74	Oleksandra Kovyrivova	134
Francisco J. Cisneros	39	Oleksiy Stepanovskiy	197
Ginés Bravo	23	Olesya Dyachenko	197
Gustavo Méndez	39	Olexandr Kuzomin	144
Hasmik Sahakyan	55	Olga Korobulina	152
Helen Shynkarenko	188	Orly Yadid-Pecht	102
Ievgen Kozlov	144	Paula Cordero	48
Igor Galelyuka	134	Pavel Burak	152
Iliya Mitov	134	Peter Stanchev	134
Iryna V. Garyachevska	197	Radoslav Pavlov	177
Jorge A. Tejedor	23, 29	Rafael Gonzalo	48
Jose Luis Sanchez	15	Rafael Yusupov	78
Juan B. Castellanos	23	Roman Podraza	129
Kateryna Solovyova	169	Sandra Gómez	48
Krassimir Markov	134	Sergii Skakun	92
Krassimira Ivanova	134	Svitlana Ryzhikova	197
Levon Aslanyan	55	Viktorii Bobrovska	169
Lilia Pavlova-Draganova	177	Viktoriya Tretiyachenko	188
Lubomil Draganov	177	Vitalii Velychko	134
Luis F. de Mingo	9	Volodymyr Romanov	134
Luis Fernández	29	Yarema Zyelyk	92

BENCHMARK OF PSO-DE USING BBOB 2010

Nuria Gómez Blas, Luis F. de Mingo

Abstract: As an example, we benchmark the Particle Swarm Optimization algorithm with a Differential Evolution on the noise-free Black Box Optimization Benchmark 2010 testbed. Each candidate solution is sampled uniformly in $[-5, 5]^D$, where D denotes the search space dimension, and the evolution is performed with a classical PSO algorithm and a classical DE/x/1 algorithm according to a random threshold. The maximum number of function evaluations is chosen as 10^5 times the search space dimension. This paper shows how to evaluate the performance of a given optimization algorithm using the BBOB 2010.

Keywords: Benchmarking, Black-box optimization, Direct search, Evolutionary computation, Particle Swarm Optimization, Differential Evolution

Categories: G.1.6 [Numerical Analysis]: Optimization-global optimization, unconstrained optimization ; F.2.1 [Analysis of Algorithms and Problem Complexity]: Numerical Algorithms and Problems.

Introduction

Particle swarm optimization (PSO) is a global optimization algorithm for dealing with problems in which a best solution can be represented as a point or surface in an n -dimensional space. Hypotheses are plotted in this space and seeded with an initial velocity, as well as a communication channel between the particles. Particles then move through the solution space, and are evaluated according to some fitness criterion after each timestep. Over time, particles are accelerated towards those particles within their communication grouping which have better fitness values. The main advantage of such an approach over other global minimization strategies such as simulated annealing is that the large number of members that make up the particle swarm make the technique impressively resilient to the problem of local minima [7, 8, 9].

Equations used in the particle swarm optimization training process are the following ones, where c_1 and c_2 are two positive constants, R_1 and R_2 are two random numbers belonging to $[0, 1]$ and w is the inertia weight. This equations define how the genotype values are changing along iterations.

$$v_{in}(t+1) = wv_{in}(t) + c_1R_1(p_{in} - x_{in}(t)) + c_2R_2(p_{gn} - x_{in}(t))$$

$$x_{in}(t+1) = x_{in}(t) + v_{in}(t+1)$$

Previous equations will be modified until a stop condition is achieved, that is, a lower mean squared error or a maximum number of iterations is reached.

Differential Evolution (DE) is an evolutionary algorithm [10, 11, 12] that uses a differential mutation procedure that consists in the addition of the weighted difference of two population vectors to a third vector. Many variants of the differential mutation procedure exist. Choosing between these variants and setting parameters requires preliminary testing as [11] admits that the results of the algorithm are dependent on the chosen strategy and the choice of parameter. DE/local-to-best/1 is a variant where instead of the base vector x_{i1} being chosen in the

population vector, it is chosen to lie between the vector considered and the best vector so far, thus the update of the velocity is written as follows, where F is a constant in the range $[0, 2]$:

$$\mathbf{v}_i = \mathbf{x}_i + F(\mathbf{x}_{\text{best}} - \mathbf{x}_i) + F(\mathbf{x}_{i_2} - \mathbf{x}_{i_1}),$$

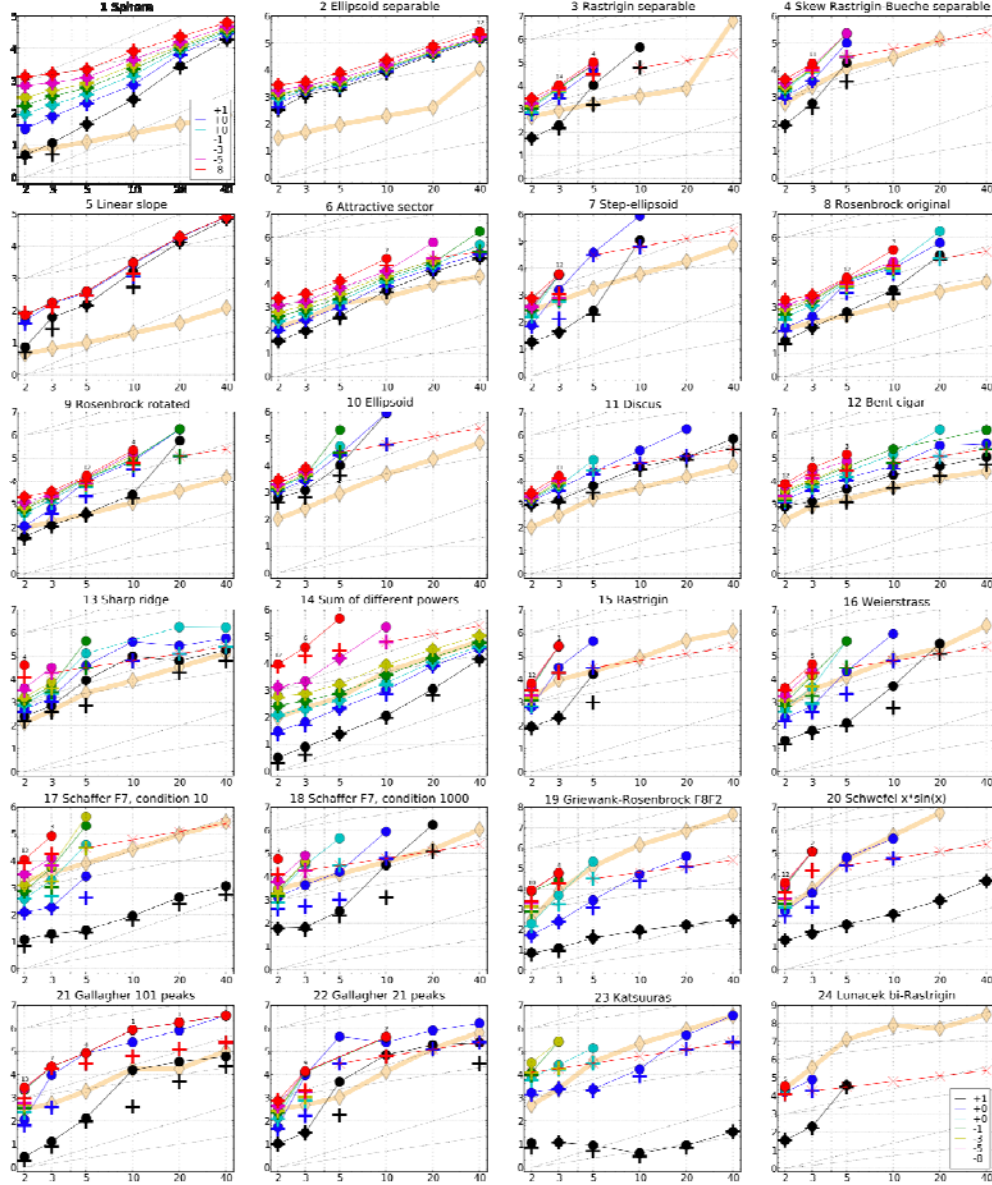


Figure 1: Expected Running Time (ERT, ●) to reach $f_{\text{opt}} + \Delta f$ and median number of f -evaluations from successful trials (+), for $\Delta f = 10^{l+1,0,-1,-2,-3,-3,-5}$ (the exponent is given in the legend of f_1 and f_{24}) versus dimension in log-log presentation. For each function and dimension, $\text{ERT}(\Delta f)$ equals to $\#FEs(\Delta f)$ divided by the number of successful trials, where a trial is successful if $f_{\text{opt}} + \Delta f$ was surpassed. The $\#FEs(\Delta f)$ are the total number (sum) of f -evaluations while $f_{\text{opt}} + \Delta f$ was not surpassed in the trial, from all (successful and unsuccessful) trials, and f_{opt} is the optimal function value. Crosses (×) indicate the total number of f -evaluations, $\#FEs(-\infty)$, divided by the number of trials. Numbers above ERT-symbols indicate the number of successful trials. Y-axis annotations are decimal logarithms. The thick light line with diamonds shows the single best results from BBOB-2009 for $\Delta f = 10^{-8}$. Additional grid lines show linear and quadratic scaling.

Method

We have used a uniform sampling in $[-5, 5]^D$, where D denotes the dimension of the search space. The experiments according to [3] on the benchmark functions given in [2, 4] have been conducted using a C-code. A maximum of $10^5 \times D$ function evaluations has been used.

f_1 in 5-D, $N=15$, $mFE=10000$				f_1 in 20-D, $N=15$, $mFE=10000$				f_2 in 5-D, $N=15$, $mFE=10000$				f_2 in 20-D, $N=15$, $mFE=10000$							
Δf	# ERT	10%	90%	RT _{success}	# ERT	10%	90%	RT _{success}	# ERT	10%	90%	RT _{success}	# ERT	10%	90%	RT _{success}			
10	15	1.1e1	4.8e1	4.8e1	15	2.2e3	1.1e4	2.6e4	10	15	2.1e3	1.2e4	5.1e3	2.1e3	15	3.3e3	2.2e4	4.2e4	
1	15	2.0e2	9.0e1	2.9e2	15	6.7e3	3.2e3	9.6e3	1	15	2.6e3	1.3e3	2.6e3	2.5e3	15	4.1e3	3.3e3	3.4e3	
in-1	15	3.8e2	2.8e2	2.2e2	15	5.0e3	7.0e2	1.1e4	1	15	3.1e3	2.2e3	4.2e3	4.1e3	15	4.3e3	3.5e3	4.2e3	
in-3	15	7.0e2	4.8e2	2.1e2	15	1.3e4	8.3e2	1.6e4	1	15	3.8e3	3.1e3	5.1e3	4.6e3	15	4.9e3	4.0e3	4.5e3	
in-5	15	1.2e3	1.1e3	1.3e3	15	1.6e4	1.3e3	1.6e4	1	15	4.6e3	3.7e3	7.1e3	4.8e3	15	5.3e3	4.5e3	7.0e3	
in-8	15	2.3e3	2.0e3	2.5e3	15	2.2e4	2.1e3	2.6e4	1	15	8.1e3	8.1e3	8.1e3	8.1e3	15	7.8e3	8.5e3	7.9e3	
f_3 in 5-D, $N=15$, $mFE=10000$				f_3 in 20-D, $N=15$, $mFE=10000$				f_4 in 5-D, $N=15$, $mFE=10000$				f_4 in 20-D, $N=15$, $mFE=10000$							
10	15	1.1e1	3.2e1	3.2e1	10	1.1e3	1.3e3	2.3e3	10	11	1.0e3	1.3e3	1.3e3	10	2.5e3	3.3e3	3.3e3		
1	7	4.9e3	3.1e3	1.1e3	1	4	1.0e3	1.3e3	1.3e3	1	4	1.0e3	1.3e3	3.0e3	1.6e3	2	2.3e3	3.0e3	4.4e3
in-1	8	7.8e3	6.3e3	2.1e3	1	2	2.3e3	3.0e3	5.5e3	1	2	2.3e3	3.0e3	4.4e3	2.7e4	2	2.3e3	3.0e3	4.4e3
in-3	3	8.0e3	1.1e4	2.2e3	1	2	2.3e3	3.0e3	4.4e3	1	2	2.3e3	3.0e3	4.4e3	2.7e4	2	2.3e3	3.0e3	4.4e3
in-5	3	8.2e3	1.0e4	1.7e3	1	2	2.3e3	3.0e3	4.4e3	1	2	2.3e3	3.0e3	4.4e3	2.7e4	2	2.3e3	3.0e3	4.4e3
in-8	4	1.0e4	1.4e4	2.4e3	1	2	2.3e3	3.0e3	4.4e3	1	2	2.3e3	3.0e3	4.4e3	2.7e4	2	2.3e3	3.0e3	4.4e3

Table 1: Shown are, for a given target difference to the optimal function value Δf : the number of successful trials (#); the expected running time to surpass $f_{opt} + \Delta f$ (ERT, see Figure 1); the 10%-tile and 90%-tile of the bootstrap distribution of ERT; the average number of function evaluations in successful trials or, if none was successful, as last entry the median number of function evaluations to reach the best function value (RT_{success}). If $f_{opt} + \Delta f$ was never reached, figures in *italics* denote the best achieved Δf -value of the median trial and the 10% and 90%-tile trial. Furthermore, N denotes the number of trials, and mFE denotes the maximum of number of function evaluations executed in one trial. See Figure 1 for the names of functions.

The simulations for 2; 5; 10; 20 and 40 D were done with the C-code and took 2 hours and a half. No parameter tuning was done and the crafting effort CrE [3] is computed to zero.

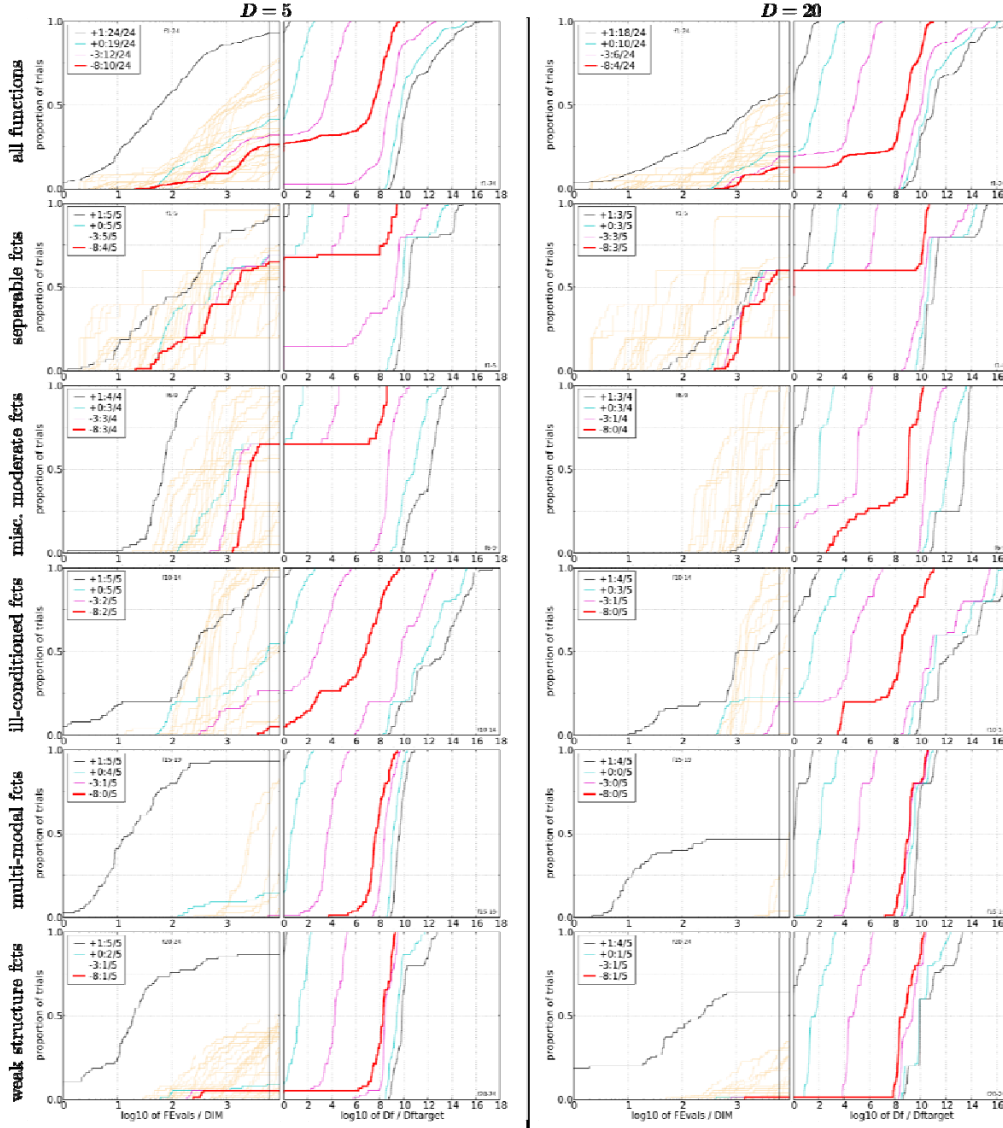


Figure 2: Empirical cumulative distribution functions (ECDFs), plotting the fraction of trials versus running time (left subplots) or versus Δf (right subplots). The thick red line represents the best achieved results. Left subplots: ECDF of the running time (number of function evaluations), divided by search space dimension D , to fall below $f_{opt} + \Delta f$ with $\Delta f = 10^k$, where k is the first value in the legend. Right subplots: ECDF of the best achieved Δf divided by 10^k (upper left lines in continuation of the left subplot), and best achieved Δf divided by 10^{-8} for running times of $D, 10D, 100D \dots$ function evaluations (from right to left cycling black-cyan-magenta). The legends indicate the number of functions that were solved in at least one trial. FEvals denotes number of function evaluations, D and DIM denote search space dimension, and Δf and Df denote the difference to the optimal function value. Light brown lines in the background show ECDFs for target value 10^{-8} of all algorithms benchmarked during BBOB-2009.

Results

Results from experiments according to [2] on the benchmarks functions given in [1, 3] are presented in Figures 1, 2 and 3 and in Tables 1 and 2. The algorithm solves some of the moderate functions f1, f2, f5, f6, f14 and f21. Else, f8, f9, f11, f12, f13 are partially solved for dimensions 20.

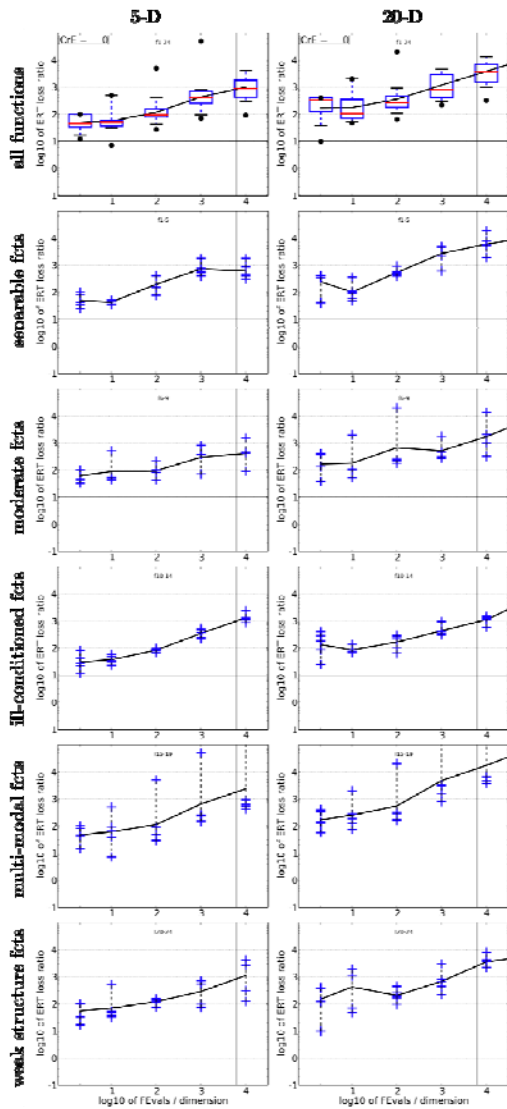


Figure 3: ERT loss ratio versus given budget FEvals. The target value f_t for ERT (see Figure 1) is the smallest (best) recorded function value such that $ERT(f_t) \leq FEvals$ for the presented algorithm. Shown is FEvals divided by the respective best ERT(f_t) from BBOB-2009 for functions f_1-f_{24} in 5-D and 20-D. Each ERT is multiplied by $\exp(CrE)$ correcting for the parameter crafting effort. Line: geometric mean. Box-Whisker error bar: 25-75%-ile with median (box), 10-90%-ile (caps), and minimum and maximum ERT loss ratio (points). The vertical line gives the maximal number of function evaluations in this function subset.

Table 2: ERT loss ratio (see Figure 3) compared to the respective best result from BBOB-2009 for budgets given in the first column. The last row RL_{US}/D gives the number of function evaluations in unsuccessful runs divided by dimension. Shown are the smallest, 10%-ile, 25%-ile, 50%-ile, 75%-ile and 90%-ile value (smaller values are better).

		f_1-f_{24} in 5-D, $maxFE/D=6158$					
#FEs/D		best	10%	25%	med	75%	90%
2		1.2	1.7	3.1	4.4	10	10
10		0.72	3.2	3.5	4.9	5.8	50
100		2.8	4.3	7.5	9.5	15	40
1e3		7.1	9.4	24	38	70	91
1e4		9.1	29	43	87	1.7e2	4.0e2
RL_{US}/D		6e3	6e3	6e3	6e3	6e3	6e3
		f_1-f_{24} in 20-D, $maxFE/D=6152$					
#FEs/D		best	10%	25%	med	75%	90%
2		1.0	3.6	11	31	40	40
10		4.7	5.1	7.2	10	32	2.0e2
100		6.4	11	18	26	45	2.8e2
1e3		22	29	41	80	3.0e2	4.7e2
1e4		32	94	1.4e2	3.7e2	6.7e2	1.4e3
1e5		99	2.0e2	3.5e2	6.2e2	3.7e3	6.2e3
RL_{US}/D		6e3	6e3	6e3	6e3	6e3	6e3

Conclusion

We have presented the results of the Particle Swarm Optimization algorithm with a Differential Evolution term, that does use information gathered during search for guiding its next steps following a social behavior not a genetic one. Those results provide a baseline comparison that every adaptive algorithm should outperform. Results have been obtained using the Black Box Optimization Benchmark 2010, which provides useful tools to analyze data in a graphical way.

Bibliography

- [1] S. H. Brooks. A discussion of random methods for seeking maxima. *Operations Research*, 6:244–251, 1958.
- [2] S. Finck, N. Hansen, R. Ros, and A. Auger. Real-parameter black-box optimization benchmarking 2009: Presentation of the noiseless functions. Technical Report 2009/20, Research Center PPE, 2009.
- [3] N. Hansen, A. Auger, S. Finck, and R. Ros. Real-parameter black-box optimization benchmarking 2009: Experimental setup. Technical Report RR-6828, INRIA, 2009.
- [4] N. Hansen, S. Finck, R. Ros, and A. Auger. Real-parameter black-box optimization benchmarking 2009: Noiseless functions definitions. Technical Report RR-6829, INRIA, 2009.
- [5] M. J. D. Powell. The NEWUOA software for unconstrained optimization without derivatives. *Large Scale Nonlinear Optimization*, pages 255–297, 2006.
- [6] J. Nelder and R. Mead. The downhill simplex method. *Computer Journal*, 7:308–313, 1965.
- [7] T Jayabarathi, Sandeep Chalasani, Zameer Ahmed Shaik, Nishchal Deep Kodali; "Hybrid Differential Evolution and Particle Swarm Optimization Based Solutions to Short Term Hydro Thermal Scheduling", *WSEAS Transactions on Power Systems Issue 11, Volume 2*, pp. , ISSN: 1790-5060, 2007.
- [8] Piao Haiguo, Wang Zhixin, Zhang Huaqiang, "Cooperative-PSO-Based PID Neural Network Integral Control Strategy and Simulation Research with Asynchronous Motor Controller Design", *WSEAS Transactions on Circuits and Systems Volume 8*, pp. 136-141, ISSN: 1109-2734, 2009.
- [9] Lijia Ren, Xiuchen Jiang, Gehao Sheng, Wu B;"A New Study in Maintenance for Transmission Lines", *WSEAS Transactions on Circuits and Systems Volume 7*, pp. 53-37, ISSN: 1109-2734, 2008.
- [10] Kenneth Price. Differential evolution vs. the functions of the second ICEO. In *Proceedings of the IEEE International Congress on Evolutionary Computation*, pages 153–157, 1997.
- [11] Kenneth Price, Rainer M. Storn, and Jouni A. Lampinen. *Differential Evolution: A Practical Approach to Global Optimization (Natural Computing Series)*. Springer- Verlag New York, Inc., 2005. ISBN 3540209506. URL <http://portal.acm.org/citation.cfm?id=1121631>.
- [12] K.V. Price. Differential evolution: a fast and simple numerical optimizer. In *Fuzzy Information Processing Society, 1996. NAFIPS. 1996 Biennial Conference of the North American*, pages 524–527, 1996. doi: {10.1109/NAFIPS.1996.534790}.

Authors' Information

Nuria Gómez Blas – Associate professor U.P.M Crtra Valencia km 7, Madrid-28031, Spain; e-mail: ngomez@eui.upm.es

Research: DNA computing, Membrane computing, Education on Applied Mathematics and Informatics

Luis F. de Mingo – Associate professor U.P.M Crtra Valencia km 7, Madrid-28031, Spain; e-mail: lfmingo@eui.upm.es

Research: Artificial Intelligence, Social Intelligence, Education on Applied Mathematics and Informatics

INTELLIGENT P-SYSTEMS

Alberto Arteta , Angel Luis Castellanos, Jose Luis Sanchez

Abstract: Membrane computing is a recent area that belongs to natural computing. This field works on computational models based on nature's behavior to process the information. Recently, numerous models have been developed and implemented with this purpose. P-systems are the structures which have been defined, developed and implemented to simulate the behavior and the evolution of membrane systems which we find in nature. However no other technology has been used to simulate the behavior of the living cells. In this paper we present a proposal for a set of reactive robots that receives, process and learn from the living cells through the use of p-systems. When analyzing the properties of the proposed robot. we realize that we get interesting points compared to traditional p-systems.

Keywords: Autonomous robots, membrane computing, p-systems, artificial intelligence.

Introduction

Natural computing is a new field within computer science which develops new computational models. These computational models can be divided into three major areas:

1. Neuronal networks.
2. Genetic Algorithms
3. Biomolecular computation.

Membrane computing is included in biomolecular computation. Within the field of membrane computing a new logical computational device appears: The P-system. These P-systems are able to simulate the behavior of the membranes on living cells. This behavior refers to the way membranes process information. (Absorbing nutrients, chemical reactions, dissolving, etc)

Membrane computing formally represents, through the use of P-systems, the processes that take place inside of the living cells. In terms of software systems, it is the process within a complex and distributed software. In parallel computational models, p-systems might be as important as the Turing machine is in sequential computational models.

In this paper, we design an autonomous robot that it is capable to simulate the behavior the living cells to solve known problems through the use of p-systems. Although this has been done so far by traditional p-systems [1], we propose a new model. We will design several autonomous robots that behave as the living cells when processing information. This way these robots can obtain solutions to known problems. The interesting part here is that we state that performance improves by reducing computational complexity. Traditional P-systems are the structures used within membrane computing to do the living cells simulation. P-systems evolve in a parallel a non deterministic way. By using an autonomous robot, we will eliminate the non-determinism as the evolution will be orientated to obtain results in a faster way.

In the paper we are going to go through different topics: (1) Introduction to P-systems theory; (2) Description of autonomous robot; (3) Theoretical model of a robot processing information from the living cells; (4) Traditional P-system vs P-systems with robots; (5) Conclusions and further work.

Introduction to P-systems theory

In this section we will study into detail all of the theories related to the paradigm of the P-systems. A P-system is a computational model inspired by the way the living cells interact with each other through their membranes. The elements of the membranes are called objects. A region within a membrane can contain objects or other membranes. A p-system has an external membrane (also called skin membrane) and it also contains a hierarchical relation defined by the composition of the membranes. A multiset of objects is defined within a region (enclosed by a membrane). These multisets of objects show the number of objects existing within a region. Any object 'x' will be associated to a multiplicity which tells the number of times that 'x' is repeated in a region.

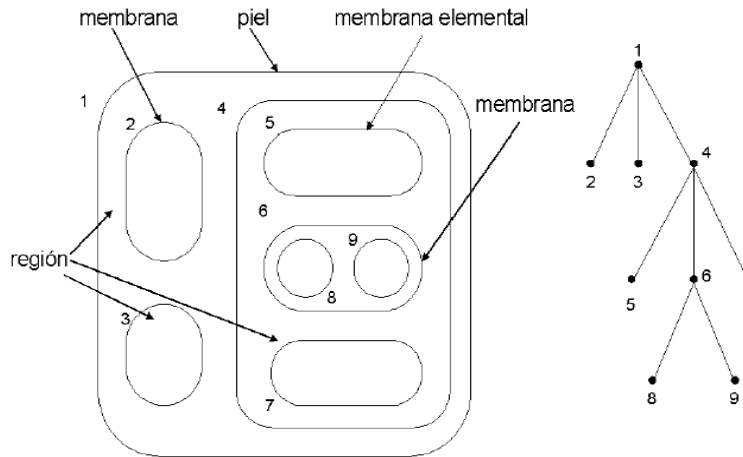


Fig. 1. The membrane's structure (left) represented in tree shape (right)

According to Păun 's definition, a transition P System of degree n , $n > 1$ is a construct: [Păun 1998]

$$\Pi = (V, \mu, \omega_1, \dots, \omega_n, (R_1, \rho_1), \dots, (R_n, \rho_n), i_0)$$

where:

V is an alphabet; its elements are called objects;

μ is a membrane structure of degree n , with the membranes and the regions labeled in a one-to-one manner with elements in a given set ; in this section we always use the labels $1, 2, \dots, n$;

ω_i $1 \leq i \leq n$, are strings from V^* representing multisets over V associated with the regions $1, 2, \dots, n$ of μ

R_i $1 \leq i \leq n$, are finite set of evolution rules over V associated with the regions $1, 2, \dots, n$ of μ ; ρ_i is a partial order over R_i $1 \leq i \leq n$, specifying a priority relation among rules of R_i . An evolution rule is a pair (u, v) which we will usually write in the form $u \rightarrow v$ where u is a string over V and $v = v'$ or $v = v' \delta$ where v' is a string over $(V \times \{here, out\}) \cup (V \times \{in_j \mid 1 \leq j \leq n\})$, and δ is a special symbol not in V . The length of u is called the radius of the rule $u \rightarrow v$

i_0 is a number between 1 and n which specifies the output membrane of Π

Let U be a finite and not an empty set of objects and N the set of natural numbers. A *multiset of objects* is defined as a mapping:

$$M : V \rightarrow \mathbb{N}$$

$$a_i \rightarrow u_i$$

Where a_i is an object and u_i its multiplicity.

As it is well known, there are several representations for multisets of objects.

$$M = \{(a_1, u_1), (a_2, u_2), (a_3, u_3), \dots\} = a_1^{u_1} \cdot a_2^{u_2} \cdot a_n^{u_n} \dots$$

Evolution rule with objects in U and targets in T is defined by $r = (m, c, \delta)$

where $m \in M(V), c \in M(V \times T)$ and $\delta \in \{to\ dissolve, not\ to\ dissolve\}$

From now on 'c' will be referred to as the consequent of the evolution rule 'r'

The set of evolution rules with objects in V and targets in T is represented by $R(U, T)$.

We represent a rule as:

$x \rightarrow y$ or $x \rightarrow y\delta$ where x is a multiset of objects in $M((V) \times Tar)$ where $Tar = \{here, in, out\}$ and y is the consequent of the rule. When δ is equal to "dissolve", then the membrane will be dissolved. This means that objects from a region will be placed within the region which contains the dissolved region. Also, the set of evolution rules included on the dissolved region will disappear.

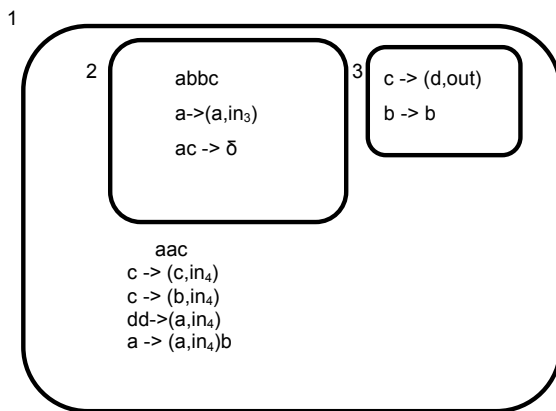


Fig 2 P-system with 3 regions and multiset of objects in each region

P-systems evolve, which makes it change upon time; therefore it is a dynamic system. Every time that there is a change on the p-system we will say that the P-system is in a new transition. The step from one transition to another one will be referred to as an evolutionary step, and the set of all evolutionary steps will be named computation. Processes within the p-system will be acting in a massively parallel and non-deterministic manner. (Similar to the way the living cells process and combine information). We will say that the computation has been successful if:

1. The halt status is reached.
2. No more evolution rules can be applied.
3. Skin membrane still exists after the computation finishes.

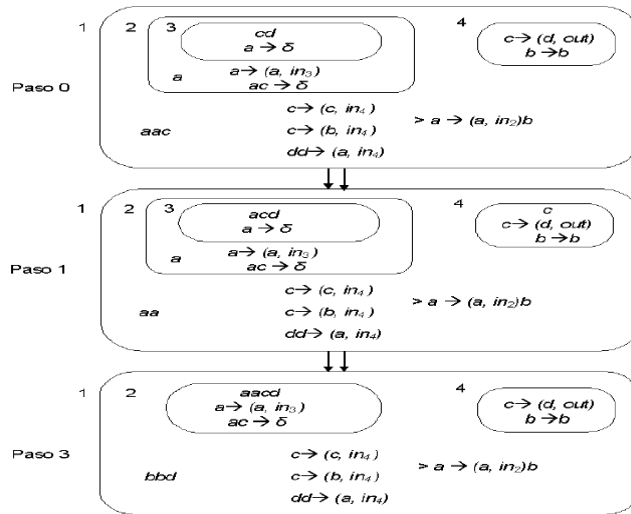


Fig 3. Example of the evolution of a p-system

Currently There are some simulations of p-systems in different languages and hardware[Arroyo, 2001], [Arroyo, 2003]

Autonomous robots

Nowadays humans and robots are used to interact with each other. We could say that robots try to simulate humans' actions. Part of those humans' actions is making decisions. If a robot can make decision, can we certainly state that the robot is intelligent? Today we still have no a unique definition of what intelligence means. What we know is that the concept: "intelligence" [Brooks, 1986] is related to:

- Understanding
- Problem solving
- Learning.

Even if we are able to create robots that solve problems we cannot assure that they understand or learn. That is the reason why the definition of intelligence does not cover all the possible meanings.

Although we cannot say that robots are intelligent, we can certainly establish a classification for them:

In general terms we can separate the robots into two classes:

- Deliberative robots
- Reactive Robots

The first kind is the traditional one. This types of robots work based on scheduled actions and known information.

Reactive robots interact with the environment, they observe, process information, and make decisions based on the environment they are.

From our perspective the second type is the interesting one. These robots can adapt to the environment they live and make decisions based on that. Not only making decisions but also solving problems. Adaptability is other characteristic of intelligence.

Another characteristic of traditional robotics is the use of a centralized system. This centralized system stores all the information of the environment. The information is represented in a symbolic way. After processing the information is possible to calculate the next action. The problem here is when the environment changes constantly. That is why traditional robots become useless in this scenario. The aim is trying to build a robot able to adapt to its environment regardless the type of environment it is. In that way, the robot will have some basic actions and it will create a bigger knowledge by learning from the environment. The learning process comes from the reactions that robots do.

Learning process is a complex one.

If a robot is created to experiment in a lab, it is possible to determine all the situations in where the robot is going to be. On the contrary, if the robot created wants to be useful in the real world, there is no way for us to determine how many different situations the robot is going to face. Thus, we must be able to create a robot that learns and therefore be able to react correctly when a new circumstance arises.

These are the different ways for implementing the learning process.

There are four types of learning.

- Auto-organized: It works with random variables.
- Supervised: any action has different data and variables.
- Hybrid: A combination of the two previous ones
- Feedback: The response from the environment influences on its actions.

The robots we propose will learn by receiving feedback from the environment.

Robots working as p-systems

Once we have seen the characteristics of p-systems and autonomous robots separately, we are going to propose a community of robots that act as a p-system but let us say as a learning p-system.

The scenario we propose is:

Since we have regions in the living cells we are going to place a robot in every region delimited by a membrane.

The way to do this is:

In a the regions, given a set of membranes $M = \{m_i \mid i \in \mathbb{N}, 1 \leq i \leq n\}$ where m_i is a membrane, we allocate a robot:

Thus, we will need to define a function that for each membrane or region

$$f_{robot} : M \rightarrow A$$

$$f_{robot}(m_i) = R_i \quad \forall i \in \mathbb{N} \quad i \leq n, \quad n \text{ number of membranes}$$

The p-system considers three major stages:

1. Static structure of the p-system
2. Dynamic behavior of the p-system
3. Synchronism between membranes.

1. The static structure of the p-systems is updated by the existence of a robot r_i

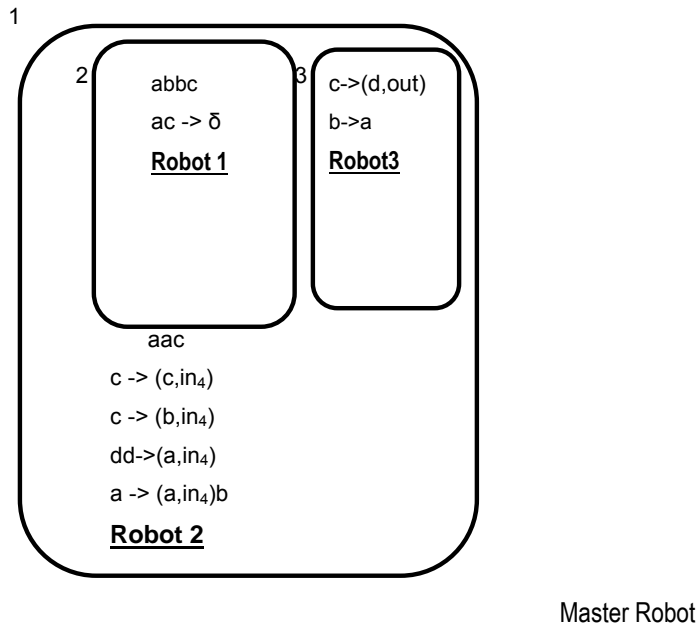


Fig 4 P-system controlled by reactive robots. The robot outside controls the evolution of the p-system.

Every robot controls a region. The robot is storing the information about the information processing for every region. The p-system evolves following the robots patterns.

For every region the robot store information and makes decisions based on the information is stored. Moreover the robot outside the p-system controls the execution steps of the p-system.

Robots learn from the times that evolution rules are applied and the results obtained.

The aim of this p-system supervised by robots is not just to find solutions to known problems, but also improving performance on that.

According to Paun's model all the rules application processes occur in every region in a parallel manner. Moreover the process is non deterministic.

Every robot R_i collects information about the membrane m_i in where he is allocated.

- The evolution rules that are applied in m_i and number of times that every rule r_i is applied to obtain a maximal multiset [Arteta, 2008]

For every region the robot R_i stores records as: $((r_1,1), (r_2,2), \dots, (r_i,7), \dots, (r_m,3))$

- The robot outside the p-systems:
 - Process the information arriving from all the robots.
 - It keeps the number of computation steps until the program finishes.
 - Stores information of the communication phase between membranes
 - It tells the robots what decisions to make in order to reduce computation steps. (Deterministic way)

The more time the p-system works the better results are obtained. At the end the robot will tell the robot in each membrane which evolution rules should be applied and also how many times those rules must be applied in order to obtain fast results. This will reduce the execution time.

Traditional p-systems versus p-systems with robots

Implementing membrane computing with robots shows a few advantages compared to the traditional ones.

The election for the evolution rules to be applied would be taken according the robot's indications. At first, robots will not be very accurate on the elections but during the learning process programs will take shorter to finish and to find solutions. Finding solutions in a faster way requires certain degree of what we call intelligence. P-systems with robots can be referred as intelligent p-systems. Although this is the main advantage we obtain when using p-systems with robots, there are certain disadvantages too:

- The need of auxiliary space to store information about the rules' election
- The need of extra time to calculate the times that the evolution rules have to be applied
- In the beginning the results might take longer than when using traditional p-systems. This might occur because as there are not enough information from the p-system computation, the decisions made by the robot about the rules could be worse than electing rules in a non deterministic way (as the traditional p-systems do)
- When changing non-determinism by intelligent elections, Paun's biological model is not useful anymore to implement the living cells which means that p-system of robots needs cannot be used to implement the living cells model.

Conclusions

This paper presents an update of what we know about membrane computing. P-systems are the structures that implement Paun's biological model. Membrane computing has proved to be able to reduce computational complexity when solving known problems as the "*knapsack problem*". What we propose is a new model based on some of the inherent properties of p-systems plus the learning capability which is achieved by a set of autonomous robots. Although non determinism is not a property of our system anymore, we can certainly state that this new model of p-systems can obtain optimal results to known problems due to the capacity of learning.

This idea is in an early stage because there are no formal implementations of membrane computing yet. However the idea is promising. A theoretical p-system that is able to learn by using intelligent systems as robots, can make right decisions when applying evolution rules in p-systems' regions.

A further study is necessary to define formally the new model created by the combination of p-systems and robots.

Bibliography

- [Păun, 1998] "Computing with Membranes", Journal of Computer and System Sciences, 61(2000), and Turku Center of Computer Science-TUCS Report n° 208, 1998.
- [Brooks, 1986] Achieving AI through Building Robots [Periodic publication] // AI Memo. - Massachusetts : [s.n.], 1986.
- [Arroyo, 2001] "Structures and Bio-language to Simulate Transition P Systems on Digital Computers," Multiset Processing
- [Arroyo, 2003] "A Software Simulation of Transition P Systems in Haskell, Membrane Computing,"
- [A. Arteta, 2008] "Algorithm for Application of Evolution Rules based on linear diophantic equations" Synasc 2008 (IEEE), Timisoara Romania September 2008[1] A. Syropoulos, E.G. Mamas, P.C. Allilones, K.T. Sotiriades "
-

Authors' Information

Alberto Arteta Albert – Associate professor U.P.M Crtra Valencia km 7, Madrid-28031, Spain; e-mail: aarteta@eui.upm.es

Research: Membrane computing, Education on Applied Mathematics and Informatics

Angel Luis Castellanos Peñuela – Associate professor U.P.M Ciudad Universitaria, Madrid, Spain; e-mail: angel.castellanos@upm.es

Research: Education on Applied Mathematics and Informatics

Jose Luis Sanchez Sanchez – Associate professor U.P.M Crtra Valencia km 7, Madrid-28031, Spain; e-mail: jlsanchez@eui.upm.es

Research: Membrane computing

COMMUNICATION LATENESS IN SOFTWARE MEMBRANES

Miguel Angel Peña, Jorge Tejedor, Juan B. Castellanos, Ginés Bravo

Abstract: This paper presents a study about the time of the communication phase of the P System implemented as software. After presenting the membranes as software components, is studied how communicate with each other, running on the same processor or another. We study the times and delays that occur if communication takes place via network. It presents the theoretical formulas governing the study, empirical studies, and adjustment factors which depend on communications. Finally, some examples that show facts for some P System; the distribution of membranes on several computers improves global times.

Keywords: P System, Membrane Systems, and Natural Computation.

ACM Classification Keywords: F.1.2 Modes of Computation, I.6.1 Simulation Theory, H.1.1 Systems and Information Theory

Introduction

In 1998 Gheorghe Paun introduced membrane computing [Paun, 2000] as a parallel computing model based on the biological notion of the cell. On the original model, there have been several variations in order to solve various problems, and improve computation times, in order to solve complex problems such as NP-complete, in times similar to the polynomial.

The idea behind a membrane system is based on the permeability of the same, and the internal changes taking place. Depending on the elements that are working, we can distinguish two main types, those which manipulate objects, and working with strings. The behavior is similar in both cases. In parallel, each membrane performs a series of rules with objects or strings you have, resulting in other objects or strings, which, using the permeability of the membrane can move to other membrane if they indicate their transformation rules. The great advantage of these systems is that each membrane is run independently of the others, so the runtime does not depend on the number of membranes.

From this model, many researchers have developed software implementations, with different points of view. Some of them have been based on hardware architectures, such as [Petreska, 2003], [Fernández, 2005] or [Martínez, 2006]. Others were based on software simulation, in different languages, such as [Cordón-Franco, 2004] or [Malita, 2000] in Prolog, [Suzuki, 2000] LISP based, [Arroyo, 2003] Haskell based, [Balbontín-Noval, 2002] Scheme based, [Nepomuceno-Chamorro, 2004] Java based. These simulators use only one processor to perform operations. For more performance software implementations should use the idea of a distributed architecture, as proposed [Ciobanu, 2004] in C++ and [Syropoulos, 2003] in Java. What is not mentioned in any of these implementations is how to make the distribution of membranes per processor. To carry out this distribution is necessary to know the time it takes communication to make the best possible distribution.

P- System definition

The first definition of a P System was made for Paun [Paun, 2000], a defined a Transition P System:

Definition 1. A Transition P System is $\Pi = (V, \mu, \omega_1, \dots, \omega_n; (R_1, \rho_1), \dots, (R_n, \rho_n); i_0)$, where:

V is an alphabet (composed of objects),

μ is the membrane structure with n membranes

ω_i are the multiset of symbols for the membrane i .

R_i are the evolution rules for the membrane i .

P_i are the priority of rules for the membrane i .

i_0 indicates a membrane, which is the membrane of system output.

Running P-System is made through configurations. The following phases are making on each configuration of each membrane in parallel and no deterministic way:

Determining the set of utility rules: On this micro-step are evaluating all evolution rules of the membrane in order to determine which are useful. A rule is useful when all membranes exists in the P-System that are indicated in its consequent

Determining the set of applicable rules: It will be necessary evaluate all the evolution rules of the membrane for identify those that meet that its predecessor is contained in the multiset of objects of the region.

Determining the set of actives rules: Intersection of two previous sets conform the entrance group to this micro-step. Each one of rules of the set must be processed, to determine which meets the condition of active rule. To determine if a rule meets that condition, is necessary check there is not another rule with higher priority that belongs to the useful and applicable rules set.

Non deterministic distribution of objects of the region between its active rules and application: In this micro-step, copies of present objects in the multiset of the region are distributed between active evolution rules of the same. Copies of objects that are assigned to each rule, match with those of the multiset that results from scalar product of a number between minimum and maximum bound of applicability of those rule and its predecessor. This distribution process is made on a non-deterministic way. Moreover, at the end of it, objects no assigned to any rule forms a multiset and they will be characterized because they do not contained to any predecessor of rules. The result of the distribution is a multiset of actives rules, where multiplicity of each rule defines the times that would be applied, and therefore, indirectly through its predecessors, objects are assigned to the multiset of the region. Objects used are eliminated and generate new objects that are destined for a membrane.

Transferring of multiset of generated objects or communication phase: In this micro-step, the new objects generated on the previous micro-step whose indicator of destiny membranes was 'in' or 'out', must be transferred to its corresponding membranes. Each membrane, will have unused objects in its application, together with those that result for the applying of rules and that have this membrane as destiny.

Implementation software

At [Gomez, 2007] is defined a framework for Transition P-System. Following this model has been done a software implementation on java. Each computer, with one processor, will contain a number of membranes that need not be equal. Furthermore, each computer will be known where all the membranes are. Computers will be connected via Ethernet. On each processor, will be running in a secuencial way each one of the phases that were indicated previously, on each one of membranes, in a secuencial way too. At finished each phase it will make synchronization between processors with the purpose that all of them will begin the next phase simultaneously. We used some computer at 0,7 GHz, with 0,5 MB of RAM. They were connecting with a 100 Ethernet.

Communication Phase

With the application of rules, is possible that one membrane generates objects that are destined to another membrane taking advantage the characteristic of permeability of the membrane. Because that, the new generated object must be sended, on communication phase, to its destiny membrane. On implementation over distributed software, is possible that this destiny membrane is in the same or in another machine. Because that,

the sending method will be different for each one of two cases that we find. In the case of sending over the same processor, and therefore the share memory, will take place immediately, with the processor in charge of the whole process. In the case of source and target membranes, are on distinct processors, is needed send objects through the network.

Because there are two types of connections between membranes – internal and external to the processor-, times of the communication phase not only depend on the number of objects to interchange, but also of disposition of membranes on the processors.

To analyze the communication between processors, firstly, we should define the message to send through the net. Although there may be multiple formats to the communication message, we go to use a simplification of it:

- The first four bytes represent the destiny membrane of objects. It reserves the highest number of membranes, to indicate that the message is used for a different use, like to synchronize executing, dissolution of membranes,.. The use of four bytes limits the number to 2^{24} membranes.
- The next byte represents the object of whom will exist n repetitions to add. Therefore, it exists a limitation to a vocabulary of 2^8 objects.
- The next two bytes, represent the repetitions that are being transferred, of previously indicated object. Therefore, it can transfer, only 2^{16} equal objects.
- Symbol and repetitions block can appear many times as it likes, until find the special object corresponding to value 28.

Hence, message size depends on quantity of distinct objects that must be sending, but like minimum will be 8 bytes. It has been made many tests to compare times it takes to send a determinate number of bytes. For taking time, it has been synchronized watches on both computers, on a level to seconds, and it has been made 100000 repetitions in a row of each test, in order that synchronism error was negligible. Results are show on figure 1 – left. On it observes that distinct teste on the same conditions, presents distinct results. It is because when executions are made over software, several factors influence the time:

- The processor executes several tasks and depending of Operative System and task manager, it can delay execution of process of execution of P-System.
- It can perform other input-output operations
- May be necessary to release memory and the interchange from swap area.
- There may be network congestion, as indicated by Ciobanu [Ciobanu, 2004].
- The Java Virtual Machine can introduce similar delays too

However, removing data that clearly show excessive delays, it can obtain values to adjust (Figure 1 right):

$$\text{Time} \approx 0,0023 \text{ Bytes}$$

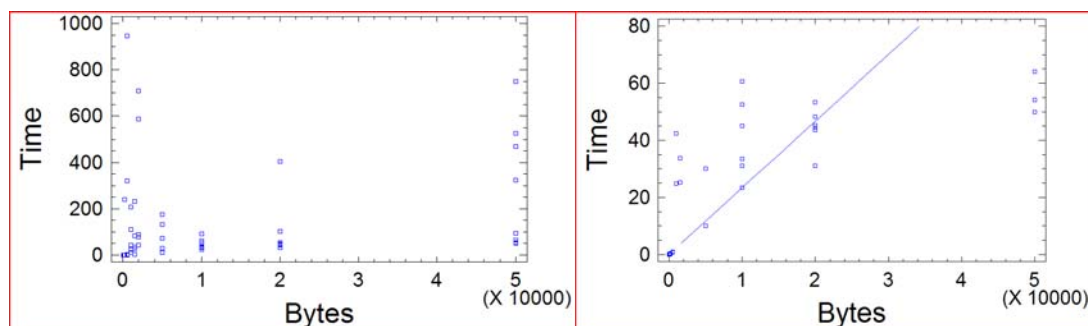


Figure 1. Time it takes to send the bytes over a network without removing delays (left) and eliminating (right).

The result shows that the time is practically lineal with respect to number of bytes, as pointed all hypotheses. That little variation is due to bytes number of header that added distinct existing protocols over network (TCP, IP...). In particular, this model explained a 88,5351% of variability. The correlation coefficient is 0.940931, indicating a relatively strong relation between variables.

Analyzing maximum values (Figure 2), we obtain Time $\approx 0,0032$ Bytes. The statistic 'R-square' indicates to model explain 82,7288% of variability on Time, after transformation of logarithmic scale to linearize model. The correlation coefficient is 0.909554, indicating a relatively strong relation between variables.

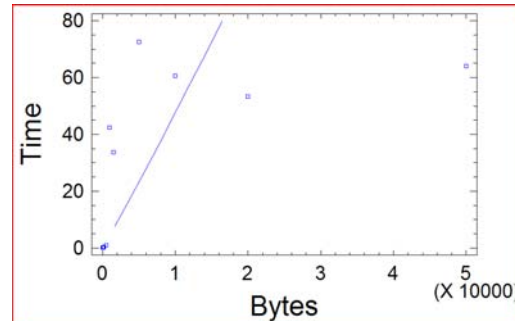


Figure 2. Maximum communication time.

When communication must be done over the same computer, transmission time is null.

Hence, we conclude, to send all equal objects of one membrane to another, it is needed a maximum time (on seconds) approximate:

$$T_{max} \approx 0,0256 \text{ object}$$

To know the needed time on communication phase is not enough with multiply the number of distinct objects to send by the time of each one, for each one of the membranes. Experimentally, we obtained values that rapidly change due to the same problems previously mentioned. Taking into account these deviations and the membranes are all connected with the same number of membranes and send the same number of objects, we get:

$$T \approx 58,42 + 8,16 \text{ Membrane} + 3,16 \text{ Object}$$

In [Tejedor, 2007], consider the communication time depends only on the transmission time of bytes and a constant. With this formula from empirical data find that this constant is dependent on the number of membranes that communicate. This is a direct result that the communication phase included the composition of the new multiset of objects, and this is done in each membrane.

Distribution of membranes

Taking the communication times and [Tejedor, 2007] studies, on the same P system and a certain number of processors, execution times depend on the distributions of the membranes are made. In Figure 3 we can see that both distributions present different times in their implementation. If we consider that all processors have three membranes, and consider the execution time of the application phase equal, it will not show differences. Internal communications are negligible with respect to external communications. Just analyze communications between processors.

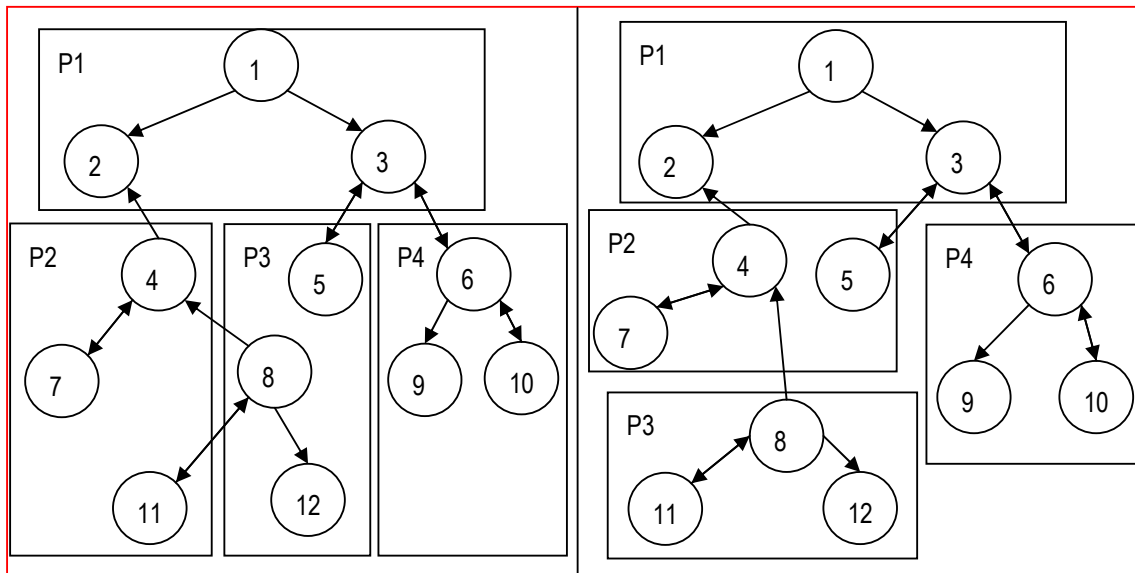


Figure 3. Two distributions of PSystem in four processors. The edges indicate how the symbols are sent from one membrane to another according to the rules.

We assume that communications between membranes only send one objects. The times to communicate will:

$$P1 \rightarrow T = 58,42 + 8,16 * 2 + 3,16 = 77,90 \quad P1 \rightarrow T = 58,42 + 8,16 * 2 + 3,16 = 77,90$$

$$P2 \rightarrow T = 58,42 + 8,16 * 2 + 3,16 = 77,90 \quad P2 \rightarrow T = 58,42 + 8,16 * 2 + 3,16 = 77,90$$

$$P3 \rightarrow T = 58,42 + 8,16 * 3 + 3,16 = 86,06 \quad P3 \rightarrow T = 58,42 + 8,16 * 1 + 3,16 = 69,74$$

$$P4 \rightarrow T = 58,42 + 8,16 * 1 + 3,16 = 69,74 \quad P4 \rightarrow T = 58,42 + 8,16 * 1 + 3,16 = 69,74$$

One can see that if messages are sent in parallel, or following a token ring architectures proposed by [Tejedor, 2007] or [Bravo, 2008], time in the second case is less.

Conclusion

It has been shown empirically that the time required for the communication phase will depend on the number of objects that are sent, and the different target membrane. We also observe that these times will be extended by different actors outside the system. Given these times, with a good distribution of the membranes in processor, will reduce communication time, and therefore, the whole times.

There remain many studies to be performed on the communication phase, such as the degree to which they affect a processor to communicate with various, or that the number of symbols is different for each membrane.

Bibliography

- [Arroyo, 2003] F. Arroyo, C. Luengo, A.V. Baranda, L.F. de Mingo. A software simulation of transition P System in Haskell. In: Lecture Notes in Computer Science 2597. Springer-Verlag, Berlin 2003, Pp: 19-32.
- [Balbotín-Noval, 2003] D. Balbotín-Noval, M.J. Pérez-Jiménez, F. Sancho-Caparrini. A MzScheme Implementation of Transition P System. IN: Membrane Computing, LNCS 2597, Spring Verlag. Berlin 2003. Pp: 57-73.
- [Bravo, 2008] G. Bravo, L. Fernández, F. Arroyo, M. Peña. Hierarchical Master-Slave Architecture for Membrane Systems Implementation. In: The 13th International Symposium on Artificial Life and Robotics. 2008. Pp: 485-490.
- [Ciobanu, 2004] G. Ciobanu, W. Guo. P System Running on a Cluster of Computers. In: Workshop on Membrane Computing, LNCS 2933. 2004. Pp: 123-2004.

- [Cordón-Franco, 2004] A. Cordón-Franco, M.A. Gutiérrez-Naranjo, M.J. Pérez-Jiménez, F. Sancho-Caparrini. A Prolog Simulator for Deterministic P Systems with Active Membranes. In: *New Generation Computing*, 22(4). 2004. Pp: 349-363.
- [Fernández, 2005] L. Fernandez, V.J. Martinez, F. Arroyo, L.F. Mingo, A Hardware Circuit for Selecting Active Rules in Transition P Systems. In: *First International Workshop on Theory and Application of P System*. September 2005. Pp: 45-48.
- [Gomez, 2007] S. Gómez, L. Fernández, I. García, F. Arroyo. Researching Framework for Simulating/Implementing P Systems. In: *Fifth International Conference Information Research and Applications (i.TECH-2007)*. Varna (Bulgary) June, 2007.
- [Malita, 2000] M. Malita. Membrane Computing in Prolog. In: *Preproceedings of the Workshop on multiset Processing*. 2000. Pp: 159-175.
- [Martínez, 2006] V. Martínez, L. Fernández, F. Arroyo, A. Gutiérrez. Implementation HW of an Enclosed Algorithm to Rules Applications in a Transition P System. In: *8-th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing*. 2006.
- [Nepomuceno-Chamorro, 2004] I.A. Nepomuceno-Chamorro. A Java Simulator for Basic Transition P System. In: *Journal of Universal Computer Science*. 2004. Pp: 620-619.
- [Paun, 2000] G. Paun. "Computing with membranes". In: *Journal of Computer and System Sciences*, 1(61). 2000. Pp: 108-143.
- [Petreska, 2003] B. Petreska, C. Teuscher. A hardware membrane system. In: *Preproceedings of the Workshop on Membrane Computing*. 2003. Pp: 242-255.
- [Suzuki, 2000] Y. Suzuki, H. Tamaka. On a Lisp implementation of a class of P Systems. *Romanian Journal of Information Science and Technology*. 2000. Pp: 173-186.
- [Syropoulos, 2003] A. Syropoulos, E.G. Mamatas, P.C. Allilomes, K.T. Sotiriades. A distributed simulation of P Systems. In: *Preproceedings of the Workshop on Membrane Computing*. 2003. Pp: 455-460.
- [Tejedor, 2007] J. Tejedor, L. Fernández, F. Arroyo, G. Bravo. An architecture for attacking the bottleneck communication in P systems. In: M. Sugisaka, H. Tanaka (eds.), *Proceedings of the 12th Int. Symposium on Artificial Life and Robotics*. 2007. Pp: 500-505.

Authors' Information

Miguel Angel Peña – Dept. *Inteligencia Artificial, Facultad de Informática, Universidad Politécnica de Madrid, Campus de Montegancedo, 28660 Madrid, Spain; e-mail: m.pena@fi.upm.es*

Jorge Tejedor – 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: jtejedor@eui.upm.es*

Juan B. Castellanos – Dept. *Inteligencia Artificial, Facultad de Informática, Universidad Politécnica de Madrid, Campus de Montegancedo, 28660 Madrid, Spain; e-mail: jcastellanos@fi.upm.es*

Ginés Bravo – 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: gines@eui.upm.es*

A BOUNDED ALGORITHM BASED ON APPLICABILITY DOMAINS FOR THE APPLICATION OF ACTIVE RULES IN TRANSITION P-SYSTEMS

F. Javier Gil, Jorge A. Tejedor, Luis Fernández

Abstract: Transition P systems are a computational model based on the basic features of biological membranes and in the observation of biochemical processes. In this model, a membrane contains object multisets, which evolve following a determine set of evolution rules. The system changes from an initial configuration to another one performing a computation by applying these rules in a non-deterministic maximally parallel way.

Previous works about the rule application algorithms follow a common pattern: to choose a rule applicable on the objects multiset, and to apply that rule a number of times. This pattern is repeated until there is no applicable rule. In this paper, we present an algorithm that proposes a new approach that consists of the following steps: (i) To analyze the antecedent rules to establish the applicability domains; (ii) to select some applicable rules depending on the objects multiset available inside the membrane. Unlike previous algorithms, this new approach always selects an applicable rule in each iteration, avoiding iterations with non-applicable rules. In addition, this is a bounded algorithm, which is a necessary condition that facilitates decision-making in membrane architecture design.

Keywords: Natural Computing, Membrane computing, Transition P System, Rules Application Algorithms.

ACM Classification Keywords: D. Software. D.1 Programming Techniques

Introduction

Membrane computing is a branch of natural computing witch tries to abstract computing models from the structure and the functioning of biological cells, particular of the cellular membranes. The main objective of these investigations consists of developing new computational tools for solving complex, usually conventionally-hard problems. Being more concrete, Transition P systems are introduced by Gheorghe Păun derived from basic features of biological membranes and the observation of biochemical processes [Păun, 1998]. This computing model has become, during last years, an influential framework for developing new ideas and investigations in theoretical computation.

An essential ingredient of a P system is its membrane structure. Transition P systems are hierarchical, as the region defined by a membrane may contain other membranes. The basic components of the Transition P systems are the *membranes* that contain chemical elements (*multisets of objects*, usually represented by symbol strings) which are subject to chemical reactions (*evolution rules*) to produce other elements (another multiset). Multisets generated by evolution rules can be moved towards adjacent membranes (parent and children). This multiset transfer feeds back the system so that new multisets of symbols are consumed by further chemical reactions in the membranes.

The nondeterministic maximally parallel application of rules throughout the system is a transition between system states, and a sequence of transitions is called a computation. Each transition or *evolution step* goes through two sequential phases: the application of the evolution rules and communication. First, the evolution rules are applied simultaneously to the object multiset in each membrane. This process is performed by all membranes at the same time. Then, also simultaneously, all the membranes communicate with their neighbors, transferring symbol multisets.

Most membrane systems are computationally universal: “P systems with simple ingredients (number of membranes, forms and sizes of rules, controls of using the rules) are Turing complete” [Păun, 2005]. This framework is extremely general, flexible, and versatile. Several classes of P systems with an enhanced parallelism are able to solve computationally hard problems (typically, NP complete problems) in a feasible time (polynomial or even linear) by making use of an exponential space.

In this paper we propose a new algorithm for the application of evolution rules oriented towards the implementation of Transition P systems. The proposed algorithm has a linear order complexity in the number of evolution rules and provides an optimization of execution time compared to preceding contributions. Due to these characteristics, the algorithm is very appropriate for the implementation of Transition P systems in sequential devices.

After this introduction, related works are presented, where the problem that is tried to solve is covered. Then are exposed the formal definitions related to the application of rules in Transition P systems. Later appears developed the bounded algorithm based on applicability domains, including finally the efficiency analysis, and the conclusions.

Related Work

In Transition P systems, each evolution step is obtained through two consecutive phases within each membrane: in the first stage the evolution rules are applied, and at the second is made the communication between membranes. This work is centered in the first phase, the application of active rules. It exists at this moment several sequential algorithms for rules application in P systems [Ciobanu, 2002], [Fernández, 2006a], [Gil, 2008] and [Tejedor, 2007], but the performance of these algorithms can be improved. In the last mentioned work is introduced an algorithm based on the elimination of active rules: this algorithm is very interesting because is the first algorithm whose execution time is only limited by the number of rules, not by the objects multiset cardinality.

Additionally, in [Tejedor, 2006] is proposed a software architecture for attacking the bottleneck communication in P systems denominated “partially parallel evolution with partially parallel communications model” where several membranes are located in each processor, proxies are used to communicate with membranes located in different processors and a policy of access control to the network communications is mandatory. This obtains a certain parallelism yet in the system and an acceptable operation in the communications. In addition, it establishes a set of equations that they allow to determine in the architecture the optimum number of processors needed, the required time to execute an evolution step, the number of membranes to be located in each processor and the conditions to determine when it is best to use the distributed solution or the sequential one. Additionally it concludes that if the maximum application time used by the slowest membrane in applying its rules improves N times, the number of membranes that would be executed in a processor would be multiplied by the square root of N , the number of required processors would be divided by the same factor, and the time required to perform an evolution step would improve approximately with the same factor.

Therefore, to design software architectures it is precise to know the necessary time to execute an evolution step. For that reason, algorithms for evolution rules application that they can be executed in a delimited time are required, independently of the object multiset cardinality inside the membranes. Nevertheless, this information cannot be obtained with most of the algorithms developed until now since its execution time depends on the cardinality of the objects multiset on which the evolution rules are applied.

They have been proposed also parallel solutions - [Fernández, 2006b] and [Gil, 2007] -, but they do not obtain the required performance. The first algorithm is not completely useful, since its run time is not time delimited, and both solutions present efficiency problems due to the competitiveness between the rules, the high number of collisions with the requests and delays due to the synchronization required between processes.

Finally, the algorithm proposed in [Gil, 2008] obtains a good performance, but these can be improved in certain situations (specifically, when the relationship between the number of rules and the number of objects is high). Analyzing the behavior of the latter shows that this algorithm performs several iterations in which objects are not consumed due to the competitiveness between the rules. The bounded algorithm based on applicability domains is designed to avoid these unnecessary iterations, thus obtaining better performance.

Formal definitions related to rules application in P systems

First, this section formally defines the required concepts of objects multisets, evolution rules, evolution rules multiset, and applicability benchmarks (maximal and minimal) of a rule on the objects multiset. Second, on the basis of these definitions are specified the requirements for the new algorithm for the application of the evolution rules.

Multiset of Objects

Definition 1: Multiset of objects. Let a finite and not empty set of objects be O and the set of natural numbers N , is defined as a multiset of object m as a mapping:

$$\begin{aligned} m : O &\rightarrow N \\ o &\rightarrow n \end{aligned}$$

Possible notations for a multiset of objects are:

$$\begin{aligned} m &= \{(o_1, n_1), (o_2, n_2), \dots, (o_m, n_m)\} \\ m &= o_1^{n_1} \cdot o_2^{n_2} \cdot \dots \cdot o_m^{n_m} \end{aligned}$$

Definition 2: Set of multisets of objects over a set of objects. Let a finite set of objects be O . The set of all the multisets that can be formed over set O is defined:

$$M(O) = \{m : O \rightarrow N \mid m \text{ is a Multiset over } O\}$$

Definition 3: Multiplicity of object in a multiset of objects. Let an object be $o \in O$ and a multiset of objects $m \in M(O)$. The multiplicity of an object is defined over a multiset of objects such as:

$$\begin{aligned} | \cdot |_o : O \times M(O) &\rightarrow N \\ (o, m) &\rightarrow |m|_o = n \mid (o, n) \in m \end{aligned}$$

Definition 4: Weight or Cardinal of a multiset of objects. Let a multiset of objects be $m \in M(O)$. The weight or cardinal of a multiset of objects is defined as:

$$\begin{aligned} | \cdot | : M(O) &\rightarrow N \\ m &\rightarrow |m| = \sum_{\forall o \in O} |m|_o \end{aligned}$$

Definition 5: Multiset support. Let a multiset of objects be $m \in M(O)$ and $P(O)$ the power set of O . The support for this multiset is defined as:

$$\begin{aligned} Supp : M(O) &\rightarrow P(O) \\ m &\rightarrow Supp(m) = \{o \in O \mid |m|_o > 0\} \end{aligned}$$

Definition 6: Empty multiset. This is the multiset represented by $\emptyset_{M(O)}$ and which satisfies:

$$\emptyset_{M(O)} \Leftrightarrow |m| = 0 \Leftrightarrow Supp(m) = \emptyset$$

Definition 7: Inclusion of multisets of objects. Let two multisets of objects be $m_1, m_2 \in M(O)$. The inclusion of multisets of objects is defined as:

$$m_1 \subset m_2 \Leftrightarrow |m_1|_o \leq |m_2|_o \quad \forall o \in O$$

Definition 8: Sum of multisets of objects. Let two multisets of objects be $m_1, m_2 \in M(O)$. The sum of multisets of objects is defined as:

$$\begin{aligned} + : M(O) \times M(O) &\rightarrow M(O) \\ (m_1, m_2) &\rightarrow \{(o, |m_1|_o + |m_2|_o) \quad \forall o \in O\} \end{aligned}$$

Definition 9: Subtraction of multisets of objects. Let two multisets of objects be $m_1, m_2 \in M(O)$, and $m_2 \subset m_1$. The subtraction of the multisets of objects is defined as:

$$\begin{aligned} - : M(O) \times M(O) &\rightarrow M(O) \\ (m_1, m_2) &\rightarrow \{(o, |m_1|_o - |m_2|_o) \quad \forall o \in O\} \end{aligned}$$

Definition 10: Intersection of multisets of objects. Let two multisets of objects be $m_1, m_2 \in M(O)$. The intersection of multisets of objects is defined as:

$$\begin{aligned} \cap : M(O) \times M(O) &\rightarrow M(O) \\ (m_1, m_2) &\rightarrow m_1 \cap m_2 = \{(o, \min(|m_1|_o, |m_2|_o)) \quad \forall o \in O\} \end{aligned}$$

Definition 11: Scalar product of multiset of objects by a natural number. Let a multiset be $m_2 \in M(O)$ and a natural number $n \in \mathbb{N}$. The scalar product is defined as:

$$\begin{aligned} \cdot : M(O) \times \mathbb{N} &\rightarrow M(O) \\ (m, n) &\rightarrow m \cdot n = \{(o, |m|_o \cdot n) \quad \forall o \in O\} \end{aligned}$$

Evolution Rules

Definition 12: Evolution rule over a set of objects with target in T and with no dissolution capacity. Let a set of objects be O , $a \in M(O)$ a multiset over O , $T = \{\text{here, out}\} \cup \{\text{in}_j / 1 \leq j \leq p\}$ a set of targets and $c \in M(O \times T)$ a multiset over $O \times T$. An evolution rule is defined like a tuple:

$$r = (a, c)$$

Definition 13: Set of evolution rules over a set of objects and targets in T . This set is defined as:

$$R(O, T) = \{r \mid r \text{ is a rule over } O \text{ and } T\}$$

Definition 14: Antecedent of Evolution Rule. Let an evolution rule be $r \in R(O, T)$. The antecedent of an evolution rule is defined over a set of objects as:

$$\begin{aligned} \text{input} : R(O, T) &\rightarrow M(O) \\ (a, c) &\rightarrow \text{input}(r) = a \mid r = (a, c) \in R(O, T) \end{aligned}$$

Definition 15: Evolution rule applicable over a multiset of objects. Let an evolution rule be $r \in R(O, T)$ and a multiset of objects $m \in M(O)$, it is said that an evolution rule is applicable over a objects multiset if and only if:

$$\Delta_r(m) \Leftrightarrow \text{input}(r) \subset m$$

Definition 16: Set of evolution rules applicable to a multiset of objects. Let a set of evolution rules be $R \in P(R(O, T))$ and a multiset of objects $m \in M(O)$. The set of evolution rules applicable to a multiset of objects is defined as:

$$\Delta^* : P(R(O, T)) \times M(O) \rightarrow P(R(O, T))$$

$$(R, m) \rightarrow \Delta_R^*(m) = \{r \in R \mid \Delta_R(m) = true\}$$

Property 1: *Maximal applicability benchmark of evolution rule over a multiset of objects.* Let an evolution rule be $r \in R(O, T)$ and a multiset of objects $m \in M(O)$. The maximal applicability benchmark of a rule in a multiset is defined as:

$$\Delta[\] : R(O, T) \times M(O) \rightarrow N$$

$$(r, m) \rightarrow \Delta_r[m] = \min \left\{ \frac{|m|_o}{|input(r)|_o} \mid \forall o \in Supp(m) \wedge |input(r)|_o \neq 0 \right\}$$

Property 2: *Minimal applicability benchmark of evolution rule over a multiset of objects and a set of evolution rules.* Let an evolution rule be $r \in R(O, T)$, a multiset of objects $m \in M(O)$ and a set of evolution rules $R \in P(R(O, T))$. The minimal applicability benchmark is defined as the function:

$$\Delta[\] : R(O, T) \times M(O) \times P(R(O, T)) \rightarrow N$$

$$(r, m, R) \rightarrow \Delta_r[m] = \Delta_r \left[m - \left(m \cap \sum_{\forall r_i \in R - \{r\}} input(r_i) \cdot \Delta_{r_i}[m] \right) \right]$$

Property 3: *An evolution rule $r \in R(O, T)$ is applicable to a multiset of objects $m \in M(O)$ if and only if the maximal applicability benchmark is greater or equal to 1.*

$$\Delta_r(m) \Leftrightarrow \Delta_r[m] \geq 1$$

Property 4: The maximal applicability benchmark of a rule $r \in R(O, T)$ over an object multiset $m \in M(O)$ is greater than or equal to the maximal applicability benchmark of the rule in a subset of the object multiset.

$$\Delta_r[m_1] \geq \Delta_r[m_2] \quad \forall m_1, m_2 \in M(O) \mid m_2 \subset m_1$$

Property 5: If the maximal applicability benchmark of a rule $r \in R(O, T)$ over a multiset of objects $m \in M(O)$ is 0, then the maximal applicability benchmark of the rule r over the sum of input(r) and m is equal to the maximal applicability benchmark of the input(r) and equal to 1.

$$\Delta_r[m] = 0 \Rightarrow \Delta_r[input(r) + m] = \Delta_r[input(r)] = 1$$

Multisets of Evolution Rules

Definition 17: *Multiset of evolution rules.* Let a finite and not empty set of evolution rules be $R(O, T)$ and the set of natural numbers N , a multiset of evolution rules is defined as the mapping:

$$M_{R(O, T)} : R(O, T) \rightarrow N$$

$$r \rightarrow n$$

All definitions related to multisets of objects can be extended to multisets of rules.

Definition 18: *Linearization of evolution multiset of rules.* Let a multiset of evolution rules be $m_R = r_1^{k_1} \cdot r_2^{k_2} \cdot \dots \cdot r_q^{k_q} \in M_{R(O, T)}$ linearization of m_R is defined as:

$$\sum_{i=1}^q r_i \cdot k_i \in R(O, T)$$

Requirements of Application of Evolution Rules over Multiset of objects

Application of evolution rules in each membrane of P Systems involves subtracting objects from the objects multiset by using rules antecedents. Rules used are chosen in a non-deterministic manner. The process ends when no rule is applicable. In short, rules application to a multiset of object in a membrane is a process of information transformation with input, output and conditions for making the transformation.

Given an object set $O = \{o_1, o_2, \dots, o_m\}$ where $m > 0$, the input to the transformation process is composed of a multiset $\omega \in M(O)$ and $R \in R(O, T)$, where:

$$\omega = o_1^{n_1} \cdot o_2^{n_2} \cdot \dots \cdot o_m^{n_m}$$

$$R = \{r_1, r_2, \dots, r_q\} \text{ being } q > 0$$

In fact, the transformation only needs rules antecedents because this is the part that acts on ω . Let these antecedents be:

$$input(r_i) = o_1^{n_i^1} \cdot o_2^{n_i^2} \cdot \dots \cdot o_m^{n_i^m} \quad \forall i = \{1, 2, \dots, q\}$$

The **output** of the transformation process will be a objects multiset of $\omega' \in M(O)$ together with the multiset of evolution rules applied $\omega_R \in M_{R(O, T)}$.

$$\omega' = o_1^{n_1'} \cdot o_2^{n_2'} \cdot \dots \cdot o_m^{n_m'}$$

$$\omega_R = r_1^{k_1} \cdot r_2^{k_2} \cdot \dots \cdot r_q^{k_q}$$

The conditions to perform the transformation are defined according to the following requirements:

Requirement 1: The transformation process is described through the following system of equations:

$$n_1 = n_1^1 \cdot k_1 + n_1^2 \cdot k_2 + \dots + n_1^q \cdot k_q + n_1'$$

$$n_2 = n_2^1 \cdot k_1 + n_2^2 \cdot k_2 + \dots + n_2^q \cdot k_q + n_2'$$

$$\dots$$

$$n_m = n_m^1 \cdot k_1 + n_m^2 \cdot k_2 + \dots + n_m^q \cdot k_q + n_m'$$

That is:

$$\sum_{j=1}^q n_i^j \cdot k_j + n_i' = n_i \quad \forall i = \{1, 2, \dots, m\}$$

or

$$\sum_{i=1}^q input(r_i) \cdot k_i + \omega' = \omega$$

The number of equations in the system is the cardinal of the set O . The number of unknowns in the system is the sum of the cardinals of the set O and the number of rules of R . Thus, the solutions are in this form:

$$(n_1', n_2', \dots, n_m', k_1, k_2, \dots, k_q) \in \mathbb{N}^{m+q}$$

Meeting the following restrictions:

$$0 \leq n_i' \leq n_i \quad \forall i = \{1, 2, \dots, m\}$$

Moreover, taking into account the maximal and minimal applicability benchmarks of each rule, the solution must satisfy the following system of inequalities:

$$\Delta_{r_j}[\omega] \leq k_j \leq \Delta_{r_j}[\omega] \quad \forall j = \{1, 2, \dots, q\}$$

Requirement 2: No rule of the set R can be applied over the multiset of objects ω' , that is:

$$\Delta_r(\omega') = false \quad \forall r \in R$$

Having established the above requirements, the system of equations may be incompatible (no rule can be applied), determinate compatible (there is a single multiset of rules as the solution to the problem) or indeterminate compatible (there are many solutions). In the last case, the rule application algorithm must provide a solution that is randomly selected from all possible solutions in order to guarantee the non-determinism inherent to P systems.

Bounded Algorithm based on Applicability Domains for the Application of Active Rules in Transition P Systems

This section describes the bounded algorithm based on applicability domains the application of active rules. The initial input is a set of active evolution rules for the corresponding membrane -the rules are applicable and useful- and the initial membrane multiset of objects. The final results are the complete multiset of applied evolution rules and the multiset of objects obtained after the application of rules.

The Rule Applicability Domains

The applicability domains determination consist in a preliminary study of the antecedents of all the evolution rules, determining at each moment the set of rules that can be applied depending on the object multisets. Thus, at any time since it is known the object multisets in the membrane, it is possible to know exactly which set of rules that can be applied. Moreover, the computing of the applicability domains may be made prior to the process of application of rules, and therefore before the P system begins to evolve.

In addition, since at this point of transition between states of the P system only objects are consumed, the set of rules are reduced with each iteration of the algorithm, eliminating those rules that after each iteration are no longer applicable.

The Rule Application Algorithm

Before the application of the evolution rules, the domains of applicability are calculated. The algorithm is made up of two phases:

At the first phase an applicable rule set is randomly selected, and then applied a random number of times between one and its maximal applicability benchmark. The selected rule set is not reapplied during this first phase.

In the second phase all the applicable rule sets are applied to its maximal applicability benchmark until there is no applicable rule. Consequently, when this phase ends there it is not left any rule applicable, and the algorithm finishes generating like result the multiset of rules applied and the final multiset of objects.

In order to facilitate the explanation of the algorithm, the set of initially active rules is represented like an ordered sequence R and an auxiliary object called *applicableRuleSet*. This object has been built upon the study of the domains of applicability, and contains all sets of rules applicable at a given moment. The algorithm pseudo code is as follows:

```

(01)  $\omega' \leftarrow \omega;$ 
(02)  $\omega_r \leftarrow \emptyset_{MR(U)};$ 
(03)  $applicableRuleSet.init(R[i], \omega);$ 
(04) do { // Phase 1
(05)    $r \leftarrow applicableRuleSet.getRandom();$ 
(06)    $K \leftarrow random(1, \Delta_{R[r]}[\omega']);$ 
(07)    $\omega_r \leftarrow \omega_r + \{R[r]^K\};$ 
(08)    $\omega' \leftarrow \omega' - K \times input(R[r]);$ 
(09)    $applicableRuleSet.remove(r);$ 
(10) } while (! $applicableRuleSet.empty()$ );
(11)
(12)  $applicableRuleSet.init(R[i], \omega');$  // Phase 2
(13) while (! $applicableRuleSet.empty()$ ) {
(14)    $i \leftarrow applicableRuleSet.getNext();$ 
(15)    $ma \leftarrow \Delta_{R[i]}[\omega'];$ 
(16)    $\omega_r \leftarrow \omega_r + \{R[i]^{ma}\};$ 
(17)    $\omega' \leftarrow \omega' - ma \times input(R[i]);$ 
(18)    $applicableRuleSet.remove(i);$ 
(19) }

```

As it has been previously indicated, the algorithm is made up of two phases. In first stage is offered the possibility to all the applicable set of rules to be applied between one and their maximum applicability benchmark. The rule sets can let be active in this stage due to two possible reasons: a) the rule set has been applied to its maximum applicability, or b) other precedents rule sets have consumed the necessary objects so that the rule set can be applied.

At the beginning of the second phase the applicable set of rules is recalculated. Then in a loop are selected and applied to its maximum applicability each applicable rule sets, until there is no applicable rule and the application algorithm finish their execution. Through the use of the domains of applicability, all iterations consume objects. As seen, the algorithm executes a finite and bounded number of operations.

In the next sections we are going to demonstrate the correctness of the exposed algorithm, as well as the efficiency analysis.

Algorithm Correctness

The presented algorithm is correct because:

Lemma 1: *The algorithm is finite.*

Proof: The algorithm is composed of basic operations and two loops. In these loops always reduces the number of objects in the membrane, and therefore the algorithm terminates when there is no applicable rule.

Lemma 2: *No evolution rule is applicable to ω' .*

Proof: After the execution of the second phase of the algorithm, the maximal applicability of all the rules is zero. Therefore, at the end of the algorithm execution, it is not left any applicable rule to ω' .

Lemma 3: *Any result generated is a possible solution.*

Proof: The multiset of rules applied ω_R is obtained by the multiple applications of the active rules in both phases. In addition, the second phase ends when there are no applicable rules over ω' (requirement 2), and the result generated is a possible solution.

Lemma 4: *Any solution possible is generated by the algorithm*

Proof: Phase 1 of the algorithm - from line (4) to (10) - guarantees that any possible solution can be generated. It is enough whereupon the appropriate number is generated in line 6, when the number of applications of a rule is determined. In the second phase it would be only needed to apply the last rule the appropriate number of times.

Lemma 5: *The algorithm is not determinist*

Proof: The algorithm is nondeterministic since from the same initial input is able to generate any possible solution (multiset of rules that make the algorithm finishes). In phase 1, both the selection of the set of rules, as the choice of the number of times they apply, are made randomly.

Efficiency Analysis

Examining the algorithm it is possible to observe that in the two phases, the heaviest operations are those that calculate the maximal applicability benchmark (sentences 6 and 15), the scalar product of the *input* of a rule by a whole number and the difference of two multisets (sentences 08 and 17). These operations are made in both phases in the worse case. All these operations are linearly dependant on the cardinal of the multiset support ω .

$$\#operations_per_iteration \approx 3 \cdot Supp(\omega)$$

Moreover, examining the first phase, the worst case of the algorithm occurs when - in sentence 05 - is selected a set that contains only one rule. As the rule is eliminated from the set of applicable rules, the first loop will execute at most R times. The same is true in the second phase, bringing that the number of iterations of both loops will be:

$$\#iterations = 2 \cdot |R|$$

Therefore, the number of operations executed at worst case by the algorithm is:

$$\#operations = (2 \cdot |R|) \times 3 \cdot Supp(\omega)$$

So the execution time of the algorithm at worst is bounded and linear dependant of the number of rules. Moreover, in practice, it is expected that the number of iterations of both loops is smaller, and therefore the efficiency will be higher.

Conclusions

This paper introduces a new algorithm for active rules application to a multiset of objects based on applicability domains in transition P systems. This algorithm attains a certain degree of parallelism, as a set of rules can be applied a great number of times in a single step. The number of operations executed by the algorithm is bounded, because it only depends on the number of rules of the membrane. The number of rules of the membrane is well known static information studying the P system, thus allowing the determination prior to the application of the rules. This information is essential to calculate the number of membranes that have to be located in each

processor in distributed implementation architectures of P systems to achieve optimal times with minimal resources.

We think that the presented algorithm can represent an important contribution in particular for the problem of the application of rules in membranes, because it presents high productivity and it allows estimate the necessary time to execute an evolution step. Additionally, this last one allows making important decisions related to the implementation of P systems, like the related ones to the software architecture.

Bibliography

- [Ciobanu, 2002] G. Ciobanu, D. Paraschiv, "Membrane Software. A P System Simulator". Pre-Proceedings of Workshop on Membrane Computing, Curtea de Arges, Romania, August 2001, Technical Report 17/01 of Research Group on Mathematical Linguistics, Rovira i Virgili University, Tarragona, Spain, 2001, 45-50 and *Fundamenta Informaticae*, vol 49, 1-3, 61-66, 2002.
- [Ciobanu, 2006] G. Ciobanu, M. Pérez-Jiménez, Gh. Păun, "Applications of Membrane Computing". Natural Computing Series, Springer Verlag, October 2006.
- [Fernández, 2006a] Fernández, L. Arroyo, F. Castellanos, J. et al (2006) "New Algorithms for Application of Evolution Rules based on Applicability Benchmarks". BIOCAMP 06, Las Vegas (USA)
- [Fernández, 2006b] L. Fernández, F. Arroyo, J. Tejedor, J. Castellanos. "Massively Parallel Algorithm for Evolution Rules Application in Transition P System". Seventh Workshop on Membrane Computing, WMC7, Leiden (The Netherlands). July, 2006
- [Gil, 2007] Gil, F. J. Fernández, L. Arroyo, F. et al "Delimited Massively Parallel Algorithm based on Rules Elimination for Application of Active Rules in Transition P Systems" i.TECH-2007. Varna (Bulgaria).
- [Gil, 2008] Gil, F. J. Tejedor, J. A. Fernández, "Fast Linear Algorithm for Active Rules Application in Transition P Systems" i.TECH-2008. Varna (Bulgaria).
- [Păun, 1998] G. Păun. "Computing with Membranes". In: *Journal of Computer and System Sciences*, 61(2000), and Turku Center of Computer Science-TUCS Report n° 208, 1998.
- [Păun, 2005] G. Păun. "Membrane computing. Basic ideas, results, applications". In: Pre-Proceedings of First International Workshop on Theory and Application of P Systems, Timisoara (Romania), pp. 1-8, September, 2005.
- [Tejedor, 2006] J. Tejedor, L. Fernández, F. Arroyo, G. Bravo. "An Architecture for Attacking the Bottleneck Communications in P systems". In: *Artificial Life and Robotics (AROB 07)*. Beppu (Japan), January 2007.
- [Tejedor, 2007] J. Tejedor, L. Fernández, F. Arroyo, A. Gutiérrez. "Algorithm of Active Rules Elimination for Evolution Rules Application". In 8th WSEAS Int. Conf. on Automation and Information, Vancouver (Canada), June 2007.

Authors' Information

F. Javier Gil Rubio – Dpto. de Organización y Estructura de la Información, E.U. de Informática. Natural Computing Group, Universidad Politécnica de Madrid, Spain; e-mail: jgil@eui.upm.es

Jorge A. Tejedor Cerbel – Dpto. de Organización y Estructura de la Información, E.U. de Informática. Natural Computing Group, Universidad Politécnica de Madrid, Spain; e-mail: jtejedor@eui.upm.es

Luis Fernández Muñoz – Dpto. de Lenguajes, Proyectos y Sistemas Informáticos, E.U. de Informática. Natural Computing Group, Universidad Politécnica de Madrid, Spain; e-mail: setillo@eui.upm.es

COLLISION DETECTION AND TREATMENT USING 2D RECONFIGURABLE HARDWARE

Alejandro Figueroa, Gustavo Méndez, Francisco J. Cisneros, Adriana Toni

Abstract: *The detection and treatment of collisions has been the subject of study for many years, periodically appear new techniques and algorithms to solve. It This article presents a hardware alternative to the detection of collisions between two or more surfaces in real time taking advantage of the parallelism offered by FPGAs, applying on a Spanish billiard of three balls simulator. FPGAs (Field-Programming Gate Array) provide a highly flexible environment for the programmer, since its cells can be reprogrammed and executed with great ease, which allows them to be used for an enormous range of applications.*

Keywords: *Collision detect, 2D Graphics*

ACM Classification Keywords: *1.6. Simulation and Modelling, 1.3 Computer Graphics*

Introduction

The simplest definition of a collision detection between two objects (be they surfaces or volumes) is basically whether at a given moment, there is an intersection between the two or not. With this idea, we developed the overall study of collisions in an environment and its treatment [1].

For example, in three-dimensional model representation, the objects define their body - called mesh - through a large number of vertices that form polygons. These polygons are in contact with each other, forming the mesh. When there is a collision with an obstacle within the environment of the object, an application must determine where and how the intersection has occurred that caused the collision and take action.

Detection system and treatment of collisions is necessary for a large number of applications including entertainment systems, robotics, physical applications, biomedicine, military applications, etc. All these fields of research require such applications that solve their specific problems. Today, nearly all of these applications use collision detection software techniques developed in different languages, mostly software that needs one or more CPU to run [2] [3].

The FPGA is a semiconductor device, which contains a large number of logical blocks, called CLB, including a number of look-up tables (LUT) - where combinational logic is stored - "full adders" as basic hardware (adders), bistable, etc, whose interconnection and functionality is configurable by the programmer (which adds great flexibility to the developer). Each FPGA has more than one million CLBs, suggesting the enormous operational capability of these devices. In addition, FPGAs have special blocks for memory (both on chip and external) that act as a support for all the logic of the plate and, of course, can be freely programmed in accordance with the requirements of a particular system. [5] [6]

You can load a FPGA almost any program, provided that it does not exceed the capacity of the CLB or reports.

The FPGAs are reprogrammable, such as processors, which can be used to implement designs. These designs, once loaded onto the board, may be modified, and loaded again, as often as desired [4] [4].

The work presented here is a system of detection and treatment of two-dimensional collisions in a linear environment on a FPGA. To this end, this study is simulated by an application of the Spanish pool game (three balls).

The simulation uses the implicit parallelism of FPGAs to perform all calculations concurrently. Therefore, a formal

language is needed to program, based on this inherent concurrency. VHDL has been chosen, one of the hardware description languages used in programming hardware on FPGAs.

VHDL conducts operations through assignment of signals in parallel through multithreading techniques that allow the programmer to modularize the tasks in different processes, leaving the machine that run concurrently. As in any software language, VHDL is very important for modularization, which is done by defining the entities that represent the behavior of the program.

Application and Features

The application is presented to the user via one VGA monitor, keyboard, and a speaker. With this, the user has at hand the complete management of the application.

The user controls through the keyboard's numeric pad for impact direction he wants to give the ball and simulate the impact of the cleat into the ball which will bounce with an initial velocity in the direction chosen.

The application is divided into the following modules:

- Sound Controller
- VGA Driver
- Keyboard Controller
- Main Application Module

(The modularization is detailed in the *Implementation* section)

The program has been done in several stages, in which features were added gradually. At first, he drew a polygon and speed is printed, regardless of collisions with the boundaries of the board. Thus, at first was a simple square movement, which eventually became a sphere. After obtaining the motion of the object, it was easier to detect collisions, within the limits of the board. The treatment of these collisions will be detailed below (paragraph *implementation*). To the first white ball, direction was added from the keyboard, getting, finally, a more appropriate interface with which to control the start of the simulation. We subsequently added the friction between the pool table and the ball. It is also explained in the *logic of the System* section.

Once finished, the other two balls had to be added to the environment and make them collide. Indeed, this was the most difficult point with work, because internally, this resulted in several state machines working concurrently and synchronizing with each other. The detailed explanation about the WSF is detailed in the *logic of the System* section.

The last feature added to the application was a splash screen starting with the presentation of the game. It was through RAM at the beginning of the FPGA, reading the content stored in another memory and displays on screen (*Implementation* section in detail).

We have encountered a number of problems when implementing this program. In the experiments, we observed that the machine behaved satisfactorily when simulating shown and one or two balls simultaneously, but, however, sometimes failed to add the third ball. After many hours of experimentation, it was possible to isolate the problem and place it in concerning state machines. It concluded that the idea was flawed, but the parallelism of the main board, sometimes of a FSM signals propagated to another, with different clock signals (discussed further in the *logic of the system* section) were lost, resulting in erroneous behavior at times. It was decided to create a more consistent code that better optimize the capabilities of the FPGA without, at the same time, lost information caused by the lag between watches.

Logic System

The programming allows the user FPGAs have a logic high capacity for development, because of the huge amount of CLB that stores.

When you start to develop the work, you must take into account a number of conditions that are imposed by working on FPGAs. One concerns the graphical representation on VGA. Because the FPGA print in the monitor 2 pixel for each pixel in the Y axis (while in the X axis only paints one), we adapted the logical design of our system and values doubled when working on the Y axis. For this reason, the balls are not drawn as spheres, but as ellipses, whose Y component is twice the radius of the component X. The final representation on the screen is that of a regular field.

To try it this peculiarity, all signals were doubled to represent both the value on the axis X and the Y axis, because for the detection of collisions needed a complete record of the current positions of each object, and this requires real-time updating of the values.

Movement of the balls

The movement of the balls is done pixel by pixel (two pixels in the Y coordinate for each pixel in X). The speed of each ball is simulated using a parameterized clock, and modifying in real time the frequency of the clock, getting simulate the friction of the balls with the board, or wear after a collision. According to the value vector having the *direction* of the ball, the process updates the signals that indicate the position of the center of the ball.

Each ball has its own process of movement, which has a particular clock, for every ball should move independently. Each watch is parameterized by a vector, so that you can change the frequency at run time (note that increasing the value of that vector, decreases the clock frequency at which the process works, as seen in the snippet below).

```

PROCESS OF CLOCK SIGNAL FROM THE WHITE BALL
(SIMILAR TO THE PROCESS WILL WATCH THE OTHER BALL)
process (clock,reset)
begin
    if reset = '1' then
        WhiteClockCounter <="000000000000000000000000";
        WhiteClock<='0';
    elsif (clock'event and clock = '1') then
        WhiteClockCounter <= WhiteClockCounter + '1';
        if WhiteClockCounter>=WhiteParam then
            WhiteClock <=not WhiteClock;
            WhiteClockCounter <="000000000000000000000000";
        end if;
    end if;
end process

```

where *WhiteClock* is the clock signal the cue ball; *clock* is the clock of the FPGA,

WhiteClockCounter is the counter that counts the number of *clock* cycles that take *action*.

WhiteParam is the parameter that modifies the frequency. Compared with *WhiteClockCounter* and, if it has reached that number of cycles, it makes a transition from *WhiteClock*.

When the ball is at rest (state *S_ini*) the clock parameter is assigned a constant value. This value is small enough for the ball to go out with some initial velocity, but not as fast as for the player was not able to perceive.

In the process, in each clock cycle is increased the value of the vector, getting a longer clock cycle, giving the impression that the ball is stopping. To stop the ball, set a higher threshold for the vector, so if the state machine detects that it has exceeded that threshold, will transition to idle and the ball stops.

In the case of yellow and red balls, do not give an initial value vector that parameterizes the clock, because the ball will stand to suffer the impact of another ball. Upon the impact, the vector of the incised ball gets the value of the vector of ball that hit her. This generates a feeling that the ball is thrown to hit the same speed as the ball that coincided with it.

Finite state machines (FSM)

The Spanish pool consists of three balls. A white one the user hits, and two balls (red and yellow), which can only be hit by another ball. The behavior of the balls is implemented by three state machines. Each specifies the behavior of a ball. The state machines communicate with each other to report the state of the table, that is, the possible collisions between balls, walls, etc.

Different FSM differ little from each other, the behavior of the yellow ball is the same of the red and white, except that in the FSM of the cue ball must be controlled starting direction the user gives the ball. The three state machines work with the clock signal produced by the FPGA itself.

The implementation of each of the state machine in VHDL is by two concurrent processes. The first one is responsible for carrying out synchronously, according to the clock of the FPGA, the state transitions. The second process performs the operations defined for each state. It is responsible for monitoring the collisions, charge signals in the registers, etc.

PROCESS OF CHANGE OF STATUS:

```
process (clock,reset)
begin
    if (reset='1') then
        state <= Sini;
    elsif(clock'event and clock='1') then
        state<=nextState;
    end if;
end process;
```

PROCESSES. PSEUDOCODE

```
process (state,reset)
begin
if (reset = '1') then if (reset = '1 ') then
    // Initialize SIGNALS
elseif (clock'event and clock = '1 ') then
    case state is
        when SIni=>
            // INACTIVE STATUS
            // TRANSITION TO S0 IF AND ONLY IF THE USER
            Hit the ball
        when S0=>
            // IF VECTOR PARAMETER> = THRESHOLD, TRANSITION
            Sini
            // IF NO, TRANSITION TO S1
        when S1=>
            // Intermediate state. always, Transition to S2;
        when S2 =>
            // Collision detection
            // UPDATE VALUES OF DIRECTION OF THE BALL
            // If there is another collision with ball, flag and active
            change your address
            // TRANSITION TO S0
        when others =>
            // TRANSITION TO S0.
    end case;
end if;
end process;
```

The basic and most important property of the FPGAs is the natural parallelism offered to the developer. In the system, such parallelism is used to make balls move simultaneously.

Initially, the three state machines are *asleep*, that is, in its initial state (S_ini). When the users determines a direction using the keypad and confirm the release, this selection is loaded in the register which controls the direction of the cue ball and the state machine is *activated*, making a transition to the first state. The ball begins to move, as specified in paragraph *movement of the ball*. Collisions with both walls and other balls are treated in the state $S2$. If the ball detects a collision with the red or yellow ball, the signal corresponding to active mode *flag* is activated and received by the state machine of the impacted ball, charging in its address register orientation calculated from the direction of the ball incident (as explained in the *Implementation* section, subparagraph *collisions*) making a transition to its first state, ie the ball *wakes up* and starts to move.

The outline of the state machine is as follows:

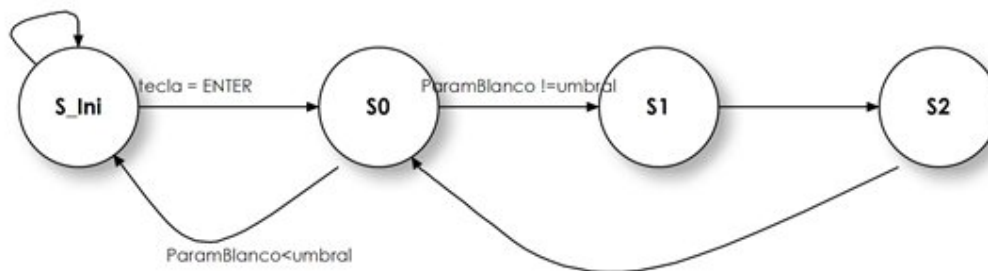


Fig 1. FSM white ball

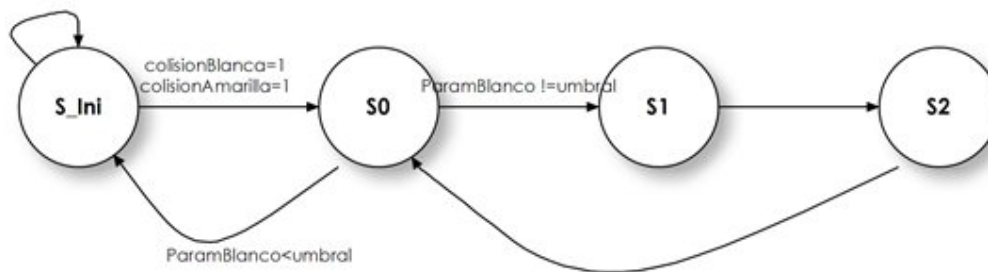


Fig 2. FSM yellow and white ball

It is seen that the three state machines are similar, have the same states and the same transitions, being the only difference between them how to be the first transition (from S_ini to $S0$).

The operation of state machines is as follows:

- **S0:** This state is the first state of movement of the ball, which moves during a cycle. This statement makes a check of the clock signal particular parameterization of the ball, to see if the threshold that will force the ball to stop has been reached. If, indeed, being reached, makes a transition to inactivity, S_ini . If the check is counterfeit, it makes the transition to $S1$.
- **S1:** This state has the unique function of intermediate state. In this state, the ball does not move. It makes a transition to $S2$ in any case. This state is necessary, because thanks to the modification of the *Enable* movement, the FPGA plays $S0$, $S1$ and $S2$ as independent states and, overall, as a FSM.
- **S2:** The state that performs the audit of collisions. By order, checks if the ball hits the walls that delimit the table or whether, on the contrary, collide with another ball. For this test uses a sequence of *If - then-elsif-else* to check all cases. If a collision is detected, updates the new direction of the incident ball and

hit the ball activates the corresponding flag to notify the FSM of the ball and, finally, increases the vector that parameterizes the ball, to simulate the impact wear. After making all these calculations, it makes a transition back to S0.

The only state that differs from state machines is the initial state, which in the case of the white ball makes transitions to itself every cycle until the user hits the ball, when the FSM goes to S0. In the case of the other balls, this transition is only made when the impact *flags* are activated.

Implementation

Graphic representation

All the elements that represent the program are housed in the same process, which works with a special clock signal whose frequency is suitable for work on a VGA terminal.

The idea is to control, by means of two counters, the position of every pixel of your monitor. These counters (*hcnt* for the pixel count of the X axis, and *vcnt* for Y-axis) increase in each VGA clock cycle. With these counters, you can paint every pixel. Sequences are chained *if-then-else* to check the current values of these counters and assign a color to each region between them. Thus, the board is defined from the lines that delimit. The following snippet of code has been greatly simplified:

```
if ((hcnt<0 or hcnt>268) or (vcnt<18 or vcnt>301)) then
    rgb<="000000000"; --black
elsif ((hcnt>=0 and hcnt<269) and ((vcnt>17 and vcnt<33) or (vcnt>287 and vcnt<302))) then
    rgb<="011010000"; --brown
elsif ((vcnt>=24 and vcnt<295) and ((hcnt>=0 and hcnt<8) or (hcnt>261 and hcnt<269))) then
    rgb<="011010000"; --brown
else
    rgb<="000100000"; --green
end if;
```

With few lines of code, you can specify the color that should have the entire VGA monitor. The RGB signal is responsible for assigning a pixel color.

Lines are used 1, 2, 3, 4 of Figure 3 to define the edges of the table and use them to paint the table and collision handling.

For the initial screen using an image memory with a representative image, whose relationship pixel - memory location is performed similarly to the board. It also develops an VHDL entity that defines the behavior of a RAM. When the board starts, loads in memory an initial vector with the image you want to load and then the allocation is made of the corresponding pixel.

To choose the direction you use the number keys and the enter key to confirm and hit the ball in that direction. The representation of the direction is drawing the point of impact of bat on the ball (point 5 in the Fig 3).

The balls are created as a vertical ellipse twice the horizontal radius to compensate for both radio and create the effect of circumference.

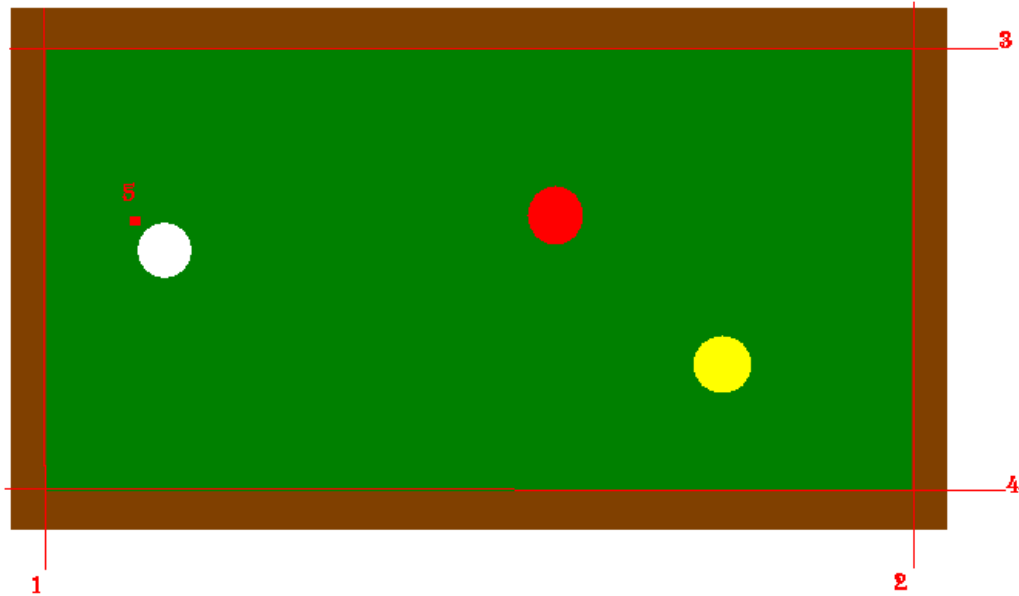


Fig 3. Board of game

Simulation

This section details how objects are defined internally affecting the simulation.

The main problem faced by a programmer to work with FPGAs is the arduous task of getting the FPGA code correctly interpreted the way you want. When, moreover, it is a system that modifies their values dynamically, we face the need to constantly store, in a series of signals, the updated values of the attributes and object in the system and define them for each conditional branch so that the system has a number of concrete data and no problems of interpretation.

The three balls are defined by two signals each, which store the position of its center in each component (X and Y) so that we know where the ball is located within the board. These two signals are essential in the calculation of collisions. Each clock cycle, during the simulation, these signals are updated depending on the direction of the ball. It is important to re-emphasize that the movement of the balls is done pixel by pixel, two by two in the case of Y.

The direction of each ball is also implemented with a signal, a vector of three bits, to specify the eight possible directions you can take a ball in the simulation. This vector is also of great importance for the calculation of collisions, it determines the output direction of the ball incident and incised.

Collisions

This section details how to detect, treat and resolve the application of the balls collisions with both other balls as the walls.

The motion simulation is pixel by pixel (explained in paragraph *movement of the balls* in the *Logic of the system* section). With this premise, we eliminate the problem of interpenetration.

The algorithmic scheme is based on the idea of covering each ball with a *Bounding-Box* that delimits. Thus, the balls are treated as if they were virtual square, significantly simplifying the logic needed to resolve these collisions.

As explained above, the application stores all the information of the objects in signals. It has a signal that indicates the radius of the balls. Is to create a Bounding-Box dimensions $2 * Radio$.

Collisions with fixed objects. Wall

These collisions are the simplest, as the walls no change after the collision.

As already indicated, is a comprehensive management of the values of the positions, and the limits of the board. Note that these limits are constantly required to paint the board.

Because the determination is carried out in the state machine, you can make that during that cycle the ball remains stopped until the collision is resolved. The user does not appreciate this stop. During that cycle (coinciding with the state *S2*) checks if the ball (surrounded by *Bounding-Box*) hits on some walls. If a collision is detected by means of a switch, is seen with which direction the ball hits and based on it, it checks if the ball is approaching or moving away and resolves its address output. Likewise, updates the parameter friction to simulate the wear of the collision and make the ball to slow.

Collisions with moving objects. Balls

These collisions are equally easy to detect, as in the case of the walls, all the information is stored and updated, but they are difficult to simulate, because they must synchronize two or more state machines in the number of balls involved in the collision. The way to deal with these collisions is very similar to the case of the walls. In the state of treatment of collisions of the state machine (*S2*) we establish the conditional branches to verify all possible cases. If it detects that the position of the moving ball collides with another ball (which can stand or move, too), the program does the following:

1. Calculates the new direction of the ball.
2. Activates the relevant *flag* to *raise* adequate state machine.
3. Calculates the address you should start the ball impacted.
4. Go to the transition to *S0*.

All these points are made within a switch that is calculated based on the direction of the ball incident.

Should be noted that VHDL does not allow writing a signal from different processes, as the FPGA, when executes these processes concurrently, will find a conflict of multiple signals at the entrance to a record, etc. This force us to create different signals within each process, which has the function of modifying the signal that cannot be written from there. You can, however, read a signal from several sites, which translates into the same output device, connected to different sites. Therefore, to modify a signal from *outside*, create as many processes as auxiliary signals exist having to modify that signal. The inherent problems involved this is in excess of logic that may make some signals are lost, resulting in erroneous behavior.

Thus, the state machines of the balls that have been impacted have two possible answers:

- 1) If they were in the inactive state (*S_ini*) then they must read the *flag* on, which indicates the type of collisions they have suffered and the auxiliary signal to be loaded into its vector direction, load the new address and moving the state forward.
- 2) If they were in motion, they will detect the collision in the state *S1*, modify the direction and speed (significant increase of the parameter for this purpose), check that has not reached the threshold speed limit (in which case they should go to inactive state and stop) and continue execution.

The ball incidents also modify the parameter of friction to slow down and continuing the simulation.

Sound Module

The game makes a sound each time one of the balls collide with another ball or any of the bands of the table, and when the shot is made with the cue.

To make sounds with the FPGA, we use the audio CODEC Serial AK4520A which is included in the FPGA. The Codec has the following inputs:

- *MCLK* or main clock.
- *LRCK* or channel selector.
- *SCLK* or clock of serial data transmission.
- *STDI* or serial input data.

Each of these signals is handled by its own counter to achieve the often necessary for the issuance of a note LA. Every time there is a connection, make a noise with a duration of 10^6 cycles of the FPGA clock.

Conclusions

It has been developed a system of detection and treatment of dynamic collisions, capable of resolving collisions of mobile and static surfaces, which are managed through independent state machines, through a mechanism of warning *flags*, and applied to a Spanish pool game. It was also seen that the implementation of this mechanism, beyond the cost of maintaining the state machines, does not require much logic area of the FPGA, which can be adapted to more complex circuits and even to multiple FPGA systems.

Bibliography

- [1] G. Baciú and S. K. Wong. Image-based techniques in a hybrid collision detector. In IEEE Trans. On Visualization and Computer Graphics, 2002.
- [2] A. Gress and G. Zachmann. Object-space interference detection on programmable graphics hardware. In In SIAM Conf. On Geometric Design and Computing. 2003.
- [3] D. Knott and D.K.Pai. Cinder: Collision and interference detection in real-time using graphics hardware. In Proc. Of Graphics Interface, 2003.
- [4] Modelsim. <http://www.model.com>.
- [5] Xilinx. <http://www.xilinx.com>.
- [6] NAN Xi, GONG Longqing, TIAN Wei, LI Xiao. Design and Implementation of Reconfigurable System Based on FPGA. Modern Electronics Technique, 2009

Authors' Information

Alejandro Figueroa Meana - Facultad de Informática Universidad Complutense de Madrid,
e-mail: afmeana@gmail.com

Gustavo Méndez Muñoz - Facultad de Informática Universidad Complutense de Madrid.
e-mail: gustavillo85@gmail.com

Francisco J. Cisneros de los Ríos – Natural Computing Group. Universidad Politécnica de Madrid, Boadilla del Monte, 28660 Madrid, Spain: e-mail: kikocisneros@gmail.com

Adriana Toni – Facultad de Informática Universidad Politécnica de Madrid. e-mail: atoni@fi.upm.es

BACTERIAL TECHNOLOGY TO BUILD COMPUTERS: A SURVEY

Paula Cordero, Sandra Gómez, Rafael Gonzalo

Abstract: *Synthetic biology is an emerging discipline that combines knowledge from various disciplines including molecular biology, engineering, and mathematics to design biological devices whose goal is to extend or modify the behavior of organisms and engineer them to achieve desired functions for broad applications. This paper provides a brief survey about the methods for altering the behavior of individual elements and the construction of complex networks in single-cell. In this review, we shown and analyze the recent advances in synthetic biology towards engineering complex living single-cell by designing genetic circuits to perform new tasks in bacteria behavior.*

Keywords: *Synthetic biology, bacterial engineering, bacterial computation, natural computing*

ACM Classification Keywords:

Introduction

It has been observed for several years the emergence of synthetic biology, a new area of biological research that aims to engineer biological devices with desired functions for broad applications. A large set of advances in this field highlight the potential of this discipline to impact diverse areas, including environment, bioremediation and computation. Synthetic biology aims to create novel behaviors through the engineering of genetic elements and the integration of basic elements into circuits that implement more complex functions that do not occur in nature [Weiss, 2005]. The design of synthetic biology applications includes living organisms, hardware and software components representing the instruments and the application to acquire and process signals generated by the biological component of the system. Early efforts aimed at altering the behavior of individual elements have now evolved to focus on the construction of complex networks in single-cell and multicellular systems. Synthetic biology needs to develop more sophisticated design strategies because offers valuable quantitative insight into naturally occurring information processing activities.

From the information codification point of view, synthetic biology has tackled a great change of concept in the field of computing. Molecular computing consists of representing the problem's information with organic molecules and making them react within a test tube in order to solve a problem. This concept is based on the work made by Leonard M. Adleman, where the first DNA computation based on Operations to hard combinatorial problem solved using desoxirribonucleic acid molecules [Adleman, 1994]. In this point, all the computations carried out in vitro based the codification of the problem solutions on representing them into organic molecules and as a result it was required the presence of a scientist to execute the experiments and give meaning to the solution obtained in the context of the problem domain. Here, the great difference provided by synthetic biology appears. In particular, in the area of bacteria computing the codification of the information is implemented at genetic level, this way the organisms are equipped with autonomous computing abilities due to the information have biological functionality in this case. At higher level, these computations have a biological meaning for organisms that carry out it and as a result applications in complex systems with specific purposes can be applied; this is the main aim of synthetic biology.

This paper is structured in the following way: firstly, a background about engineering bacteria is shown in order to present a framework of the concepts and the importance of this field. Secondly, a set of prominent developments are introduced to show the fast growth of the synthetic biology and the revolutionary change of concept that is

producing in the field of computing. Finally a set of perspectives and conclusions about the potential applications and synthetic biology areas for improving are shown.

Background: Engineering Bacteria

Cells are an important element of nature that serves as a model to abstract efficient and complex functions. Cells can be programmed by identifying three functional layers: an input layer, an information processing layer, and an output layer [Kobayashi, 2004]. In the input layer, they are inherently able to detect small concentrations of chemicals or combinations of chemicals in their environment to process in the processing layer and respond usually with an amplified signal across the output layer. This abstraction allows viewing this functionality as a device that can be manipulated and programmed and therefore performed as a computational device. Therefore, one of the most important aims of the synthetic biology is to design bacteria in order to achieve miniature computers, which could be programmed by designing genetic circuits into bacteria with specific functionalities.

This biological devices or miniature computers represents genetic circuits composed of biochemical reactions and genes of the same manner those electronic circuits work with boolean logic signals and gates. These devices can be engineered with sets of one or more biochemical reactions and are based on inducible promoters which allow us to switch on/off the expression of a certain gene. Every gene is a long double DNA strand which codifies particular information. Gene expression is the main principle of synthetic biology. This is the process by which information from a gene is used in the synthesis of a functional gene product. These products are often proteins, but in non-protein coding genes the product is a functional RNA. Several steps in the gene expression process may be modulated, including the transcription, and post-translational modification of a protein. The initiation of gene transcription is controlled by two sequences located upstream of the gene, the promoters. RNA polymerase recognizes these promoters as a signal to start transcription. This initiating step is the main site at which the rate of gene transcription is controlled. The ability of RNA polymerase to recognize and bind promoters can be altered by accessory proteins which can be activators or repressors. By controlling this process it is possible to achieve biological devices with a broad spectrum of applications.

Synthetic biology can be inspired by chemistry or by engineering to synthesize biomolecules, to develop synthetic analogs and to apply them in the framework of biological systems; and to design of new, complex, bioinspired systems respectively [Pleiss, 2006]. In this field bacteria computing is developed and in order to define these biological devices is necessary a design strategy. As in the design of computer systems, the strategy consists of dividing the complex biological system in a set of parts or reusable modules with defined properties and functions. Each of these parts represents a biological device. Several devices can be assembled in even more complex devices until the final system is constructed. This architecture allows a high level of abstraction because it allows separation of design and construction.

Therefore, complex systems with predictable properties such as computers can be constructed from only a small number of different, standardized basic parts. In accordance, the programmed bacteria could be able to communicate with others, and as a result to operate in a coordinate manner. For instance, *Escherichia coli* bacteria are designed with quorum-sensing proteins that allow emitting or receiving signals when the concentration level exceeds a certain limit. These useful techniques are applied in the design of communication systems in order to achieve more complex architectures.

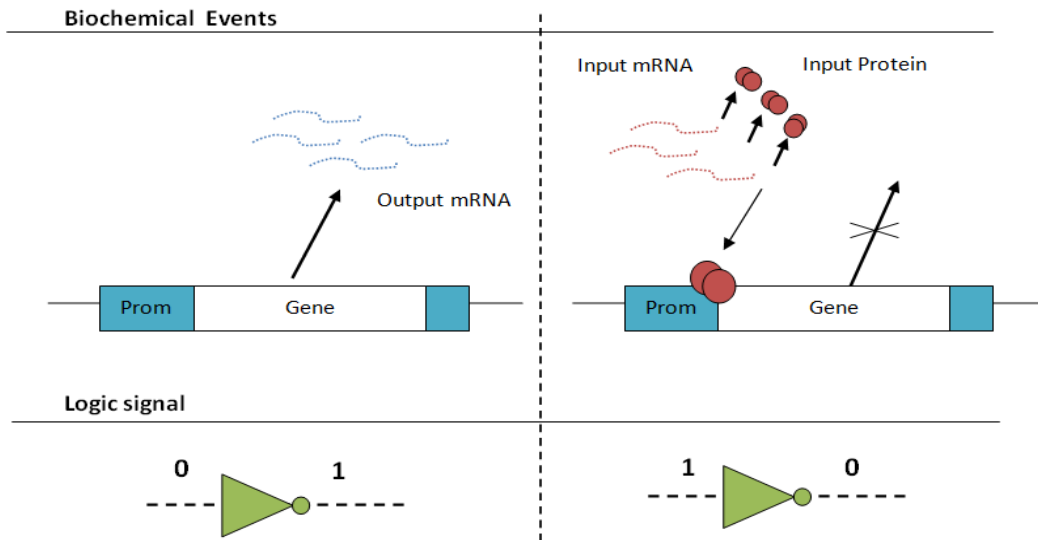


Figure 1: Genetic circuit's modules. A logic signal is represented by a certain mRNA concentration. The gate has a single input signal, the mRNA. Firstly the mRNA is absent and as a result the gene is transcribed into mRNA as signal output. Secondly, when the mRNA input signal is presented a protein is translated, this protein binds to an operator of the gene's promoter, and RNA polymerase is prevented from transcribing the gene. Applying Boolean logic to the system, where high protein concentrations represent '1' and low concentrations represent '0', it can be seen that the above system is behaving like an inverter. The gate is used to determine the intracellular state of the cell. Source: [Weiss, 2003]

Developing new and more sophisticated design techniques becomes necessary, as synthetic biology evolves. From this point of view, as electronic development carries out systematic comparisons of different possible designs based on user specifications in order to determine an optimal design, the need to conduct the same design process review in the field of synthetic biology comes up. An important design decision is the division of the features implemented at hardware/software level to make the study of the problems arising from the convergence of technologies in heterogeneous systems easier. In the context of synthetic biology applications, the design of wetware components, living organisms, and software/hardware components is required due to the need of obtaining at software level the signaling processing generated by the biological hardware components involved. It is recognized at higher level that the component wetware is heterogeneous due to transcription, translation and proteomic components handled in this kind of devices represent different areas of design. In this sense, it is necessary to use electronic engineers' knowledge about design methods of complex heterogeneous systems in order to optimize its efficiency, achieve more flexible biological applications and reduce the production costs of devices.

In the work developed by the group of David A. Ball in 2010 [Ball, 2010] three distinct solutions to a specification, namely the detection of distinct combinations of chemical signals. Figure 1. In the first option, hybrid promoters that contain binding sites for transcription factors responsive to the inputs are used to control the expression of fluorescent proteins. As the input signals of devices increase, the number of promoters needed to model the system increase also, this situation could be a disadvantage. The number of components needed to detect the different combinations of inputs may require more promoters which can be arranged. The second proposal implements the logic design at the protein level. This is accomplished by coupling each input to the expression of a non-fluorescent fragment of a fluorescent protein. This will get more flexibility to get larger and more complex

devices. However, the number of fluorescent proteins used in this modeling is also limited. By using protein self assembly, the logic design at this level provides to the device with a lower activation response at the transcription level which may be interesting, it depends on the device required. Finally the third level focuses on the study of the fluorescence spectrum for the identification of the input molecules. This option is to embed the logic in the electronic layer. In this case each input directly activates the expression of one of the different fluorescent proteins and the inputs present are determined by processing the pattern of fluorescence that is obtained. By implementing the logic outside of the biological system, the number of molecules possible to distinguish between is greatly increased, limited only by the number of transduction mechanisms and reporters.

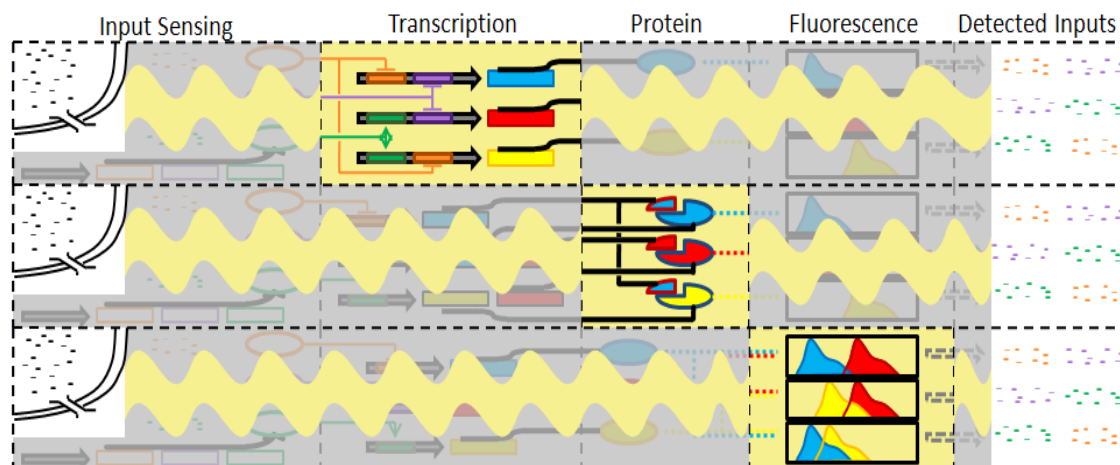


Figure 2. Implementation of logic in different design domains. The figure gives an overview of how each approach processes the environmental inputs. Wavy yellow lines indicate signal transduction, and yellow boxes highlight where the logic occurs. Source: [Ball,2010]

Applications

The main motivation of synthetic biology arises from the notion of programmable cells capable of resolving highly complex problems. Basic elements, for example promoters, ribosome binding sites and transcriptional repressors were combined to form small behaviors with specified modules. Devices as logic gates, switches or oscillators have been implemented by using this kind of techniques. These and other modules can be used to regulate gene expression, protein function, metabolism and cell-cell communication.

In the last few years many challenges have been tackled in order to achieve the main objectives of synthetic biology. Many developments and efforts have been combined to design and build approaches to use in fields such as bioremediation, biomedical therapies, molecular fabrication of biomaterials, sustainable energy production etc.

A first interesting work was proposed in [Weiss, 1999]. Weiss et al., presents a design paradigm for gene-expression based digital logic implemented in vivo. The proposed modular abstraction enables the construction of complex digital logic circuits into genetic regulatory networks using a library of interchangeable components. The chemical activity of such genetic network in vivo implements the computation specified by the digital circuit. Logic signals are implemented by rates of DNA binding proteins since they can function as transcriptional repressors. This manner the flow of logical information is represented as the effect of one protein on the transcription rate of

another. Gates are represented by structural genes which codifies output proteins. These genes are fused to promoter/operator regions that are regulated by input proteins.

An implementation of a specific device, an oscillator denominated the “repressilator”, was presented in [Elowitz, 2000]. This approach is composed of two plasmids consists to oscillating network in *Escherichia coli* bacteria that use three transcriptional repressor systems. The network periodically induces the synthesis of green fluorescent protein (GFP). The larger plasmid contains the oscillatory circuit of the repressors (LacI, tetR and CI). The first repressor inhibits the transcription of the second repressor gene which in turn inhibits the expression of a third gene completing the cycle, as is illustrated in the Figure 3. The observation of GFP permits monitoring the repressilator, which oscillates at a regular interval. The periodicity of the GFP cycle was much longer than the periodicity of cell division by the bacteria, which indicates the signaling mechanism outlived the lifetime of any given cell [Campbell, 2005].

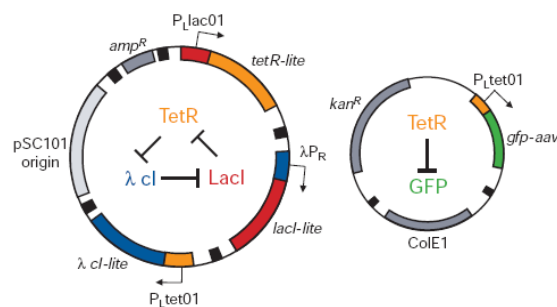


Figure 3: The repressilator network: cyclic negative-feedback loop composed of three repressor genes and their corresponding promoters, as shown schematically in the centre of the left-hand plasmid. Source: [Elowitz, 2000]

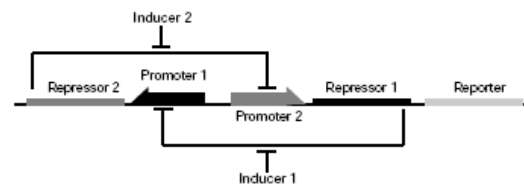


Figure 4: Toggle switch design: repressor 1 inhibits transcription from Promoter 1 and is induced by Inducer 1. Repressor 2 inhibits transcription from Promoter 2 and is induced by Inducer 2. Source: [Gardner, 2000]

Another interesting work based on logic circuit, published in the same year and that uses an *Escherichia coli* was published in [Gardner, 2000]. This work presents the design and construction of a genetic bistable toggle switch in *Escherichia coli*. Its design is simple: two promoters and two constitutive promoters. When the black gene is active, the gray gene and the reporter gene are silenced as illustrated in the Figure 4. In this sense the toggle exhibits bistability.

An original application “bacteria ‘photograph’” was reached by Levskaya et al., in [Levskaya, 2005]. In this work is showed a smart a light pattern as a high-definition chemical image. This system is switched between different states by red light and consists of a synthetic sensor kinase that allows a lawn of bacteria to function as a biological film, such that the projection of a pattern of light on to the bacteria produces a high-definition two-dimensional chemical image. This spatial control of bacterial gene expression could be used to ‘print’ complex biological materials, for example, and to investigate signaling pathways through precise spatial and temporal control of their phosphorylation steps.

To encourage the development of applications in this field and attract engineers to form interdisciplinary groups was created the iGEM - Synthetic Biology Summer Competition. This competition consist of teams of multidisciplinary groups of students whose goal is to design a genetic circuit with an interesting feature with its mathematical model and verify its operation with an implementation experimentally in the laboratory. This initiative was born of SynBioComm, an organization that aims to create a community of researchers in synthetic

biology through conferences such as the European Conference on Synthetic Biology and the participation in the iGEM, which is also promoted annually by the Massachusetts Institute of Technology.

In this context, the Pico-Pumbler project shows the results obtained by the iGEM. Pico-Pumbler presents a set of plumbers bacteria [PicoPlumber, 2009]. In this development *E. coli* is engineered to detect the breach responding to an inducer molecule (IPTG) released from this site. Bacterium is sensed by the inducer IPTG and swim towards breach so that can repair it. Pipes released he inducer only in the case of a breach in the wall. To do this, is used quorum sensing. This mechanism of some microorganisms permits that a bacterium can detect the presence of other bacteria in the neighborhood. The density of bacteria increases substantially close to the leak. Hence, by the quorum sensing signal, bacteria know that they have reached the leaking site, and therefore, they can start producing the glue proteins. Bacterium lyses after a certain time interval, during which they have produced a sufficient amount of the glue proteins. This project is designed in a modular architecture that permits that individual gene modules can be mathematically modeled and their design improved, before being built and tested and of this manner ensures a challenging research project.

Perspectives and Conclusions

The synthetic biology strategy consists of applying the knowledge of biological systems in order to design new biological devices with new and improved properties. This strategy is similar than the challenges that allowed organic chemistry to develop new organic compounds with interesting non-natural properties. This challenge concerns the adaptation of engineering practices to the design of biologically-inspired systems and the improvement of its development and standardization in order to design more complex system with a large set of functionalities. The development of practical synthetic biology devices will require a system-level analysis and a design approach that have yet to be explored.

A new meaning is given to the concept of information codification; the information is encoded into organic molecules with meaning as long the scientist as the organisms which are the support of the computation. As a result it is possible to design bacteria with the ability of autonomous computing in order to apply these functionalities in different fields like environment, new materials, industrial processes, energy and biomedicine. These applications are the main challenges of synthetic biology and in order to achieve them it is necessary to improve some of the synthetic biological devices design techniques. Most of synthetic biology efforts have to be focused on the following aspects: the identification, characterization and design of synthetic systems with programmable and controlled behavior. The development of efficient practices in order to integrate modules in a more complex system, the capacity to replace natural genomes by synthetic genomes is one of the main challenges to achieve. The characterization and standardization of biological modules in order to obtain engineers specialized in both areas: the design of basics modules and complex systems and finally the improvement of the no controlled noise of the biologically-inspired systems.

Bibliography

- [Adleman, 1994] Adleman L., Molecular Computation of Solutions to Combinatorial Problems. In: Science Journal. 266 (11): 1021-1024, 1994.
- [Andrianantoandro, 2006] Andrianantoandro E., Basu S., Karig D.K., Weiss R. Synthetic biology: new engineering rules for an emerging discipline. In: Molecular Systems Biology 2. Article number: 2006.0028, 2006.
- [Ball, 2010] Ball D, Lux M, Graef R, Peterson M, Valenti J, Dileo J, Peccoud J. Co-design in Synthetic Biology: A System-level Analysis of the Development of an Environmental Sensing Device. In: Pacific Symposium on Biocomputing. 15:385-396, 2010

- [Bulter, 2004] Bulter, T., Lee, S.-G., Wong, W., et al., Design of artificial cell-cell communication using gene and metabolic networks. In: Proc. Natl. Acad. Sci. USA 101, 2299–2304. 2004
- [Campbell,2005] Campbell M. Meeting Report: Synthetic Biology Jamboree for Undergraduates. In: Cell Biology Education, Vol. 4, 19–23, Spring, 2005.
- [Elowitz, 2000] Elowitz, M.B., Leibler, S. A synthetic oscillatory network of transcriptional regulators. In: Nature 403, 335–338, 2000
- [Endy, 2005] Endy D. Foundations for engineering biology. In: Nature 438: 449-453. 2005.
- [Gardner, 2000] Gardner, T.S., Cantor, C.R., Collins, J.J. Construction of a genetic toggle switch in Escherichia coli. In: Nature 403, 339–342, 2000
- [Kobayashi, 2004] Kobayashi H, Kaern M, Araki M, Chung K, Gardner TS, et al. Programmable cells: interfacing natural and engineered gene networks. In: Proceedings of the National Academy of Sciences of the United States of America 101: 8414-8419, 2004
- [Levskaya, 2005] Levskaya A, Chevalier A., Tabor J., Simpson Z., Lavery L.,et al. Engineering Escherichia coli to see light. In: Nature 438(7067):441-442, 2005
- [PicoPlumber, 2009] University of Aberdeen. PicoPlumber Project: Self-healing pipe system using engineered Escherichia coli. In: iGEM 2009.
- [Pleiss, 2006] Pleiss J., The promise of synthetic biology. In: Applied Microbiology and Biotechnology. Springer 73 Vol 4. Pags:735–739, 2006
- [Purnick, 2009] Purnick P.E.M., Weiss R. The second wave of synthetic biology: from modules to systems. In: Nature Reviews Molecular Cell Biology . 10, 410-422, 2009
- [Weiss, 1999] Weiss R, Homsy G, Knight TF.Jr. Toward in-vivo digital circuits. In: DIMACS Workshop on Evolution as Computation, Princeton Univ. Springer-Verlag, 1999
- [Weiss, 2003] Weiss R., Basu S., Hooshangi S., Kalmbach A., Karig D., Mehreja R., Netravali I. Genetic circuit building blocks for cellular computation, communications, and signal processing. In: Natural Computing 2, 47–84, 2003.
- [Weiss, 2005] Weiss R., McDaniel R. Advances in synthetic biology: on the path from prototypes to applications. In: Current Opinion in Biotechnology 2005, 16:476–483, 2005.

Authors' Information



Paula Cordero – Natural Computing Group. Universidad Politécnica de Madrid, Boadilla del Monte, 28660 Madrid, Spain: e-mail: p.cordero@alumnos.upm.es



Sandra Gómez - Natural Computing Group, Universidad Politécnica de Madrid, Carretera de Valencia Km 7., Madrid, Spain: email: sgomez@eui.upm.es



Rafael Gonzalo – Artificial Intelligence Department. Facultad de Informática. Universidad Politécnica de Madrid, Boadilla del Monte, 28660 Madrid, Spain. e-mail: rgonzalo@fi.upm.es

CHAIN SPLIT OF PARTIALLY ORDERED SET OF K-SUBSETS

Hasmik Sahakyan, Levon Aslanyan

Abstract: An application oriented class of partially ordered sets is considered. Let $P(n, k)$ denotes the set of all k -tuples with strictly increasing elements from the set $N = \{1, 2, \dots, n\}$ and $1 \leq k \leq n$. Some properties of $P(n, k)$ is studied in terms of partially ordered sets. An algorithm that constructs a set of non intersecting increasing chains that cover all elements of $P(n, 3)$ is brought. The number of these chains is the minimal possible: it equals to the width of $P(n, 3)$, i.e. the largest cardinality of an antichain. Analogous to the Hansel's well known algorithm for identification of monotone Boolean functions, the chains constructed for $P(n, 3)$ can be used for identification of monotone functions defined on $P(n, 3)$.

Keywords: partially ordered sets, chain split.

ACM Classification Keywords: G.2.1 Discrete mathematics: Combinatorics

Introduction

An application oriented class of partially ordered sets is considered. Let $P(n, k)$ denotes the set of all strictly increasing k -tuples of elements that are from the set $N = \{1, 2, \dots, n\}$, and for some $k, 1 \leq k \leq n$. Properties of $P(n, k)$ and consequently of $P(n, 3)$ is studied. An algorithm that constructs a set of non intersecting increasing chains that cover all elements of $P(n, 3)$ is brought. Analogous to the Hansel's well known algorithm for identification of monotone Boolean functions, based on partitioning of the set of vertices of the cube into non intersecting chains, - the chains, constructed for $P(n, 3)$ can be used for identification of monotone functions given in $P(n, 3)$. The study of $P(n, 3)$ is also motivated by its tight relation with the 3-hypergraphs.

Partially Ordered Sets

This section brings introduction to the partially ordered sets ([ST, 2008], [E, 1997]).

Definition 1. A partially ordered set (or poset) is an ordered pair (P, \leq) , consisting of a set P and a relation \leq on P satisfying the following three properties:

- (1) for all $x \in P$, $x \leq x$ (reflexivity).
- (2) for all $x, y \in P$, if $x \leq y$ and $y \leq x$, then $x = y$ (anti-symmetry).
- (3) for all $x, y, z \in P$, if $x \leq y$ and $y \leq z$, then $x \leq z$ (transitivity).

The notation $x < y$ is used when both $x \leq y$ and $x \neq y$.

Definition 2. An element x of a poset P is minimal if there is no element $y \in P$ s.t. $y < x$. Similarly, x is maximal if there is no element $z \in P$ s.t. $x < z$.

Two elements x and y in the poset P are comparable if $x \leq y$ or $y \leq x$; otherwise x and y are incomparable.

Definition 3. A *chain* in a poset (P, \leq) is a subset C of P which is totally ordered in P . An *antichain* is a set A of pairwise incomparable elements.

The *height* of a poset is the largest cardinality of a chain, and its *width* is the largest cardinality of an antichain. We denote the height and width of (P, \leq) by $h(P)$ and $w(P)$. In a finite poset (P, \leq) , a chain C and an antichain A have at most one element in common.

Theorem 1 (Dilworth's Theorem) *Let (P, \leq) be a finite poset. Then there is a partition of P into $w(P)$ chains.*

Let x and y be distinct elements of a poset (P, \leq) . We say that y *covers* x if $x < y$ but no element z satisfies relation $x < z < y$. The *Hasse diagram* of a poset (P, \leq) is the directed graph whose vertex set is P and whose arcs are the covering pairs (x, y) in the poset. We usually draw the Hasse diagram of a finite poset in the plane in such a way that, if x is covered by y , then the point representing y is higher than the point representing x . Then no arrows are required in the drawing, since the directions of the arrows are implicit

While Dilworth's theorem uses transitive comparisons in splitting (P, \leq) into the chains, our interest below concerns the chains consisting of pairwise covering vertices, that is chains in the Hasse diagram. A general postulation on existence of such chain splits on posets is not known. It is not hard to compose a simple poset P that can't be split into $w(P)$ increasing chains of covering vertices. Such decompositions are valid for a number of well known particular cases of posets such as the unit cube, the structure of unit cube subcubes by inclusion, etc. The poset that we investigate in this regard is the special order of k -subsets of a finite set, - the equivalent structure of the k -th layer of a unit cube.

***k*-subsets**

Let $P(n, k)$ denotes the set of all k -tuples with strictly increasing elements from the set $N = \{1, 2, \dots, n\}$, and for some $k, 1 \leq k \leq n$. $(i_1, \dots, i_k) \in P(n, k)$ iff $1 \leq i_1 < i_2 < \dots < i_k \leq n$. The number of elements of $P(n, k)$ is C_n^k . For two elements, (i_1, \dots, i_k) and (j_1, \dots, j_k) we define $<$ relation as follows: $(i_1, \dots, i_k) < (j_1, \dots, j_k)$ if and only of $i_1 < j_1, \dots, i_k < j_k$. Then $P(n, k)$ becomes a poset with minimal and maximal elements $(1, 2, \dots, k)$ and $(n - k + 1, \dots, n)$ respectively. We define the weight of (i_1, \dots, i_k) as the sum of its coordinates, $i_1 + \dots + i_k$. Now let us form the Hasse diagram of $P(n, k)$. The lowest layer of the diagram consists of the unique vertex $(1, 2, \dots, k)$. Then the i -th layer consists of all elements of $P(n, k)$ that cover some elements of the $(i - 1)$ -th layer. The highest layer contains the vertex $(n - k + 1, \dots, n)$. So the overall diagram consists of $k \cdot (n - k) + 1$ layers: we number them from 0 to $k \cdot (n - k)$.

All vertices of the i -th layer have equal weights which is $i + (1 + \dots + k)$. We introduce a notion of middle layer or layers, which is the $(k \cdot (n - k) / 2)$ -th layer for even k , or odd k and odd n ; and the $(k \cdot (n - k) \pm 1)$ -th layers for odd k and even n .

Each layer of $P(n, k)$ consists of pairwise incomparable elements, that is, it composes an antichain. According to the Dilworth theorem there is a partition of $P(n, k)$ into $w(P(n, k))$ chains. We will prove that $w(P(n, k))$ is achieved among the vertex sets of layers of $P(n, k)$. Our attention is restricted to the case $k = 3$ in regard to the framework of describing 3-hypergraph degree sequences [S, 2009], where 3 is the minimal number to check the complexity of algorithms for the hypergraph degree sequence problem [B, 1986]. $w(P(n, 3))$ is found, and increasing chains consisting of covering vertices, are constructed for $P(n, 3)$.

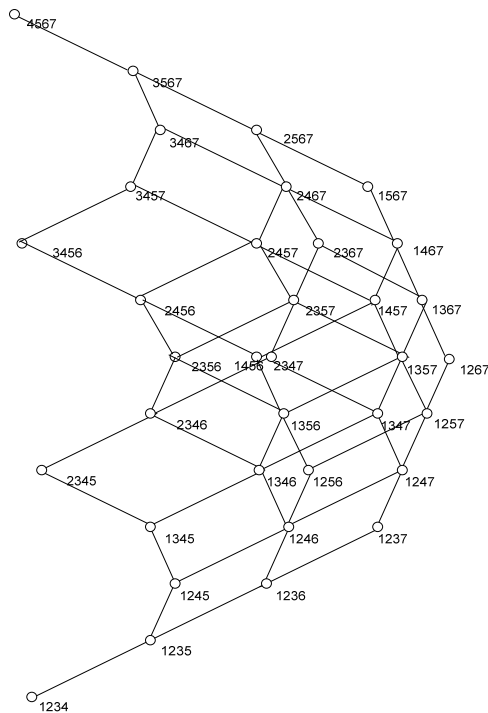


Figure 1. Hasse diagram of $P(7,4)$.

$P(n,3)$

In this section we give formula for calculating layer cardinalities of $P(n,3)$ and study some properties of $P(n,3)$ that will be used for determining the largest cardinality and to construct splitting to the chains.

Formula

Let L_l denotes the layer of $P(n,3)$ containing elements with the weight equal to l :

$L_l = \{(i_1, i_2, i_3) / i_1 + i_2 + i_3 = l, 1 \leq i_1 < i_2 < i_3 \leq n\}$. Calculation of $|L_l|$ below is done by determining the range of feasible values for each coordinate i_j .

It is easy to check that the minimal feasible value for i_1 is $\max(1, l - 2n + 1)$ and the maximal value equals $\lfloor l/3 \rfloor - 1$.

For a given feasible i_1 the minimal feasible value of i_2 is $\max(i_1 + 1, l - i_1 - n)$ and the maximal is $\lceil (l - i_1) / 2 \rceil - 1$.

For given i_1 and i_2 , i_3 is unique.

Resuming the above reasoning, we bring the formula of $|L_l|$:

$$|L_l| = \sum_{i_1 = \max(1, l - 2n + 1)}^{\lfloor l/3 \rfloor - 1} (\lceil (l - i_1) / 2 \rceil - \max(i_1 + 1, l - i_1 - n)).$$

For determining the layer of greatest cardinality which we intend, the formula given is improper, and we study further properties of $P(n,3)$.

Symmetry

The elements of $P(n,3)$ located on j -th layer have weight $j+6$. Middle layer of $P(n,3)$ is at $L_{mid} = \frac{3 \cdot (n-3)}{2}$ for odd n and there are two middle layers $L_{mid+} = \frac{3 \cdot (n-3)+1}{2}$ and $L_{mid-} = \frac{3 \cdot (n-3)-1}{2}$ for even n . $P(n,3)$ is symmetric in respect to its middle layer (layers). If j -th layer contains an element (i_1, i_2, i_3) for some j , then its "opposite" element that we define as $(n+1-i_3, n+1-i_2, n+1-i_1)$ is located on the $(3 \cdot (n-3) - j)$ -th layer. We denote by $\hat{P}(n,3)$ and $\check{P}(n,3)$ the parts of $P(n,3)$ above and below the middle layers respectively.

Partitioning

The structure of $P(n,3)$ naturally partitioned into 3 parts, denote them by $P^1(n,3)$, $P^2(n,3)$ and $P^3(n,3)$. $P^1(n,3)$ and $P^3(n,3)$ consists of the first and last $n-3$ layers of $P(n,3)$ respectively, and $P^2(n,3)$ consists of the remaining $n-3+1$ layers.

Consider a layer i from the part $P^1(n,3)$. It is simple to indicate one specific vertex $(1, 2, i+3)$ on this layer, which is used in forthcoming considerations. Symmetrically, $P^3(n,3)$ contains opposite to i layer $3(n-3) - i$ and the vertex $(n-i-2, n-1, n)$ on it. Our main attention is to the middle part $P^2(n,3)$. We count the layer widths of $P^2(n,3)$ from layers 1 to $n-3+1$, and indicate the vertex $(1, i+1, n)$ for the layer $i+1$. Obviously middle layer or layers belong to $P^2(n,3)$.

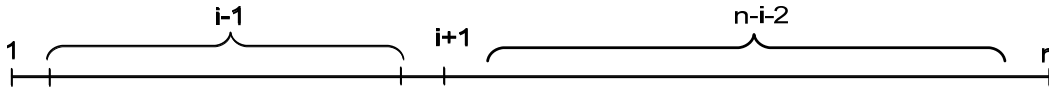
Quantities in $P^1(n,3)$ and $P^2(n,3)$

Elements of i -th layer of $P^1(n,3)$ can be generated starting from $(1, 2, i+3)$: a group of elements is generated by increasing the second coordinate 2 and decreasing the third one $i+3$ simultaneously. Then consider $(2, 3, i+1)$ and generate elements by increasing 3 and decreasing $i+1$. In general consider $(1+j, 2+j, i+3-2j)$ while $1+j \leq \left\lfloor \frac{1+2+i+3}{3} \right\rfloor - 1$, that is $j \leq \left\lfloor \frac{i}{3} \right\rfloor$, and generate elements by increasing the second coordinate and decreasing the third. It follows that the number of such elements increases with i , and therefore $P^1(n,3)$ has maximal number of elements on its last $(n-3-1)$ -th layer. Completely analogous is the situation for $P^3(n,3)$.

Quantities in $P^2(n,3)$

Now consider the middle zone $P^2(n,3)$.

We construct the vertices (a, b, c) of a particular layer $i+1$ in this area. This layer as we know contains the vertex $\alpha_{i+1} = (1, i+1, n)$ which will be the origin of our constructions.



Construction will be done by several groups. The groups are defined as sets of vertices that have first coordinate fixed, for example, for α_{i+1} it is 1. For the first coordinate a we denote the corresponding group by G_{i+1}^a . In G_{i+1}^a let b_{\min} be the smallest possible value for the second coordinate, denoted it by b . The third coordinate which we denote by c , is determined in a unique way by a and b given. Let c_{\max} is the greatest possible value for the third coordinate for fixed a (it is determined by b_{\min}).

For a given a define group operation for generating all elements of the group. First compute b_{\min} and c_{\max} for a , then the group operation increases b_{\min} by one (shifts the position to the right) and decreases c_{\max} by one (shifts the position to the left). Evidently this group consists of all elements of layer $i + 1$, having a its first coordinate. All the groups by different first coordinates are non intersecting. Thus $\bigcup_a G_{i+1}^a$ represents the layer $i + 1$ of $P^2(n,3)$. Moreover, it is easy to calculate the group sizes when we know b_{\min} and c_{\max} : it simply equals $1 + \lfloor (c_{\max} - b_{\min} - 1) / 2 \rfloor$ or the same $\lfloor (c_{\max} - b_{\min} + 1) / 2 \rfloor$.

Further we do two types of actions – compute the group sizes for all feasible a 's, and compute and compare the groups of neighbor layers $i + 1$ and i . Last action intends to determine the layer of maximum size in $P^2(n,3)$.

We start from α_{i+1} . Then increase a of α_{i+1} by one. To keep the vertex in the same layer we decrease the second coordinate b by one, the third, c remains the same, - currently it is n . Repeating this operation while new b is greater than new a , we get series of vertices of layer $i + 1$. These vertices have the property that $b = b_{\min}$ and $c = c_{\max} = n$ for the a fixed. Determine the values of a and b at the end of these series. 2 cases are possible: a) a and b meet at $a = (i - 1) / 2$ and $b_{\min} = (i - 1) / 2 + 1$ for even $i - 1$. b) a and b meet at the positions $(i - 2) / 2$ and $(i - 2) / 2 + 2$ when $i - 1$ is odd. In case b) the value $(i - 2) / 2 + 1$ between a and b is not used. Then increasing a and decreasing c by one we get the element $((i - 2) / 2 + 1, (i - 2) / 2 + 2, n - 1)$, which generates an additional group. Denote this group by G_{i+1}^* . All groups constructed at this stage are called $G1_{i+1}$ groups.

Now a still can be increased while it reaches the great possible value for a , that is: $a = \lfloor (1 + i + 1 + n - 3) \rfloor / 3$. Continue increasing a . At this stage increasing a causes increasing also b . Increase a and b by one (this is the smallest b , that is b_{\min}) and decrease c by two (this is c_{\max}). Repeating these operation, which ends at a triple (a, b, c) with $a = \lfloor (1 + i + 1 + n - 3) \rfloor / 3$, we get new groups of elements, that we call $G2_{i+1}$ groups.

Below the table represents two series of groups, first for layer $i + 1$ and second for i . Consider the case when $i - 1$ is odd.

Layer $i + 1$	$G1_{i+1}^1$	$G1_{i+1}^2$...	$G1_{i+1}^{(i-2)/2}$	G_{i+1}^*	$G2_{i+1}^{(i-2)/2+2}$...
Layer i	$G1_i^1$	$G1_i^2$...	$G1_i^{(i-2)/2}$	$G2_i^{(i-2)/2+2}$	$G2_i^{(i-2)/2+3}$...

An important notion is that all groups $G1_i^1, G1_i^2, \dots, G1_i^{(i-2)/2-1}$ are congruent to groups $G1_{i+1}^2, G1_{i+1}^3, \dots, G1_{i+1}^{(i-2)/2}$ correspondingly, their sizes are equal and they might be eliminated in comparisons of layers i and $i + 1$. The case when $i + 1$ is even is similar to this one.

Groups and their sizes

$G1_{i+1}^{1+j}$ The group consists of elements where $a=1+j$, $b=i+1-j$ and $c=n:(1+j, i+1-j, n)$. Possible values for j are $0, 1, \dots, (i-1)/2$ for even $i-1$ and $0, 1, \dots, (i-2)/2$ for odd $i-1$. The subgroup of each j is generated by the group operation. So the subgroup of j contains:

$$1 + \left\lfloor \frac{n - (i+1-j) - 1}{2} \right\rfloor = \left\lfloor \frac{n - i + j}{2} \right\rfloor \text{ elements.}$$

The last subgroup starts with the element $(1 + (i-1)/2, 2 + (i-1)/2, n)$, or the same $((i+1)/2, 1 + (i+1)/2, n)$ for even $i-1$ and $((i+1)/2, 2 + (i+1)/2, n)$ for odd $i-1$.

So we get $\sum_{j=0}^{\frac{i-1}{2}} \left\lfloor \frac{n - i + j}{2} \right\rfloor$ elements for even $i-1$ and $\sum_{j=0}^{\frac{i-2}{2}} \left\lfloor \frac{n - i + j}{2} \right\rfloor$ for odd $i-1$.

G_{i+1}^* This group exists only for odd $i-1$. It starts with the element $(2 + (i-2)/2, 3 + (i-2)/2, n-1)$, or the same $(1 + i/2, 2 + i/2, n-1)$, and generates $\left\lfloor \frac{n - 1 - (2 + i/2) - 1}{2} \right\rfloor$ elements. So $|G_{i+1}^*| = \left\lfloor \frac{n - 4 - i/2}{2} \right\rfloor$.

$G1_{i+1}$ is the union of all these sets: $G1_{i+1} = \left(\bigcup_j G1_{i+1}^{1+j} \right) \cup G_{i+1}^*$.

$G2_{i+1}$ group described above consists of elements, where $a=1+(i-1)/2+j=(i+1)/2+j$, $b=2+(i-1)/2+j=1+(i+1)/2+j$ and $c=n-2j$, where possible values for j are $1, \dots$, while $(i+1)/2+j \leq \left\lfloor \frac{i+n-1}{3} \right\rfloor$, that is, $j \leq \left\lfloor \frac{i+n-1}{3} \right\rfloor - \frac{i+1}{2}$, - for even $i-1$ and $i/2+j \leq \left\lfloor \frac{i+n-1}{3} \right\rfloor$,

$j \leq \left\lfloor \frac{i+n-1}{3} \right\rfloor - \frac{i}{2}$, for odd $i-1$. The subgroup of each j is generated by the group operation. So the

subgroup of j contains $1 + \left\lfloor \frac{n - 2j - (1 + (i+1)/2 + j) - 1}{2} \right\rfloor$ elements.

$$|G2_{i+1}| = \sum_{j=1}^{\left\lfloor \frac{i+n-1}{3} \right\rfloor - \frac{i+1}{2}} \left\lfloor \frac{n - 2j - ((i+1)/2 + j)}{2} \right\rfloor \text{ for even } i-1. \text{ For odd } i-1$$

$$|G2_{i+1}| = \sum_{j=1}^{\left\lfloor \frac{i+n-1}{3} \right\rfloor - \frac{i}{2}} \left\lfloor \frac{n - 2j - (i/2 + j)}{2} \right\rfloor$$

All the above reasoning prove the following theorem:

Theorem: $G1_{i+1}$ and $G2_{i+1}$ are non intersecting groups that cover the $(i+1)$ -th layer of $P^2(n,3)$.

Our next goal is to find the areas of increasing cardinalities among the neighbor layers. Compose $G1$ and $G2$ groups for the i -th layer of $P^2(n,3)$. We will consider the case of even $i-1$ only. The case of odd $i-1$ can be done in an analogous way.

$G1_i^{1+j}$ is the group of elements where $a = 1 + j$, $b = i - j$ and the third is $c = n : (1 + j, i - j, n)$, where possible values for j are $0, 1, \dots, (i - 1)/2 - 1$. The subgroup of each j is generated by the group operation. So the subgroup of j contains:

$$1 + \left\lfloor \frac{n - (i - j) - 1}{2} \right\rfloor = \left\lfloor \frac{n - i + j + 1}{2} \right\rfloor \text{ elements. The last subgroup starts with the element } ((i - 1)/2, (i - 1)/2 + 2, n). \text{ So we get } \sum_{j=0}^{\frac{i-1}{2}-1} \left\lfloor \frac{n - i + j + 1}{2} \right\rfloor \text{ elements.}$$

Here we have an additional group.

$G1_i^*$ starts with the element $((i - 1)/2 + 1, (i - 1)/2 + 2, n - 1) = ((i + 1)/2, 1 + (i + 1)/2, n - 1)$, which generates $|G1_i^*| = \left\lfloor \frac{n - 1 - (i + 1)/2}{2} \right\rfloor$ elements by the group operation. Then $G1_i$ is the union of these subgroups: $G1_i = \left(\bigcup_j G1_i^{1+j} \right) \cup G1_i^*$.

$G2_i$ This group contains elements where $a = 1 + (i - 1)/2 + j = (i + 1)/2 + j$, $b = 1 + (i + 1)/2 + j$ and $c = n - 1 - 2j$, where possible values for j are $1, \dots$, while $\frac{i + 1}{2} + j \leq \left\lfloor \frac{n + i - 2}{3} \right\rfloor$, that is $j \leq \left\lfloor \frac{n + i - 2}{3} \right\rfloor - \frac{i + 1}{2}$. Then the subgroup of each j is generated by the group operation. So the subgroup of j contains $1 + \left\lfloor \frac{n - 1 - 2j - (1 + (i + 1)/2 + j) - 1}{2} \right\rfloor$ elements.

$$|G2_i| = \sum_{j=1}^{\left\lfloor \frac{n+i-2}{3} \right\rfloor - \frac{i+1}{2}} \left\lfloor \frac{n - 2j - (1 + (i + 1)/2 + j)}{2} \right\rfloor$$

Calculate the differences: $|G1_{i+1}| - |G1_i|$ and $|G2^{i+1}| - |G2^i|$.

$$|G1_{i+1}| - |G1_i| = \sum_{j=0}^{\frac{i-1}{2}} \left\lfloor \frac{n - i + j}{2} \right\rfloor - \sum_{j=0}^{\frac{i-1}{2}-1} \left\lfloor \frac{n - i + j + 1}{2} \right\rfloor - \left\lfloor \frac{n - 1 - (i + 1)/2}{2} \right\rfloor = \left\lfloor \frac{n - i}{2} \right\rfloor - \left\lfloor \frac{n - 1 - (i + 1)/2}{2} \right\rfloor$$

$$|G2^{i+1}| - |G2^i| = \sum_{j=1}^{\left\lfloor \frac{n+i-1}{3} \right\rfloor - \frac{i+1}{2}} \left\lfloor \frac{n - 2j - ((i + 1)/2 + j)}{2} \right\rfloor - \sum_{j=1}^{\left\lfloor \frac{n+i-2}{3} \right\rfloor - \frac{i+1}{2}} \left\lfloor \frac{n - 2j - 1 - ((i + 1)/2 + j)}{2} \right\rfloor$$

Consider cases:

a) 3 is divisor of $n + i - 1$, it follows that $\left\lfloor \frac{n + i - 2}{3} \right\rfloor = \left\lfloor \frac{n + i - 1}{3} \right\rfloor - 1$

b) $(n+i-1)/3$, 1 remainder, it follows that $\left\lfloor \frac{n+i-2}{3} \right\rfloor = \left\lfloor \frac{n+i-1}{3} \right\rfloor$

c) $(n+i-1)/3$, 2 remainder, it follows that $\left\lfloor \frac{n+i-2}{3} \right\rfloor = \left\lfloor \frac{n+i-1}{3} \right\rfloor$

Consider b) or c)

$$|G_{2_{i+1}}| - |G_{2_i}| = \sum_{j=1}^{\left\lfloor \frac{n+i-1}{3} \right\rfloor - \frac{i+1}{2}} \left(\left\lfloor \frac{n-2j - ((i+1)/2 + j)}{2} \right\rfloor - \left\lfloor \frac{n-2j-1 - ((i+1)/2 + j)}{2} \right\rfloor \right) =$$

$$\sum_{j=1}^{\left\lfloor \frac{n+i-1}{3} \right\rfloor - \frac{i+1}{2}} \left(\left\lfloor \frac{n-(i+1)/2 - 3j}{2} \right\rfloor - \left\lfloor \frac{n-(i+1)/2 - 1 - 3j}{2} \right\rfloor \right)$$

1) $n-(i+1)/2$ is even, then it follows that $n-(i+1)/2-3j$ is even for even j and is odd for odd j .

1a) $n-(i+1)/2$ is even and j is even, and then it follows that

$$\left\lfloor \frac{n-(i+1)/2 - 3j}{2} \right\rfloor = \left\lfloor \frac{n-(i+1)/2 - 1 - 3j}{2} \right\rfloor + 1.$$

1b) $n-(i+1)/2$ is even and j is odd, then it follows that $\left\lfloor \frac{n-(i+1)/2 - 3j}{2} \right\rfloor = \left\lfloor \frac{n-(i+1)/2 - 1 - 3j}{2} \right\rfloor$

So in case 1a) $|G_{2_{i+1}}| - |G_{2_i}| = \sum_{\text{even } j}^{\left\lfloor \frac{n+i-1}{3} \right\rfloor - \frac{i+1}{2}} 1$, approximately the half of the upper index.

2) $n-(i+1)/2$ is odd, then it follows that $n-(i+1)/2-3j$ is odd for even j and is even for odd j .

2a) $n-(i+1)/2$ is odd and j is even, and then it follows that

$$\left\lfloor \frac{n-(i+1)/2 - 3j}{2} \right\rfloor = \left\lfloor \frac{n-(i+1)/2 - 1 - 3j}{2} \right\rfloor.$$

2b) $n-(i+1)/2$ is odd and j is odd, and then it follows that

$$\left\lfloor \frac{n-(i+1)/2 - 3j}{2} \right\rfloor = \left\lfloor \frac{n-(i+1)/2 - 1 - 3j}{2} \right\rfloor + 1.$$

So in case 2b) $|G_{2_{i+1}}| - |G_{2_i}| = \sum_{\text{odd } j}^{\left\lfloor \frac{n+i-1}{3} \right\rfloor - \frac{i+1}{2}} 1$, approximately the half of the upper index.

Further analysis of all possible cases provides that the $(n+1)/2$ -th (for odd n) and $n/2$ -th and $(n/2+1)$ -th (for even n) layers of $P^2(n,3)$, - serve as layers of the largest cardinality for $P(n,3)$. In both cases these are the middle layers.

Chain Split

In this part our goal is to split $P(n,3)$ into disjoint chains of covering pair sequences. Then each chain must contain exactly one element of the antichain of largest cardinality, and consequently will pass through the middle layer/layers. Due to the symmetry property it is sufficient to have chain constructions only for $\hat{P}(n,3)$ or $\check{P}(n,3)$ and then the extended construction is by symmetry.

Notice that the antichain of the largest cardinality contains the element $(1, (n+1)/2, n)$ for odd n and the 2 antichains of the largest cardinality contains $(1, n/2, n)$ and $(1, n/2 + 1, n)$ respectively, for even n .

Algorithm

1. **Ordering of elements.** Consider lexicographic order of elements on layers;
2. **Constructing chain fragments in $\check{P}(n,3)$.** Consider a recurrent procedure. First chain starts with the element $(1,2,3)$. Any current chain starts with the smallest unused element of the lowest layer that still contains unused elements and goes up until it reaches the layer L_{mid} for odd n (L_{mid+} for even n). From this point we go up by increasing the third component until it reaches n or the middle layer. If we come in some step across an element which is already used in previous chains, then we go back and increase the second component by one and then continue increasing the third. If also the increase of second component moves the element to the used one, then we go back and increase the first component by one, and continue increasing the second, etc. until the chain reaches the middle layer or finds a deadlock. For an element e from L_{mid} (L_{mid+}) we denote by $C(e)$ the chain reaching this element. The first chain that started at $(1,2,3)$ reaches $(1, (n+1)/2, n)$ for odd n and $(1, (n/2 + 1), n)$ – for even n .

As an example consider $P(9,3)$. 159, 168, 249, 258, 267, 348, 357, 456 lists the elements of layer L_{mid} .

Chains in $\check{P}(9,3)$ constructed by the algorithm are:

$$C(159) = \{123, 124, 125, 126, 127, 128, 129, 139, 149, 159\},$$

$$C(168) = \{134, 135, 136, 137, 138, 148, 158, 168\},$$

$$C(249) = \{234, 235, 236, 237, 238, 239, 249\}, \quad C(267) = \{145, 146, 147, 157, 167, 267\}$$

$$C(258) = \{245, 246, 247, 248, 258\}, \quad C(357) = \{156, 256, 257, 357\},$$

$$C(348) = \{345, 346, 347, 348\}, \quad C(456) = \{356, 456\}.$$

3. **Extending chains to the $\hat{P}(n,3)$.** Complete chains of $P(n,3)$ are constructed in a way of extending the chains of $\check{P}(n,3)$ into the $\hat{P}(n,3)$ area. We use the symmetry property of $P(n,3)$ in the following way. For each element $e = (i_1, i_2, i_3)$ of L_{mid} , it is easy to check that its "opposite" to $e = (i_1, i_2, i_3)$ – the element $e^{op} = (n+1-i_1, n+1-i_2, n+1-i_3)$ also belongs to L_{mid} , for odd n . When n is even, for each element $e = (i_1, i_2, i_3)$ of $L_{(mid-)}$, its "opposite" element $e^{op} = (n+1-i_1, n+1-i_2, n+1-i_3)$ belongs to $L_{(mid+)}$, and vice versa. For the above example: $159^{op} = 159$, $168^{op} = 249$, $249^{op} = 168$, $267^{op} = 348$, $258^{op} = 258$, $357^{op} = 357$, $348^{op} = 267$, $456^{op} = 456$.

Then continuation of a chain $C(e)$ of $\tilde{P}(n,3)$ into the $\hat{P}(n,3)$ area considers the chain denoted by $C^{up}(e)$, that consists of all "opposite" elements $C(e^{op})$ of the $C(e)$ taking in inverse order.

$$C^{up}(159) = \{149^{op}, 139^{op}, 129^{op}, 128^{op}, 127^{op}, 126^{op}, 125^{op}, 124^{op}, 123^{op}\} =$$

$$\{169, 179, 189, 289, 389, 489, 589, 689, 789\}, C^{up}(168) = \{178, 278, 378, 478, 578, 678\},$$

$$C^{up}(249) = \{259, 269, 279, 379, 479, 579, 679\}, C^{up}(267) = \{367, 467, 567\},$$

$$C^{up}(258) = \{268, 368, 468, 568\}, C^{up}(357) = \{358, 458, 459\},$$

$$C^{up}(348) = \{349, 359, 369, 469, 569\}, C^{up}(456) = \{457\}.$$

So we get the chains:

$$\{123, 124, 125, 126, 127, 128, 129, 139, 149, 159, 169, 179, 189, 289, 389, 489, 589, 689, 789\}$$

$$\{134, 135, 136, 137, 138, 148, 158, 168, 178, 278, 378, 478, 578, 678\}$$

$$\{234, 235, 236, 237, 238, 239, 249, 259, 269, 279, 379, 479, 579, 679\}$$

$$\{145, 146, 147, 157, 167, 267, 367, 467, 567\}$$

$$\{245, 246, 247, 248, 258, 268, 368, 468, 568\}$$

$$\{156, 256, 257, 357, 358, 458, 459\}$$

$$\{345, 346, 347, 348, 349, 359, 369, 469, 569\}$$

$$\{356, 456, 457\}$$

And the whole construction given by the algorithm is illustrated in the figure 2.

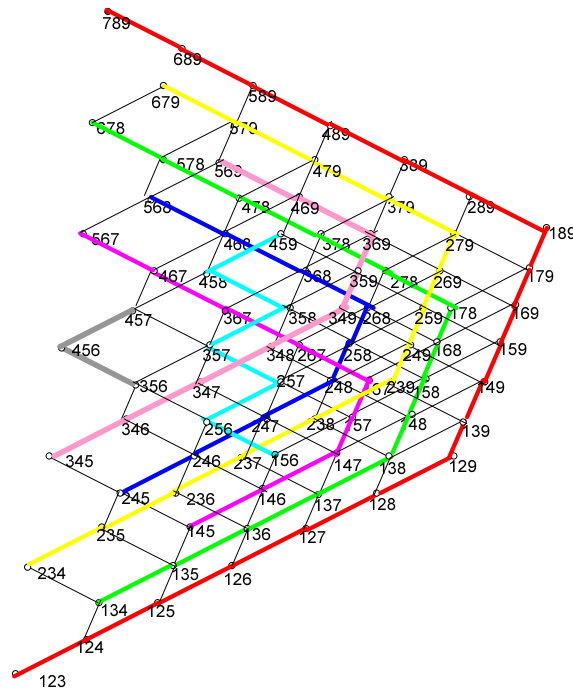


Figure 2

Correctness of the algorithm. First is the claim about the deadlock free of algorithms when growing the chains in $\tilde{P}(n,3)$. Then, it is to prove that chains constructed in step 2 cover all the elements of $\tilde{P}(n,3)$ symmetrically.

This two steps are done by induction on n taking into account the structure of l -th layer of $P(n,3)$ that is a union of $\leq l$ layers of $P(n-i,2)$, for $i = 1, \dots$

Final comparisons and computation of the chains are by formulas given above.

Conclusion

We constructed non intersecting increasing chains that cover all elements of $P(n,3)$. Theoretical outcome is that in addition to the chains by Dilworth's theorem we prove the existence of chains consisted of covering elements, - as an analogy to the Hansel chains for binary cubes. The practical outcome is the monotone recognition of subsets C_k of elements of k -th layer of the n dimensional unit cube by the use of queries about the involvement of several vertices into the C_k .

Bibliography

- [E, 1997] K. Engel. Sperner Theory, Encyclopaedia of Mathematics and Its Applications, Vol. 65, Cambridge University Press, New York, NY, 1997, p. 417
- [ST, 2008] R. Steven. Lattices and Ordered Sets, 2008, ISBN 0-387-78900-2, 305 pp.
- [B, 1986] D. Billington, Lattices and degree sequences of uniform hypergraphs, ARS Combinatoria, vol.21 A (1986), pp.9-19
- [S, 2009] H. Sahakyan. Numeral characterization of n-cube subset partitioning, Discrete Applied Mathematics, vol.157 (2009), issue 9, 2191-2197.
- [AS, 2009] . L. Aslanyan and H. Sahakyan, Chain split and computation in practical rule mining, Information Science and Computing, International book series no. 8., Classification, forecasting, data mining, 2009, pp.132-135.

Authors' Information



Hasmik Sahakyan – *Leading Researcher, Institute for Informatics and Automation Problems, NAS RA, P.Sevak St. 1, Yerevan 14, Armenia, e-mail: hasmik@jgia.sci.am*



Levon Aslanyan – *Head of Department, Institute for Informatics and Automation Problems, NAS RA, P.Sevak St. 1, Yerevan 14, Armenia, e-mail: lasl@sci.am*

UPPER BOUND ON RATE-RELIABILITY-DISTORTION FUNCTION FOR SOURCE WITH TWO-SIDED STATE INFORMATION

Mariam Haroutunian, Arthur Muradyan

Abstract: Different models of the source with side information can be considered when side information is known to the encoder, the decoder, both of them, none of them. In this paper, we investigate a generalized model of the discrete memoryless source with two-sided state information introduced by Cover and Chiang in [Cover-Chiang, 2002], which includes the data compression problems mentioned above as special cases. We study the rate-reliability-distortion function, which is understood as the minimum code rate for the encoding of the source messages under the requirement that the decoder reconstructs the messages at a desired distortion level with the error probability exponentially decreasing with the codeword length. In other words, the rate is considered as a function of a fixed distortion level and the error exponent. In this paper the upper bound on the rate-reliability-distortion function is obtained. The upper bounds on rate-reliability-distortion functions of the source with side information are derived as special cases for four possible situations - one of which coincides with known result while the three others were unknown.

Keywords: source with side information, rate-reliability-distortion function

ACM Classification Keywords: H.0 Information Systems - Conference proceedings

Introduction

The state information problems were intensively studied. The model when state information is available to the decoder was analyzed by Wyner and Ziv in [Wyner-Ziv, 1976] where the rate-distortion function was derived which shows dependence of minimal rate on a required distortion introduced by Shannon in [Shannon, 1959]. The lossless source coding problem when state information is available at the decoder was investigated by Slepian-Wolf [Slepian-Wolf, 1973].

The applications of these problems include distributed sensor networks [Xiong-Liveris, 2004], digital upgrade of analog television signals, play-back of the compressed sound in the presence of background noise where decoder is fed with a background correlated signal to improve the quality of decoding. The source coding problem when state information is known to both encoder and decoder is studied in [Viswanathan-Berger, 1997]. The study includes applications in video coding where the pixel value at a given location depends on a pixel at the same location in a previous frame. Here, the previous frame can be considered as a side information for the

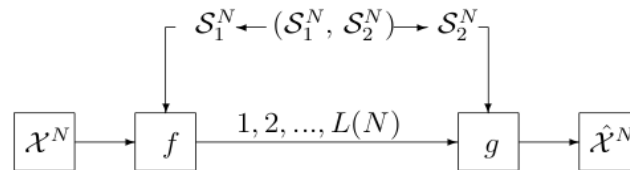


Figure 1. Source with two-sided state information

coding of the present frame.

A generalized model of sources where the encoder and the decoder have correlated state information (Figure 1) was considered by Cover and Chiang in [Cover-Chiang, 2002] where the rate-distortion function was derived. It

was proved that rate-distortion functions of the source with state information in four possible situations can be obtained from the generalized formula.

We study the rate-reliability-distortion function introduced by Haroutunian and Mekoush [Haroutunian-Mekoush, 1984] which describes the dependence of rate on reliability and distortion level. The idea was then adopted and extended for multiuser source coding problems [Haroutunian et al, 1998, Haroutunian-Maroutian, 1991, Maroutian-1990, Meulen et al, 2000]. This function is a generalization of the rate-distortion function, since it tends to that function when the error exponent (reliability - E) tends to 0. The inverse order dependence of these parameters is studied by Marton in [Marton, 1974].

In this paper, an upper bound of the rate-reliability-distortion function is derived for the source with two sided state information. The limit of this bound for $E \rightarrow 0$ coincides with the rate-distortion function, obtained in [Cover-Chiang, 2002]. As a special case, we derive the upper bounds on the rate-reliability-distortion functions for four possible situations of the source with side information, one of them coincides with the rate-reliability-distortion function of the DMS [Haroutunian et al, 2008], while the three others were unknown.

In the next section we give the definitions of the concepts extended for the considered generalized model. Description of the main theorem along with corollaries is given in section 3. Proof of the theorem is given in section 4.

Notations and Definitions

Capital letters are used for random variables S_1, S_2, U, X, \hat{X} taking values in the finite sets S_1, S_2, U, X, \hat{X} , respectively, and lower case letters s_1, s_2, u, x, \hat{x} for their realizations. Small bold letters are used for N -length vectors $\mathbf{x} = (x_1, \dots, x_N) \in X^N$.

A generalized model representing the source with two-sided state information is depicted in Figure 1.

S_1 and S_2 are the state information, known to the encoder and the decoder taking values from the set S_1 and S_2 , X is an i.i.d. random variable taking values in the finite set X (the alphabet of messages of the source). The finite set \hat{X} different from the set X , represents the reproduction alphabet of the receiver, in general case.

The generating probability distribution of the source with two-sided state information is given as

$$P^* = P_1^* \circ P_2^* = \{P^*(x, s_1, s_2) = P_1^*(x, s_1)P_2^*(s_2 | x, s_1), x \in X, s_1 \in S_1, s_2 \in S_2\}.$$

We consider the memoryless source, which means that the probability of N -length vector of message $\mathbf{x} = (x_1, \dots, x_N) \in X^N$ and state information vectors $\mathbf{s}_1 = (s_{11}, \dots, s_{1N}) \in S_1^N$, $\mathbf{s}_2 = (s_{21}, \dots, s_{2N}) \in S_2^N$ is defined as the product of component probabilities

$$P^{*N}(\mathbf{x}, \mathbf{s}_1, \mathbf{s}_2) = \prod_{n=1}^N P^*(x_n, s_{1n}, s_{2n}).$$

Let

$$d : X \times \hat{X} \rightarrow [0, \infty)$$

be the given distortion between the source and the reconstructed message. The distortion measure for the vectors $\mathbf{x} \in X^N$ and $\hat{\mathbf{x}} \in \hat{X}^N$ is defined as the average of the components' distortions

$$d(\mathbf{x}, \hat{\mathbf{x}}) = \frac{1}{N} \sum_{n=1}^N d(x_n, \hat{x}_n).$$

The code (f_N, g_N) is a pair of the encoding and the decoding functions

$$f_N : X^N \times S_1^N \rightarrow \{1, 2, \dots, L(N)\}$$

and

$$g_N : \{1, 2, \dots, L(N)\} \times S_2^N \rightarrow \hat{X}^N$$

where $L(N)$ is the *volume* of the code.

The task of the system is to ensure reconstruction of the source messages at the receiver at a given distortion level Δ and with a small error probability. Our problem is estimating of the minimum of the code volume.

To define error probability for given code (f_N, g_N) and $\Delta \geq 0$ consider the following set of triples:

$$A = \{\mathbf{x} \in X^N, \mathbf{s}_1 \in S_1^N, \mathbf{s}_2 \in S_2^N : g_N(f_N(\mathbf{x}, \mathbf{s}_1), \mathbf{s}_2) = \hat{\mathbf{x}}, d(\mathbf{x}, \hat{\mathbf{x}}) \leq \Delta\}.$$

The error probability of the code (f_N, g_N) , for the source probability distortion P^* , Δ and N is defined as:

$$e(f_N, g_N, P^*, \Delta, N) \leq 1 - P^{*N}(A).$$

A positive number R is called the (E, Δ) -*achievable rate* for a given P^* , $E > 0$ and $\Delta \geq 0$ if there exists a code (f_N, g_N) such that

$$\frac{1}{N} \log L(N) \leq R + \varepsilon$$

and error probability is exponentially small

$$e(f_N, g_N, P^*, \Delta, N) \leq \exp\{-N(E - \delta)\}.$$

for every $\varepsilon > 0$, $\delta > 0$ and sufficiently large N . The minimum (E, Δ) -achievable rate is denoted by $R(E, \Delta, P^*)$ and is called the *rate-reliability-distortion* function.

For notion of *types*, *mutual information* $I_{P, Q_1}(U \wedge X)$, *divergence* $D(P \| P^*)$, we refer to [Cover-Thomas, 1991, Csiszár-Körner, 1981, Csiszár-1998].

We introduce the following probability distributions with some auxiliary finite set U :

$$\begin{aligned} Q_1 &= \{Q_1(u | x, s_1), x \in X, s_1 \in S_1, u \in U\}, \\ Q_2 &= \{Q_2(\hat{x} | u, s_2), \hat{x} \in \hat{X}, u \in U, s_2 \in S_2\}, \\ P &= P_1 \circ P_2 = \{P(x, s_1, s_2) = P_1(x, s_1)P_2(s_2 | x, s_1), x \in X, s_1 \in S_1, s_2 \in S_2\}, \\ PQ(x, \hat{x}) &= \sum_{s_1, s_2, u} P(x, s_1, s_2) Q_1(u | x, s_1) Q_2(\hat{x} | u, s_2). \end{aligned}$$

Following estimates [Cover-Thomas, 1991, Csiszár-Körner, 1981] are used in the paper. For any type $P_1 \in P_N(X, S_1)$

$$(N+1)^{-|X||S_1|} \exp\{NH_{P_1}(X, S_1)\} \leq |T_{P_1}^N(X, S_1)| \leq \exp\{NH_{P_1}(X, S_1)\} \quad (1)$$

and any conditional type Q_1 and $\mathbf{u} \in T_{P, Q_1}(U)$

$$(N+1)^{-|X||S_1||U|} \exp\{NH_{P, Q_1}(X, S_1 | U)\} \leq |T_{P, Q_1}^N(X, S_1 | \mathbf{u})| \leq \exp\{NH_{P, Q_1}(X, S_1 | U)\}. \quad (2)$$

The number of probability distributions on X, S_1, S_2 is upper estimated as follows:

$$|P_N(X, S_1, S_2)| < (N + 1)^{|X| |S_1| |S_2|}. \quad (3)$$

Formulation of Results

Let $\alpha(E, P^*) = \{P : D(P \| P^*) \leq E\}$ and

$$R'(E, \Delta, P^*) = \max_{P \in \alpha(E, P^*)} \min_{Q_1, Q_2 \in Q(P, \Delta)} [I_{P, Q_1}(U \wedge S_1, X) - I_{P, Q_2}(U \wedge S_2)],$$

where the minimization is carried under following distortion constraint

$$Q(P, \Delta) = \{(Q_1, Q_2) : \sum_{x, \hat{x}} d(x, \hat{x}) P Q(x, \hat{x}) \leq \Delta\}.$$

Theorem. R' is the upper bound of the rate-reliability-distortion function for any $E > 0, \Delta > 0$ and P^*

$$R(E, \Delta, P^*) \leq R'(E, \Delta, P^*).$$

The proof of the theorem is given in the next section.

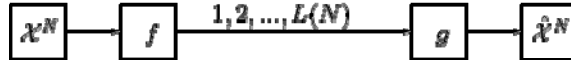
Corollary 1. We obtain the rate-distortion function $R(\Delta, P^*)$ established in [Cover-Chiang, 2002] when

$E \rightarrow 0$ for any $\Delta \geq 0$ and probability distribution P^*

$$\lim_{E \rightarrow 0} R'(E, \Delta, P^*) = \min_{Q_1, Q_2 \in Q(P^*, \Delta)} [I_{P^*, Q_1}(U \wedge S_1, X) - I_{P^*, Q_2}(U \wedge S_2)].$$

Corollary 2. We obtain the upper bounds on the rate-reliability-distortion functions for four possible situations of the source with side information as special cases.

Case 1: No state information at the sender and receiver: $S_1 = \emptyset, S_2 = \emptyset$



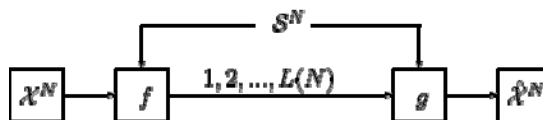
Here $(S_1, X) = X, I(U \wedge S_2) = 0$, with distributions $P(x), Q_1(u | x)$ and $Q_2(\hat{x} | u)$, $X \rightarrow U \rightarrow \hat{X}$ forms Markov chain. Therefore

$$\min_{Q_1(u|x)} I_{P, Q_1}(U \wedge X) \geq \min_{Q_2(\hat{x}|u)} I_{P, Q_2}(\hat{X} \wedge X), \text{ with equality iff } U = \hat{X}.$$

For this case we get the formula established in [Haroutunian et al, 2008]:

$$R'(E, \Delta, P^*) = \max_{P \in \alpha(E, P^*)} \min_{Q_1(u|x) Q_2(\hat{x}|u)} I_{P, Q_1}(U \wedge X) = \max_{P \in \alpha(E, P^*)} \min_{Q(\hat{x}|x)} I_{P, Q_1}(\hat{X} \wedge X).$$

Case 2: State information on both sides is the same: $S_1 = S_2 = S$



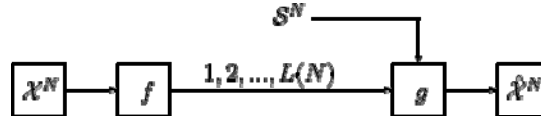
With distributions $P(x|s), Q_1(u|x, s)$ and $Q_2(\hat{x}|u, s)$, $X \rightarrow U \rightarrow \hat{X}$ forms Markov chain conditioned on S . Therefore

$$\min_{Q_1(u|x,s)} I_{P,Q_1}(U \wedge X | S) \geq \min_{Q_2(\hat{x}|x,s)} I_{P,Q_2}(\hat{X} \wedge X | S), \text{ with equality iff } U = \hat{X}.$$

We obtain the upper bound of rate-reliability-distortion function

$$\begin{aligned} R'(E, \Delta, P^*) &= \max_{P \in \alpha(E, P^*)} \min_{Q_1(u|x,s) Q_2(\hat{x}|u,s)} \left[I_{P,Q_1}(U \wedge S, X) - I_{P,Q_1}(U \wedge S) \right] = \\ &= \max_{P \in \alpha(E, P^*)} \min_{Q_1(u|x,s) Q_2(\hat{x}|u,s)} \left[I_{P,Q_1}(U \wedge S) + I_{P,Q_1}(U \wedge X | S) - I_{P,Q_1}(U \wedge S) \right] = \\ &= \max_{P \in \alpha(E, P^*)} \min_{Q_2(\hat{x}|x,s)} I_{P,Q_2}(\hat{X} \wedge X | S). \end{aligned}$$

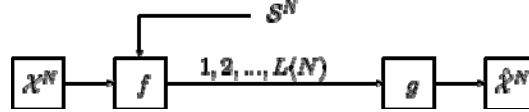
Case 3: State information on the receiver: $S_1 = \emptyset, S_2 = S$



Here $(S_1, X) = X$, and with distributions $P(x|s), Q_1(u|x)$ and $Q_2(\hat{x}|u, s)$, the upper bound of rate-reliability-distortion function is:

$$R'(E, \Delta, P^*) = \max_{P \in \alpha(E, P^*)} \min_{Q_1(u|x) Q_2(\hat{x}|u,s)} \left[I_{P,Q_1}(U \wedge X) - I_{P,Q_1}(U \wedge S) \right].$$

Case 4: State information on the sender: $S_1 = S, S_2 = \emptyset$



Here $I(U \wedge S_2) = 0$, with distributions $P(x|s), Q_1(u|x, s)$ and $Q_2(\hat{x}|u)$ $X \rightarrow U \rightarrow \hat{X}$ forms Markov chain and hence

$$\min_{Q_1(u|x,s)} I_{P,Q_1}(U \wedge X) \geq \min_{Q_2(\hat{x}|u)} I_{P,Q_2}(\hat{X} \wedge X), \text{ with equality iff } U = \hat{X}.$$

Since $U = \hat{X}$ and \hat{X} is independent of S we also have $I_{P,Q_1}(U \wedge S | X) = 0$. Taking into account these properties we get the upper bound of rate-reliability-distortion function:

$$\begin{aligned} R'(E, \Delta, P^*) &= \max_{P \in \alpha(E, P^*)} \min_{Q_1(u|x,s) Q_2(\hat{x}|u)} I_{P,Q_1}(U \wedge S, X) = \\ &= \max_{P \in \alpha(E, P^*)} \min_{Q_1(u|x,s) Q_2(\hat{x}|u)} \left[I_{P,Q_1}(U \wedge X) + I_{P,Q_1}(U \wedge S | X) \right] = \\ &= \max_{P \in \alpha(E, P^*)} \min_{Q_1(u|x,s) Q_2(\hat{x}|u)} I_{P,Q_1}(U \wedge X). \\ &= \max_{P \in \alpha(E, P^*)} \min_{Q_2(\hat{x}|x,s)} I_{P,Q_2}(\hat{X} \wedge X) = \\ &= \max_{P \in \alpha(E, P^*)} \min_{Q_2(\hat{x}|x)} I_{P,Q_2}(\hat{X} \wedge X). \end{aligned}$$

Proof of the Theorem

Let $J(P, Q_1) = \exp\{N[I_{P, Q_1}(X, S_1 \wedge U) + \varepsilon]\}$ for types P, Q_1 and $\varepsilon > 0$.

Lemma. For every type P and conditional type Q_1 there exists a collection of vectors

$$\{\mathbf{u}_j \in T_{P, Q_1}^N(U), j = 1, \dots, J(P, Q_1)\},$$

Such that for N large enough

$$T_P^N(X, S_1) \subset \bigcup_{j=1}^{J(P, Q_1)} T_{P, Q_1}^N(X, S_1 | \mathbf{u}_j).$$

The proof of the lemma is similar to the covering lemma from [Haroutunian et al, 2008].

The proof of the theorem is based on the construction of a code (f_N, g_N) based on the idea of *importance* of source vectors of messages of type P not farther from P^* (in sense of divergence). It is shown that (E, Δ) achievable rate of the constructed code satisfies (4).

The triple of sets for source messages and state information (available at encoder and decoder) of length N can be represented as a union of all disjoint types of vector triples:

$$X^N \times S_1^N \times S_2^N = \bigcup_{P \in P_N(X, S_1, S_2)} T_P^N(X, S_1, S_2).$$

For $\delta > 0$ and for N large enough the probability of appearance of vector triples of types beyond $\alpha(E + \delta, P^*)$ can be estimated in the following way:

$$\begin{aligned} P^{*N} \left(\bigcup_{P \notin \alpha(E + \delta, P^*)} T_P^N(X, S_1, S_2) \right) &= \sum_{P \notin \alpha(E + \delta, P^*)} P^{*N}(T_P^N(X, S_1, S_2)) \leq \\ &\leq (N + 1)^{|X| |S_1| |S_2|} \exp\{-N \min_{P \notin \alpha(E + \delta, P^*)} D(P \| P^*)\} \leq \\ &\exp\{|X| |S_1| |S_2| \log(N + 1) - N(E + \delta)\} \leq \exp\{-N(E + \delta / 2)\}. \end{aligned}$$

The first 2 inequalities follow from the definition of $\alpha(E, P)$ and type properties.

Encoding

For type P and conditional type Q_1 denote

$$C(P, Q_1, j) = T_{P, Q_1}^N(X, S_1 | \mathbf{u}_j) - \bigcup_{j' < j} T_{P, Q_1}^N(X, S_1 | \mathbf{u}_{j'}), j = 1, \dots, J(P, Q_1).$$

Step 1

Let us fix the type $P \in \alpha(E + \delta, P^*)$ and conditional types $(Q_1, Q_2) \in Q(P, \Delta)$.

From the definition of $C(P, Q_1, j)$ and from the lemma we have

$$\bigcup_{j=1}^{J(P, Q_1)} C(P, Q_1, j) = \bigcup_{j=1}^{J(P, Q_1)} T_{P, Q_1}^N(X, S_1 | \mathbf{u}_j) \supset T_P^N(X, S_1).$$

Step 2

Let $L(P, Q_1) = J(P, Q_1) / \exp\{N I_{P, Q_1}(U \wedge S_2)\}$. Randomly chosen indices of $\mathbf{u}_j \in T_{P, Q_1}^N(U)$ are uniformly distributed to $L(P, Q_1)$ bins. Denote by $B(i)$ the set of indices assigned to bin i .

Step 3

Considering \mathbf{x}, \mathbf{s}_1 encoder sends number i such that $(\mathbf{x}, \mathbf{s}_1) \in C(P, Q_1, j)$ and $j \in B(i)$.

DecodingStep 4

By receiving number i and \mathbf{s}_2 state information the decoder looks for \mathbf{u}_k such that $k \in B(i)$ and

$\mathbf{u}_k \in T_{P, Q_1}^N(U | \mathbf{s}_2)$. If there is such k , the decoder selects $\hat{\mathbf{x}}_i \in T_{P, Q_1}^N(\hat{X} | \mathbf{u}_k, \mathbf{s}_2)$. If there is no such k , or more than one k , decoder chooses preliminary fixed reconstruction vector $\hat{\mathbf{x}}_0$. If decoder receives i_0 , again $\hat{\mathbf{x}}_0$ is chosen.

The distortion between \mathbf{x} and $\hat{\mathbf{x}}_i (i = 1, \dots, L(P, Q_1))$ can be calculated in the following way:

$$d(\mathbf{x}, \hat{\mathbf{x}}_i) = N^{-1} \sum_{x, \hat{x}} d(x, \hat{x}) n(x, \hat{x} | \mathbf{x}, \hat{\mathbf{x}}_i) = \sum_{x, \hat{x}} d(x, \hat{x}) P Q(x, \hat{x}) = E_{P, Q_1, Q_2} d(X, \hat{X}) \leq \Delta.$$

So number of used vectors $\hat{\mathbf{x}}$ for fixed P and corresponding conditional types Q_1, Q_2 is equal

$$L(P, Q_1) = \exp \{ N [I_{P, Q_1}(X, S_1 \wedge U) - I_{P, Q_1}(U \wedge S_2) + \varepsilon] \}.$$

The number of vectors $\hat{\mathbf{x}}$ reconstructed under allowed distortion constraint for all P is not more than $P_N(X, S_1, S_2) L(P, Q_1)$. From the definition of (E, Δ) -achievable rate and from (3) we obtain:

$$\begin{aligned} & \frac{1}{N} \log [P_N(X, S_1, S_2) L(P, Q_1)] - \varepsilon = \\ & = I_{P, Q_1}(X, S_1 \wedge U) - I_{P, Q_1}(U \wedge S_2) + N^{-1} | X || S_1 || S_2 | \log(N+1) \leq \\ & \leq \max_{P \in \alpha(E, P^*)} \min_{Q_1, Q_2 \in Q(P, \Delta)} [I_{P, Q_1}(X, S_1 \wedge U) - I_{P, Q_1}(U \wedge S_2)]. \end{aligned}$$

The theorem is proved.

Bibliography

- [Cover-Chiang, 2002] T. M. Cover and M. Chiang, "Duality between channel capacity and rate distortion with two-sided state information", IEEE Transactions on Information Theory, vol. 48, no. 6, pp. 1629-1638, 2002.
- [Wyner-Ziv, 1976] A. D. Wyner and J. Ziv, "The rate-distortion function for source coding with side information at the decoder," IEEE Transactions on Information Theory, vol. IT-22, pp. 110, Jan. 1976.
- [Shannon, 1959] C. E. Shannon, "Coding theorems for a discrete source with a fidelity criterion," IRE National Convention Record, vol. 7, pp. 142-163, 1959.
- [Slepian-Wolf, 1973] D. Slepian and J. Wolf, "Noiseless coding of correlated information sources," IEEE Transactions on Information Theory, vol IT-19, pp 471-480, Jul. 1973.
- [Xiong-Liveris, 2004] Z. Xiong, A. Liveris, and S. Cheng, Distributed source coding for sensor networks, IEEE Signal Processing Magazine, vol. 21, no. 5, pp. 80-94, Sep. 2004.
- [Viswanathan-Berger, 1997] H. Viswanathan and T. Berger, "Sequential coding of correlated sources", Proceedings of IEEE International Symposium on Information Theory, Ulm, p. 272, Germany, June 1997.
- [Haroutunian-Mekoush, 1984] E. A. Haroutunian and B. Mekoush, "Estimates of optimal rates of codes with given error probability exponent for certain sources," (in Russian) 6th International Symposium on Information Theory, vol. 1, pp. 22-23, Tashkent, 1984.

- [Haroutunian et al, 1998] E. A. Haroutunian, A. N. Haroutunian, and A. R. Kazarian (Ghazaryan), "On rate-reliabilities-distortions function of source with many receivers," Proceedings of Joint Session 6th Prague Symposium Asymptotic Statistics and 13-th Prague Conference Information Theory, Statistical Decision Function Random Proceed, vol. 1, pp. 217-220, Prague, 1998.
- [Haroutunian-Maroutian, 1991] E. A. Haroutunian and R. S. Maroutian, "(E, Δ)-achievable rates for multiple descriptions of random varying source," Problemes of Control and Information Theory, vol. 20, no. 2, pp. 165-178, 1991.
- [Maroutian-1990] R. S. Maroutian, "Achievable rates for multiple descriptions with given exponent and distortion levels," (in Russian) Problems on Information Transmission, vol. 26, no. 1, pp. 83-89, 1990.
- [Meulen et al, 2000] E. C. van der Meulen, E. A. Haroutunian, A. N. Harutyunyan, and A. R. Ghazaryan, "On the rate-reliability-distortion and partial secrecy region of a one-stage branching communication system," Proceedings of IEEE International Symposium on Information Theory, Sorrento, Italy, p. 211, 2000.
- [Marton, 1974] K. Marton, "Error exponent for source coding with a fidelity criterion," IEEE Transactions on Information Theory, vol. 20, no. 2, pp. 197-199, 1974.
- [Haroutunian et al, 2008] E. A. Haroutunian, M. E. Haroutunian, and A. N. Harutyunyan, "Reliability criteria in information theory and in statistical hypothesis testing", Foundations and Trends on Communication and Information Theory, vol. 4, no 2-3, 2008.
- [Cover-Thomas, 1991] T. M. Cover and J. A. Thomas, "Elements of Information Theory", Wiley, New York, 1991.
- [Csisza' r- K' rner, 1981] I. Csisza' r and J. K' rner, "Information Theory: Coding Theorems for Discrete Memoryless Systems", Academic Press, New York, 1981.
- [Csisza' r-1998] I. Csisza' r, The method of types, IEEE Transactions on Information Theory, vol. 44, no. 6, pp. 2505-2523, 1998.

Authors' Information



Mariam Haroutunian – Professor of Physical-Mathematical Sciences; Scientific secretary, Institute for Informatics and Automation Problems, National Academy of Sciences, Armenia; e-mail: armar@jpia.sci.am

Major Fields of Scientific Research: Information theory, Probability theory and Mathematical statistics, Multimedia communications, multimedia security, Information-theoretic aspects of digital data-hiding, Steganography, Cryptography



Arthur Muradyan – Studying for PhD; Institute for Informatics and Automation Problems, National Academy of Sciences, Armenia; e-mail: mur_art@yahoo.com

Major Fields of Scientific Research: Information-theoretical investigation of sources and channels with compound states

USING THE GROUP MULTICHOICE DECISION SUPPORT SYSTEM FOR SOLVING SUSTAINABLE BUILDING PROBLEMS

Filip Andonov, Mariana Vassileva

Abstract. *The article discusses the implementation of decision support systems in the selection of the type of sustainable development building that best suits the criteria of all the participants in the decision making process – investors, clients and the local government.*

Keywords: *green building, decision support systems, sustainability*

Introduction

With the depletion of non-renewable sources of energy and growing environmental problems more and more technologies turn to innovative eco-friendly solutions. Green building tries to increase the effectiveness of the resources used – energy, water and materials and lower the negative impact on human health and the environment during the entire life-cycle of a building. In order to achieve this architects and engineers are searching for better location, design, structure, maintenance and disposal. The maximum energy efficiency and minimum impact on the environment and landscape can be achieved by creating eco-settlements designed using sustainable technologies.

Research shows that Bulgaria is among the countries with the lowest energy efficiency in buildings, especially in those, constructed before 1989 [1]. According to data from the Yale university, Bulgaria is on 56th position in the world by energy efficiency for 2008 [2]. Increasing this position will lead to:

1. reducing the negative impact of increasing energy prices for domestic users and increasing the comfort of the households;
2. creating new market opportunities for energy efficient facilities as well as new jobs;
3. achieving sustainable development [3].

The goal of this project is to develop a working model to support all participants/parties involved in the process of creating sustainable homes/eco-settlement in the decision making process regarding the most suitable technology for satisfying their needs and goals with the help of the decision support system Group Multichoice.

When developing sustainable building projects the choice is not limited between traditional technologies (reinforced concrete and bricks) with one technology for energy efficiency or one alternative building method (using wood for example). There are many existing options for alternative buildings, satisfying in various degrees the needs and criteria of the parties involved – investors, architects, prospective residents, municipal authorities and the general public. Thus such a project involves more than one decision maker (DM) and a number of perspectives to the problem. The participants have different criteria and usually contradicting goals.

Method applied. Description

In essence the problem is to find the most suitable building technology among a list of existing technologies, evaluating the alternatives by a set of contradicting characteristics with meaning for the model. Describing the problem in this way makes it a discrete multicriteria problem. There are several participants involved, therefore a group decision support method should be used. For the current example Group Multichoice system [4] is used. This system gives flexibility with regard to the problem solved, the degree of competence of the participants, the

methodology used, the type and number of the criteria, the size of the expert group and the way they express their preferences. The process of solving the problem can be divided into the following steps:

- determining the alternatives, criteria and their type;
- entering the values of the criteria with regard to the alternatives;
- selecting the method for solving the individual multicriteria problem and selecting the aggregating method;
- entering the DMs' preferences in the way the selected methods require;
- evaluating the result.

The last three steps are repeated until achieving a result, satisfying all participants or until the aggregating method stops.

Alternatives

For populating the set of alternatives several traditional and alternative construction methods were evaluated.

Table 1. Construction methods. Evaluated in the model

Number	Description	Code
1	Reinforced concrete and bricks	concrbrick
2	European type assembly house	european
3	Finnish type wooden house	finnish
4	Reinforced concrete underground house	underconcr
5	Wood and clay house	woodclay
6	Wood and straw house	woodstraw
7	Stone house	stone
8	Recycled tyres and compressed earth house	earthtires

Participants

The main participants or parties involved in the problem of choosing the technology for building new eco-settlements are basically three – prospective customers, investors and municipal authorities.

For the purposes of the current project a potential buyer and an architect (investor representative) were interviewed. For establishing the perspective of the municipal authorities the following documents were used: First National Action Plan for Energy Efficiency 2008 - 2010, and Directive 2006/32/EC of the European Parliament and the Council. The role of these documents was to identify the weights of the criteria, meaning their subjective assessment for the importance of every criterion, used in the decision making process.

Criteria

The criteria used to evaluate the alternatives are not the same for all participants. The values of the criteria for all alternatives are estimated on the basis of the average market prices for the last quarter of 2009 and consultations with an environmental expert.

Table 2. Criteria, used by prospective customers

No	Code	Description	Best value (min/max)	Type
1	price	Selling price	min	quantitative
2	maintain	Maintenance cost of the structure	min	quantitative
3	efficiency	Energy efficiency	max	quantitative
4	ecoimpact	Environmental impact	min	qualitative
5	landscape	Impact on landscape	min	qualitative
6	meteo	Susceptibility to weather influences	min	qualitative
8	comfort	Comfort	max	qualitative
9	light	Light	max	qualitative
10	humidity	Optimum humidity	max	qualitative
11	health	Impact on human health	max	qualitative

Table 3. Criteria, used by the investor

No	Code	Description	Best value (min/max)	Type
1	laborforce	Number of workers	min	quantitative
2	constrcost	Cost of construction	min	quantitative
3	price	Selling price	max	quantitative
4	edu	Costs of training staff	min	quantitative
5	materials	Costs of materials	min	quantitative
6	health	Impact on human health	max	qualitative
7	comfort	Comfort	max	qualitative
8	safety	Worker safety	min	qualitative

Table 4. Criteria, used by municipal authorities

No	Code	Description	Best value (min/max)	Type
1	ecoimpact	Environmental impact	min	qualitative
2	landscape	Impact on landscape	min	qualitative
3	laborforce	Number of workers	max	quantitative
4	roads	Impact on road infrastructure	min	qualitative
5	waste	Amount of building waste	min	quantitative
6	efficiency	Energy efficiency	max	qualitative
7	publicity	Public acclaim	max	qualitative

Results

After entering the data, the AHP method was selected for individual solving. The main reason to use AHP was that the participants do not have any experience with the applied methodology and the basic concept of this weighing method is relatively easy to understand, the method-specific data is easily extractable from the DMs and that makes them more confident in the result they obtain by applying it. Table 5 shows that despite the fact that the three participating sides have different criteria and preferences, they reach consensus on the first step of the interactive process and identify recycled tires and compressed earth structure as the winner. When using an interactive method for group decision support, the solving process can be interrupted on every step and the

currently preferred alternative is declared winner if all participants agree on that. In this case they have found a solution satisfactory for all of them and the process stops.

Table 5. Rankings of alternatives for participants on step 1

Position	Customer	Investor	Municipal authority
1	earthtires	earthtires	earthtires
2	woodclay	finnish	woodcley
3	finnish	woodstraw	woodstraw
4	woodstraw	european	underconcr
5	underconcr	woodcley	finnish
6	european	underconcr	european
7	stone	stone	stone
8	concrbrick	concrbrick	concrbrick

Aknowledgement

This research is partially granted by Bulgarian National Science Fund, Ministry of Education and Sciences in the frame of the project "Automated Metadata Extraction for e-documents Specifications and Standards", contract N D002(TK)-308/ 19.12.2008.

Conclusion

In conclusion it should be noted that the methodology used is applicable for introducing and popularizing modern sustainable technologies in construction, unjustifiably neglected by the general public in Bulgaria, despite their proven qualities and benefits due to the fact that there is no sufficient information about them or to lack of social prestige associated with these.

Bibliography

1. Naniova, C. "[The Demonstration Project for the Renovation of Multifamily Buildings: a joint initiative of the Bulgarian Ministry of Regional Development and Public Works and UNDP Bulgaria](#)" Sofia, Bulgaria. UNECE First Workshop on Energy Efficiency in Housing 2009
2. <http://epi.yale.edu/Bulgaria>
3. [National program for energy efficiency to 2015, www.mee.government.bg](#)
4. Andonov F., 2009, INTERACTIVE METHODS FOR GROUP DECISION MAKING, KDS 2009
5. First National Action Plan for Energy Efficiency 2008 - 2010, and Directive 2006/32/EC
www.strategy.bg/FileHandler.ashx?fileId=482
6. Directive 2006/32/EO of the European Parliament and Council,
www.seea.government.bg/documents/EE_Kraino_Potreblenie.doc

Authors' Information

Filip Andonov – ass. prof. [NBU, 02/8110610, fandonov@nbu.bg](mailto:NBU,02/8110610,fandonov@nbu.bg)

Major Fields of Scientific Research: Decision support systems, group decision support, multi-criteria analysis

Mariana Vassileva – assoc. Prof., phd, IIT-BAS, mvassileva@iinf.bas.bg

Major Fields of Scientific Research: Decision support systems, multi-criteria optimisation, multi-criteria analysis

INFLUENCE ANALYSIS OF INFORMATION TECHNOLOGIES ON PROGRESS IN CONTROL SYSTEMS FOR COMPLEX OBJECTS

Boris Sokolov, Rafael Yusupov, Michael Okhtilev, Oleg Maydanovich

Abstract: *Current status and perspectives of an interdisciplinary knowledge domain including informatics, computer science, control theory, and IT applications were analyzed. Scientific-and-methodological and applied problems of IT integration with existing and future industrial and socio-economical structures were stated.*

Keywords: *computer science, informatics, cybernetics, control theory, information technologies, control systems, information systems, industrial applications*

ACM Classification Keywords: *A.0 General Literature - Conference proceedings*

Introduction

Informatisation and information society (as the final aim of informatisation) are characterized by active development and mass introduction of information technologies into all areas of human activity, namely [Yusupov, Zabolotskii, 2000]: social life, material production, power engineering, health protection, education, science, culture, business, transport, communication, military science and so on. By now, some attempts were made in order to evaluate the role and influence of IT on the progress (enhancing effectiveness) of the above areas.

In the present paper we are trying to draw attention of specialists to the influence of **information technologies** (IT) on the progress in such an area as processes and control systems for objects of various kinds.

The main aspects of IT influence on progress in control systems for complex objects

We note that for realizing control process it is necessary to have information on control goals and objectives, state of the plant and environment. This information is formed on the basis of processes of data measurement, transmission and handling by means of sensors (receptors, detectors), communication channel and computational facilities, which are fundamental elements and subsystems of any control system. Currently general topics related to collection, processing, representation, transmission and protection of information are studied in **informatics**. Results of these investigations are realized as IT. This term means a family of methods for realizing informational processes in various fields of human activity aimed at manufacture of informational product, including those in control systems. It is obvious that operational effectiveness of control systems depends on the progress in informatics and information technologies. Constructive identification and investigation of the above dependence is an actual scientific-and-technical problem, in the framework of which leaders of modern large-scale enterprises are trying to get an answer to the question: "**in which of perspective IT's we should invest and why?**".

Business and state are ready to pay for exactly such amount of information resources that they really need for information support of management. Moreover, they proceed from such classical efficiency indices, used today in the market of computer services, as *return on investment (ROI)*, *total cost of ownership (TCO)*, and *quality of service (QoS)*. Superfluous information resources and redundant IT are frozen investments and resources (moreover, they are lost resources, with account for fast obsolescence of hardware and software facilities and equipment). Insufficient informational resources mean a loss of profit ([Perminov, 2007], [Seletkov, Dneprovskaya, 2006], [Sokolov, Yusupov, 2008a], [White, 2004]).

Discussing modern control processes and systems, we hereinafter will separate **two classes of control systems** for objects, namely: **automatic** and **automated** control systems for the corresponding objects (or groups of objects). It should be noted right away that information technologies, which are realized by means of the corresponding hardware-and-software facilities and computers, played and still play the determining (central) part in the aforementioned control systems. Moreover, historically, the tightest integration of these technologies and facilities with control systems is shown by the fact that computers were called **cybernetic machines** during the first years of their existence ([Gerasimenko, 1993], [Mertens, 2007], [Sovetov 2006], [Sokolov, Yusupov, 2008b]).

The above features objectively lead to necessity of development of **automated control systems**, by which we mean “man-machine” systems that ensure effective operation of the corresponding objects. In these systems, information gathering and processing, which is necessary for realization of control functions, are performed with the use of automation facilities and computer engineering ([Sovetov, 2006], [Starodubov, 2006]). According to realized control functions, kinds of controlled objects, and used generations of information technologies and facilities, there are various types of automated control systems (ACS) for complex objects (CP): supervisory control and data acquisition systems (SCADA); manufacturing execution systems (MES); ACS for flexible manufacturing systems (FMS); systems of computer-aided design (CAD); automated systems for scientific investigations (ASSI); integrated ACS; AS for organizational control (ASOC); branch-wise ACS (BACS); corporative ACS (CACs); enterprise resource planning system (ERP); storage-and-retrieval system (SRS); storage-and-advisory systems (SAS); management-information system (MIS).

Previous investigations demonstrated that both in Russia (in USSR, in the period 1960-1980) and abroad the most widespread and economically effective were automated systems at enterprises that manufactured various kinds of products ([Mertens, 2007], [Sovetov, 2006]). As a rule, such systems incorporate SCADA, MES, and ERP. An aggregate of the aforementioned systems form a computer-integrated manufacturing (CIM).

Directions of evolutionary progress in ACS CP, as well as control systems, were always determined by tendencies of development of related information and communication technologies and systems, which form the material basis for realization of existing and perspective technologies of automated control and have, by their nature, specifically informational character, as was already mentioned.

Fig. 1 represents, in a generalized form, evolution of basic information technologies that were used as the basis of the corresponding ACS for industrial enterprises discussed above. Let us take a quick look at the influence of these information technologies on the progress of this class of automated systems. For Russian ERP-systems the following three stages of their evolutionary development are usually separated (see Fig. 2).

Stage I (from the mid-1960's to the end of the 1970's).

At this stage of development of ERP-systems based on the computers of the IInd generation (M-20, M-220M, M-222, Minsk 22, Minsk 32, SM-4) at enterprises manufacturing goods and services in various specific fields, only some functions of dataware and production activity control were automated. Among such functions were, first of all, gathering, processing, storage, representation and analysis of some kinds of industrial and economic information. The first attempts were made to integrate various information technologies in the frameworks of the corresponding SCADA and MES.

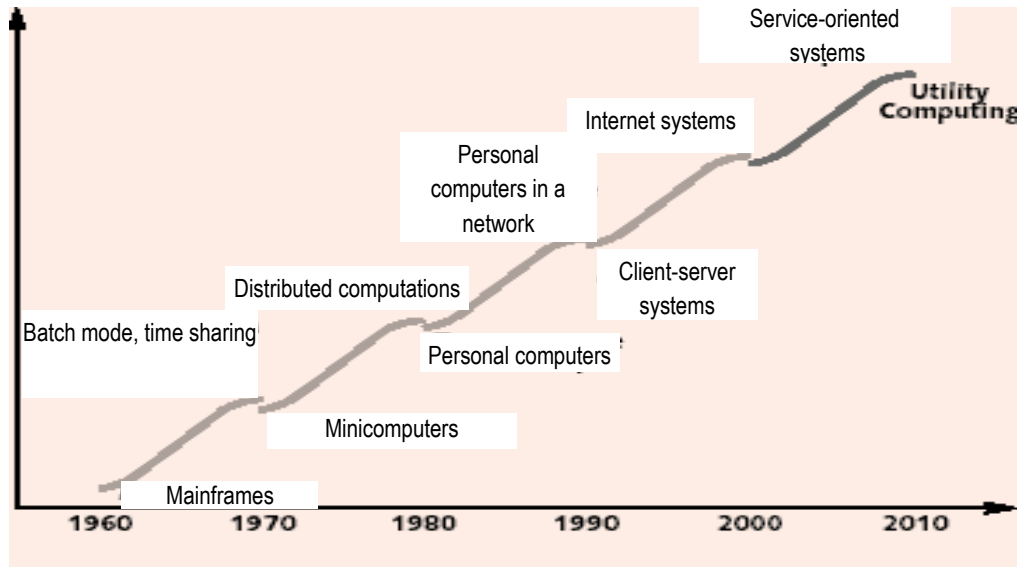


Fig. 1. Evolution of basic information technologies (Chernyak, 2003a).

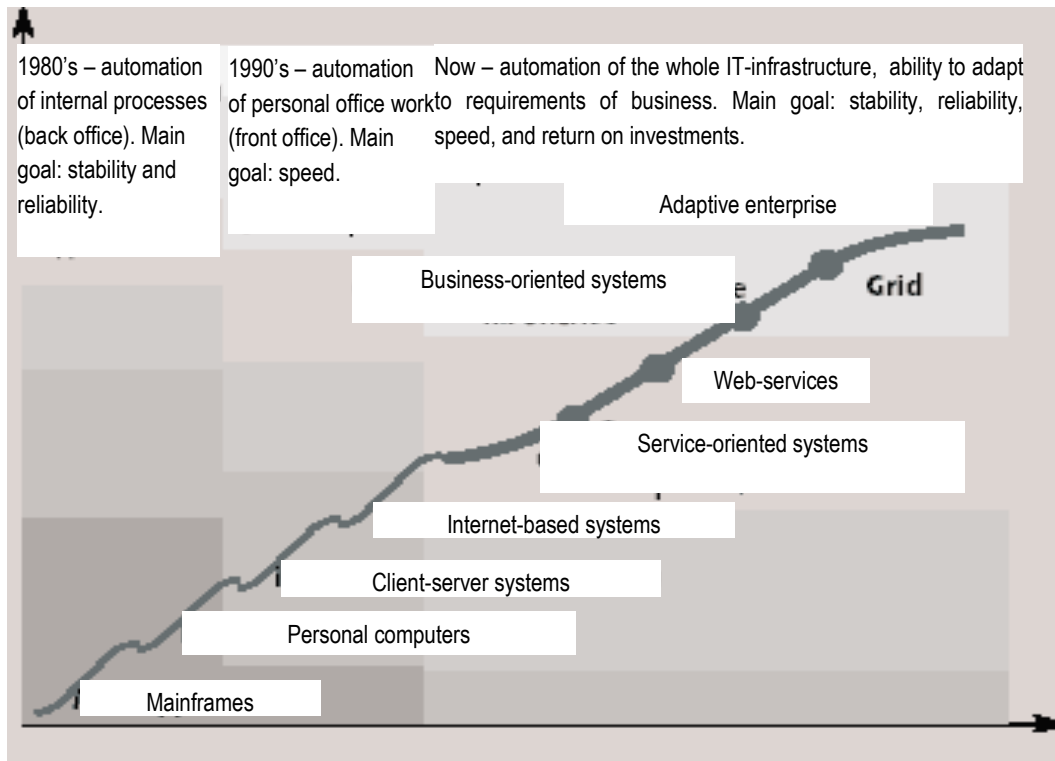


Fig. 2. Stages of evolutionary progress in automated and information systems (Chernyak, 2003a).

Stage II (from 1980 to the beginning of 1990's).

In this period, after learning the lessons of Stage I of automation and informatisation of industrial activity, works were carried out for unification of processes of using IT on the basis of wide introduction of standard automation modules (SAM). As the main element of technical basis for realization of concept of typification and unification of complex automation facilities (CAF) computers of series "Ryad" (ES) were chosen. On the basis of this concept more than 6000 ACS of various classes were created in the USSR to the beginning of 1990's [Sovetov, 2006].

Stage III (from 1990's to present stage of progress).

At this stage, evolutionary complexation and integration of information resources and technologies, proposed by Russian and foreign companies, take place on the basis of wide distribution of personal computers and modernization of previously used medium- and large-scale computers (mainframes). Information technologies that supports operation of distributed data banks and databases, as well as protocols of such local and global telecommunication networks as Intranet and Internet, provided for material basis for a new integration level for various classes of ERP-systems.

It should be noted that disintegration of the USSR in 1992 caused negative consequences for Russia, because financial support and amount of works on creation of automated systems (AS) of various classes have been greatly reduced. Instead, not very successful copying and adaptation of foreign counterparts of Russian enterprise-oriented ACS (systems of classes ERP, MRP, MRPII, and SCADA) started. One of the main problems in introduction of these systems was that Russian legislative and juridical base in accounting, financial and management spheres differs from the respective foreign base. Moreover, there was a lack of necessary technical and technological standards and information databases that determine specific features of manufacturing process. Due to these reasons, integrated automation appeared to be impossible.

Nevertheless, as time went by (in the end of 1990's) domestic replicable ERP-systems, called corporative information systems (CIS), began to appear in Russian Federation. Such domestic automated system as "Galaktika" and "Parus" can be mentioned among them. But this automation performed at industrial enterprises "from above" (at the strategic level of control) without corresponding complex automation of control processes at lower levels of control (where, in essence, wealth and surplus value are created) gave no planned effect and did not justify productive investments.

Today, owing to saturation of world market by all kind of products as well as due to general accessibility of high technologies (including those in the infosphere), the **time factor** is brought to the forefront of competitive activity. Now only those have a chance to win this competitive struggle who can successfully synchronize business-processes and manufacturing (ERP- and MES-systems) in real time (RT); develop and promote a new product (CAD/CAM/PDM-systems) in the market; have a flexible, effective and highly automated technology of control of logistic processes, which provide for decreasing cycles of delivery and off-take (Supply Chain Management, SCM); reduce order processing time (customer relationship management, CRM); ensure monitoring of resource spending in RT; realize operational control scheduling in RT (automated systems of operational supervising control, ASOSC); reduce time of return on investments (ROI-systems); reduce time needed for analysis and decision-making (OLAP-systems); provide for effective control of manufacturing cooperation in RT (e-manufacturing, co-manufacturing, m-business).

At the modern level of progress in enterprise ACS, an important role in practical realization of the above requirements must be played by **manufacturing execution systems** (MES) (they are called ACS of manufacturing processes in Russia) **and mobile (wireless) information technologies**. Now we briefly describe how these information technologies have effect on the processes of automated enterprise control, and show the main problems in distribution of the above technologies.

Today, speaking about successes in automation of complex organizational-and-technical complexes (COTC), one should mention, first of all, ERP-systems (in Russia they were previously called ACS of enterprise). The share of its successful applications in financial, administrative and trading organizations is much higher than in industry ([Len'shikov, Kumilov, 2002], [Sovetov, 2006]). At the administrative level, ERP-systems take into account any financial operation and any document, while no such a detailed supervision exists at the manufacturing levels (level of SCADA, MES). But the analysis shows that the manufacturing levels are the birthplace of the surplus value, the place of fundamental spending and main sources of economy; these levels provide for manufacturing plan and required quality of production; many factors determining efficiency and profitability of the enterprise as a whole work here. In these conditions, such a principal unit as manufacture drops out of the loop of automated control and enterprise management.

Thus being the case, today in most of implemented projects connected with creation of integrated automated control systems for industrial enterprises there is a whole strata of functions that have been covered neither by the ERP-systems nor by SCADA-systems.

Analysis of fig. 3 shows that ERP-systems do not support the level of operational control of production, being restricted by strategic scheduling only; they are not interconnected from the informational and logical viewpoint, and are not synchronized with the goals of production control in RT [Len'shikov, Kumilov, 2002]. In this layer of operational production control, which is not encompassed by information technologies, there exists a whole class of manufacturing processes that are vitally important for the enterprise, create surplus value and have an influence on its profitability on the whole.

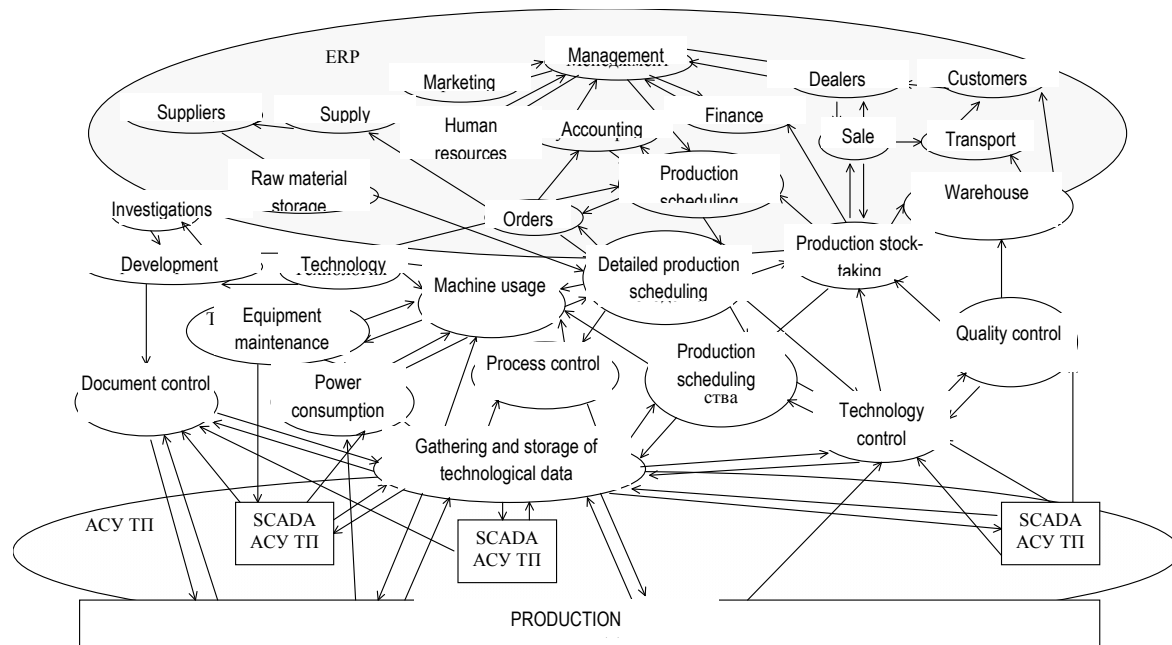


Fig. 3. Functional gap between ERP and SCADA (Len'shikov, Kumilov, 2002).

Up to now this class of processes is supported by manufacturing execution systems (MES) directed to informatisation of operational scheduling and production control, optimization of manufacturing processes and production resources, control and scheduling of fulfillment of a plan with minimal cost. Thus, speaking about integrated automated systems of enterprise control under the current conditions it is expedient to separate out the following three interconnected levels of control, namely: SCADA, MES and ERP. Each of them performs

(realizes) its own control technology and is characterized by its own intensity of information circulation, time scale, set of goals and relevant tasks.

Today enterprises often face such phenomena as return of production, delay in order fulfillment by partners, order cancellation due to low quality of raw materials, too large period needed for analyzing the cause of defect and so on. To get over these difficulties, it is necessary to provide for timely and reliable handling of information. This is possible, in its turn, only if data gathering and handling are performed immediately at the very moment of the associated event and as close to its source as possible. By the way, it should be noted that errors are most often made during simple and routine operations of data input in integrated ERP-systems (IERP) ([Len'shikov, Kumilov, 2002], [Mertens, 2007]).

In general, analysis of modern tendencies in the progress of information technologies and systems (IT and IS) shows that all leading foreign and domestic companies specializing in the field built and still do build corporative information infrastructures (including ERP-systems) only by vertical principle; they are guided by particular criteria and hardly coordinate their own conceptions with requirements of business. As a result of realization of the above tendencies, traditional approaches to automation of business-processes are today in a pre-crisis if not even in a crisis state. Moreover, difficulties of controlling modern corporative information systems go beyond the scope of administration of software environments. Necessity of integrating several heterogeneous environments into all-embracing corporative computer system and desire to overstep the limits of the company form a new complexity level. For example, in order to cope with diversity of external and internal queries to the corresponding business applications, modern IT-companies are forced to distribute their solutions in business-systems over hundreds and thousands servers. In these conditions, manual control (administration) of this diversity of information resources becomes impossible due to both organizational and financial reasons. By data of foreign analysts, only about 30% of companies' IT-budgets can be directed on development of IT-technologies, while the rest is spent on support of existing IT-technologies. If nothing will be done, this ratio reaches 5:95 to 2010 [Chernyak, 2004].

Among other shortcomings (and associated problems) brought to light up to date in the process of developing computer integrated manufacturing (CIM) created and exploited on the basis of existing IT, the following ones can be enumerated:

1. in a number of cases there is *no comprehensive analysis* of existing (non-automated) technology of gathering and processing information and decision-making; no propositions and recommendations are elaborated on its perfection (non-productive labor is automated) and transition to new intellectual information technologies; required degree of automation is not justified for any specific organization;
2. many AS (first of all, ACS for complex objects) are mostly information systems, where processes associated with decision-making are not automated, or the part of automation of these processes is negligible as compared with automation of information gathering and processing; possibilities of methods and algorithms of complex simulation and multi-objective choice are hardly used for decision justification and management;
3. there exists considerable inconsistency in goal orientation as well as in technical, mathematical, software and organizational facilities of AS on different control levels, and AS that are at the same level in the framework of a fixed hierarchical structure of the corresponding organization;
4. AS still do not provide for required orientation of each specific organization onto optimization of using available resources and growth of its efficiency on the whole; this thesis is supported by the fact that optimization problems constitute only several percent among all problems solved in AS;
5. in many AS there is no necessary software-and-mathematical facilities for performing system analysis for organization as a whole, operation of AS itself, and control of its operation quality;

6. quality of dataware still has not reached the required level; for instance, necessary filtering of information, its selection according to management level and representation in a compact form are not ensured;
7. development of software and technical facilities for “man-machine” communication, dialog communication procedures (creation of intellectual interfaces) is far behind the practice;
8. creation of AS is not interconnected appropriately with evolution problems and allotting the system with high flexibility and adaptation to variations of environment.

What are the reasons for existence of the above mentioned drawbacks (problems) related to creation and development of AS?

One of the main reasons, which have a **methodological character**, is that requirements of system approach to design of complex technical-and-organizational complexes are often ignored while developing AS. This is apparent, in particular, from the fact that automation is performed for some separate stages of information gathering and processing, and only some computational problems are solved on computers, without investigating the automation problem for control processes on the whole. In other words, there is no complex automation of the processes under consideration. As has been demonstrated by practice, automation should be applied only to well-known and fairly stable processes and technologies, for which constructive formal descriptive facilities (models), methods, algorithms and methods of solving applied problems have been developed.

Thus being the case, problems of creating and developing AS are, first of all, **model-and-algorithmic** and **informational problems** that require development of a fundamental theoretical base for their solution.

Speaking about **technical-and-technological reasons**, it should be emphasized, first of all, that traditional technology of creating AS calls for using a lot of specialists (designers, programmers, database administrators, managers, technicians and so on), which manually form the appearance of a future system, using traditional paper technology. For such a technology, developers of hardware-and-software facilities permanently encountered and still do encounter a number of hardly solvable problems, namely: problem of inadequacy of structuring of AS; problem of misadjustment of structural parts of AS; problem of inconsistency, ambiguity, redundancy (or incompleteness) of design documentation.

All of the above-listed problems stem from complexity of automated systems as objects for analysis and design. For a long time in Russia, due to bureaucratic barriers and backwardness in the field of microelectronics, available facilities for automated information processing and control have low level of unification, and were used for solving a small number of problems; this was also one of the reasons for failures in creating AS.

Among **organizational reasons**, it should be noticed once again that, as a rule, many enterprise managers reassigned all questions on coordination and control of work related to creation of AS to other functionaries, which have not necessary authority (*principle of the first manager*, which is one of the basic principles in developing AS, was not honored). Moreover, some conservatism often took place in organizations, owing to which structure and functions of AS were fit, deliberately or unconsciously, to existing technological-and-organizational structure (in other words, non-productive labor was automated).

Summarizing the aforesaid, it should be stated that modern stage of progress in science and engineering is characterized by fairly high level of development of hardware-and-software facilities for information gathering, transmission and processing, which are incorporated into any AS; these facilities are permanently modified and their technical and economical characteristics are improved.

At the same time, today, when national economics are transforming to global market economy that has dynamic network nature, increasingly more foreign and domestic specialists began to understand importance of a complex approach to automation of operation of enterprises and organizations in order to overcome the above-listed

problems in modern IT-industry ([Dmitrov, 2006], [Zatsarinnyi, Ionenkov, 2007], [Mertens, 2007], [Sovetov, 2006], [Building an adaptive enterprise, 2003], [HP, 2001], [HP, 2003], [IBM, 2004]). With this aim in view, a principally new methodology of creation and development of automated and information systems in XXI century is to be proposed.

As a basic concept for development of information and telecommunication area, leading manufacturers of computer technologies and systems propose to use the concept of “natural”, “organic” information technologies (Organic IT), which provide for permanent dynamic balance between business queries for services (business applications) and information resources of the corresponding automated systems.

Introducing the notion of *Organic IT* into terminology of modern computer science, analysts of *Forrester Research* ([Chernyak, 2004], [Chernyak, 2003a], [HP, 2001], [HP, 2003], [IBM, 2004]) would like to emphasize necessity of more organic, natural, indirect use of IT in the interests of business applications in solving the following three groups of problems:

- effective *utilization* of information resources; here the proposed IT has to admit scaling the given resources “up” and “down” without service outage; regarding reliability, modern automated systems must be similar to modern power and telephone networks;
- *integration*: Organic IT are to combine dissimilar technologies in an easy and simple way;
- *manageability*: Organic IT are to support processes of automatic installation, load balancing, diagnostics and repair, leaving possibility for an operator to intervene in the process only in worst-case situation.

As specific examples of transition to “natural” computer systems in large-scale corporations working in the field of information services, the following technologies can be mentioned: Dynamic Computing (Dell); Adaptive Infrastructure or Adaptive Enterprise (Hewlett-Packard); Computing On Demand (IBM); Autonomous Computing; Dynamic Systems (Microsoft); N1 technology (Sun Microsystems).

As applied to industrial enterprises and appropriate ERP-systems, realization of the Organic IT conception means a transition to a principally **new (fourth) stage of creation and development of the automated systems** under consideration, which are to possess the following basic properties ([Dmitrov, 2006], [Kozlovskii, 1985], [Mertens, 2007], [Rostovtsev, 1992]): self-configuration, self-perfection, self-optimization, self-diagnostics and self-repair, self-preservation, “self-consciousness” and proactivity. In other words, under discussion are flexible, adaptive, self-organizing automated industrial enterprises. By these terms we mean complex geographically distributed automated organizational-and-technical complexes that provide for manufacture of products under the conditions of operatively changed market demand and operate (because of high degree of automation of production and management processes) with a limited personal staff. These enterprises and the corresponding conceptions (methodologies) of their creation and usage are called (in Russia) flexible manufacturing automated factory (FMAF), the relevant foreign terms are Integrated Computer Aided Manufacturing (ICAM) in the USA and European Strategic Planning for Research in Information Theory (ESPRIT) in European Community ([Chernyak, 2004], [Chernyak, 2003a], [Building an adaptive enterprise, 2003], [HP, 2001], [HP, 2003], [IBM, 2004]).

Such being the case, the main goal of industrial enterprises of the next generation and the corresponding automated control systems for them is to greatly increase productivity and quality of manufactured products on the basis of implementation of adaptive automated control technologies, which provide for high flexibility and operative reaction on changing market requirements.

Today one should note prevailing role of service oriented architectures (SOA) and those based on business-application services. These technologies are oriented on permanent support of information infrastructure of communication, and on coordination in a distributed decision-making environment, which is typical for adaptive and self-organizing automatic and automated systems of new generation.

The **stage of visualization of information resources** is the most important stage of the evolution to adaptive enterprises and the corresponding automated systems. By means of virtualization, logical functions of servers, storage devices and other system components become separated from their physical functions (processors, RAM, disk drives, input-output systems, switchboards and so on); further on, they are transferred to general pool of resources, which are convenient to control in automatic and/or automated mode of operation. By now, for instance, HP proposed and realized several variants and directions of virtualization for information resources, namely ([HP, 2001], [HP, 2003], [IBM, 2004]): server virtualization, virtualization of telecommunication network, virtualization of data storage systems, virtualization of applications.

According to the idea of HP, utility data centers (UDC), created by the company, builds the basis for a transition to Darwin reference architecture. This architecture ([Chernyak, 2003b], [HP, 2001], [HP, 2003], [IBM, 2004]) is a tool for creating and developing hierarchical-and-network information structures, which makes it possible to adapt IS and IT to changing goals and objectives of business-systems. It is proposed to establish three control levels, namely ([Chernyak, 2004], [Chernyak, 2003a], [HP, 2001], [HP, 2003], [IBM, 2004]): *component level* that controls data processing center, *service level* that controls aggregated components and delivers applied servers, and *business level* that controls users and permissions to access applications.

On the whole, future Darwin reference architecture is to provide for permanent balance between business queries for services and infrastructure resources delivering these services in FMAF.

Fig. 4 presents information technologies that, together with SOA, provide for realization of the adaptive enterprise concept ([Gorodetskii, 2000], [Dmitrov, 2006], [Zatsarinnyi, Ionenkov, 2007], [Chernyak, 2003a]). On the whole, by expert estimates, creation and development of flexible adaptive integrated ACS for controlling enterprise make it possible to reach the following goals ([Kozlovskii, 1985], [Mertens, 2007], [Sovetov, 2006]): enhance productivity of labor at future industrial enterprises 8-10 times; increase production output per unit area 1.5-2 times; decrease the period of production cycle 2-10 times; increase machine utilization by 30-40%.

In conclusion of the report we shall consider, as an example of the above ideas, the influence of modern space-based information technologies on operating benefits of ACS for space-based facilities (SF). Hereinafter, by space-based information technology (SIT) we mean information technologies that provide for gathering, storage, transmission (importation), representation, processing and analysis of data at various stages of life-cycle of SF. The basic features of SIT are determined by:

- essential influence of numerous factors of space and related specific space-and-time, technical and technological restrictions that prevent direct usage of standard information-and-telecommunication methods and tools for effective solution of fundamental and applied tasks of cosmonautics;
- multilevel and cyclic nature of solving targeting and maintenance problems by SF;
- complex integration of space-based information technologies with technologies of automated (automatic) control of SF in the frameworks of the corresponding automated systems (AS).

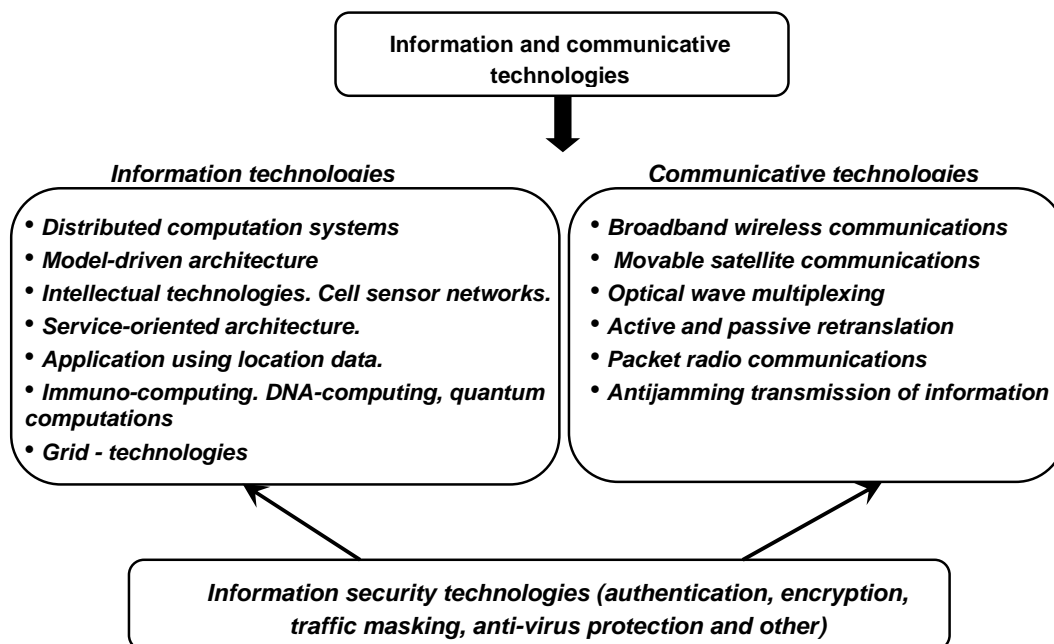


Fig. 4 Perspective information technologies (Zatsarinnyi, Ionenkov, 2007).

The influence of SIT on the essence of modern ACS for SF is apparent in the framework of the following directions [Military-space activity..., 2005]:

- Increase of globality and continuity of control of SF on the basis of creating network structures for information interchange with spacecrafts (SC) of various classes.
- Distribution of methods of situation packet telemetry, which makes it possible to form flexible telemetry programs directly on spacecraft-board.
- Substantial reduction of amount of measurements of current navigation parameters (MCNP) performed by means of ground control complex (GCC) on the basis of comprehensive utilization of navigation and time-and-frequency field produced by domestic "GLONASS" space navigation system and foreign systems.
- Creation of SF of new generation (modernization of existing SF) in order to enhance the level of their unification and multi-functionality that provides for required conditions for adaptation and self-organization of the processes of automated (automatic) control for SIT in various conditions of dynamically changed environment.
- Decentralization (space-time distribution) of the processes of gathering, representation, decision-making, storage and access to information circulating inside the control loops of SIT; it is realized by means of creating integrated distributed databases and knowledge bases with necessary level of information security and information safety.

Speaking about intrinsic efficiency indices of SIT, we restrict ourselves by a single example, which is a result of joint works performed by SPIIRAS and a special design bureau during the period from 2003 to 2008 [Okhtilev et al, 2006]. As a result of comprehensive investigations performed in the specified period, methodological and methodic basis was developed and realized for solving problems of structural and functional synthesis of intellectual information technologies (IIT) and systems for monitoring multi-structure macro-states (MS) of complex technical objects (CTP). These methods are based on their poly-model multiobjective description

obtained in the frameworks of the theory of underdetermined calculations and control of structural dynamics ([Kalinin, Sokolov, 2005], [Okhtilev et al., 2006], [Sokolov, Yusupov, 2008a]).

The proposed IIT makes it possible for a user, who is not a programmer, to perform, using a professionally-oriented language in interactive or automatic mode, intellectual processing of multi-type data and knowledge on the state of both CTP and MS using incorrect, incomplete and contradictory measurement information. This IIT is oriented to development of applications for objects that are especially difficult to control under the conditions of arising emergency and extraordinary situations and time trouble (including power consumptions, cosmonautics, petrochemical industry and transportation).

Preliminary analysis have shown that realization of the developed intellectual information technology of state monitoring (SM) for space-based facilities makes it possible to obtain the following effect for dataware systems in ACS SF [Okhtilev et al, 2006]:

- reduce time expenses spent for performing SM providing for obtaining results in real-time according to the transfer rate of measurement information (MI) with sufficient level of authenticity;
- enhance flexibility, reliability and information capacity of software used for SM; on the whole, this improves effectiveness of usage of the corresponding ACS SF;
- formalize (by means of a knowledge representation language and appropriate systems of initial data preparation) up to 90-95% of data about performing SM;
- reduce at least twice the time needed for preparation of initial data and knowledge for SM regarding SF that are being taken to information maintenance;
- reduce the period of the cycle of development and realization of information system for SM from 10 to 15 times;
- provide for saving resources spent for development of special software (SS) for the corresponding ACS SF that incorporate the SM system at hand;
- exclude up to 60-80% of all errors encountered in developing software for SM owing to the usage of verification tools for SS;
- exclude up to 80-95% of corrupted data received from CTP.

Thus, summarizing the aforesaid, it should be noted, first of all, that one of the main tendencies of development of information technologies and systems (IT and IS) in XXI century will be connected, in our opinion, with solution of the problem of comprehensive integration of these technologies and systems with existing and future industrial and socio-economical structures and the corresponding control systems. To solve successfully this interdisciplinary problem, it is necessary to solve a number of scientific-and-methodological and applied problems.

Conclusion

Today influence of computer science and IT on the progress in control theory and systems has a global nature. Specialists note that recently the second turn of convergence between general control theory (cybernetics) and computer science take place; we are observing revolutionary progress in control systems caused by the influence of IT.

In this connection, the problem under discussion deals with formation of a new interdisciplinary direction at the turn of cybernetics, telecommunication theory and general systems theory. This direction was conditionally called *neocybernetics* by the authors of the report ([Sokolov, Yusupov, 2008b], [Yusupov, 2005]).

Acknowledgements

Interdisciplinary investigations on the topics under consideration were conducted with financial support of RFFI (grants 09-07-00066, 10-07-00311, 08-08-00403), ONIT RAS (project № 2.3/03).

Bibliography

- [Bir, 1949] Bir T. Cybernetics and production control [in Russian]. Moscow: Fizmatlit, 1963, p. 22.
- [Building an adaptive enterprise, 2003] Building an adaptive enterprise. Linking business and IT, October 2003, Hewlett-Packard.
- [Chernyak, 2004] Chernyak L. Adaptability and adaptation [in Russian] Open Systems, 2004, No. 9, pp. 30-35.
- [Chernyak, 2003a] Chernyak L. SOA: a step beyond the horizon [in Russian] Open Systems, 2003, No. 9, pp. 34-40.
- [Chernyak, 2003b] Chernyak L. From adaptive infrastructure to adaptive enterprise [in Russian] Open Systems, 2003, No. 10, pp. 32-39.
- [Dmitrov, 2006] Dmitrov A. Service-oriented architecture in modern business-models [in Russian]. Moscow, 2006, 224 p.
- [Encyclopedia of Cybernetics, 1974] Encyclopedia of Cybernetics [in Russian]. Kiev: Editorial Board of USE. 1974, 406 p.
- [Gerasimenko, 1993] Gerasimenko V.A. Computer science and integration in engineering, science and cognition Foreign radio electronics, No. 5 1993, pp. 22-42.
- [Gorodetskii et al, 2000] Gorodetskii V.I., Kotenko I.V., and Karsaev O.V. Intellectual agents for detecting attacks in computer networks Proceedings of Conference on Artificial Intelligence [in Russian]. Moscow: FML Publishers, 2000. p. 23-35.
- [Wong, Sycara, 2000] H. Wong, K. Sycara, A Taxonomy of Middle Agents for the Internet, Proc. 4th Int. Conf. Multiagent Systems, IEEE CS Press, 2000.
- [HP, 2001] HP Utility Data Center. Technical White paper, October, 2001.
- [HP, 2003] HP virtualization. Computing without boundaries or constraints. Enabling an adaptive enterprise, Hewlett-Packard, 2003.
- [IBM, 2004] IBM, Autonomic Computing: IBM's Perspective on the State of Information Technology, 2004.
- [Information security..., 2006] Information security in systems of organizational management. Theoretical foundations. Vol. 1. Eds. N.A. Kuznetsov, V.V. Kul'ba. Moscow: Nauka Publishers, 2006.
- [Kalinin, Sokolov, 1995] Kalinin V.N., Sokolov B.V. Multi-model description of control processes for space facilities [in Russian] Journal of Computer and Systems Sciences International. No.1, 1995, pp. 149–156.
- [Klyuchko] Klyuchko N.V. On the notion of “control of information” Collected papers “Control of information flows” [in Russian]. ISA RAS. Moscow: URSS Publishers, 2002, pp. 189-200.
- [Kozlovskii] Kozlovskii V.A. Efficiency of versatile robotized enterprises [in Russian] / V.A. Kozlovskii, E.A. Kozlovskaya, V.M. Makarov. Leningrad: Mashinostroenie Publishers, Leningrad branch, 1985, 224 pp.
- [Kul'ba et al, 1999] Kul'ba V.V., Malyutin V.D., Shubin A.N., Vus M.A. Introduction into information agency [in Russian]. St.-Petersburg: SPbGU Publishers, 1999.
- [Len'shikov, Kuminov 2002] Len'shikov V.N., Kuminov V.V. Manufacturing execution systems (MES) as a way to effective enterprise [in Russian] World of Computer Automatics, no. 1-2, 2002, pp. 53-59.
- [Mamikonov, 1975] Mamikonov A.G. Control and information [in Russian]. Moscow: Nauka Publishers, 1975.
- [Mertens, 2007] Mertens P. Integrated information processing. Operating systems in industry. Textbook [in Russian] // Translated from German by M.A. Kostrova. Moscow: Finance and Statistics, 2007, 424 pp.
- [Military-space activity..., 2005] Military-space activity of Russia: origins, state, perspectives. Proceedings of scientific-and-practical conference [in Russian]. Saint-Petersburg: “Levsha Saint-Petersburg” Publishers, 2005, 122 p.
- [Morozov, Dymarskii, 1984] Morozov V.P., Dymarskii Ya.S. Elements of control theory for flexible manufacturing: mathematical support [in Russian]. Leningrad: Mashinostroenie, 1984, 245 pp.

- [Okhtilev et al, 2006] Okhtilev M.Yu., Sokolov B.V., Yusupov R.M. Intellectual technologies for monitoring and control of structural dynamics of complex technical plants [in Russian]. Moscow: Nauka Publishers, 2006, 410 pp.
- [Omatu et al, 1996] Omatu S., Khalid M., Yusuf R. Neuro-Control and Its Applications (Advances in Industrial Control). New York: Springer Verlag, 1996.
- [Perminov, 2007] Perminov S.B. Information technologies as a factor of economic growth [in Russian] / S.B. Perminov: [Ed. E.N. Egorov]; Central Economics and Mathematics Institute of Russian Academy of Sciences. Moscow: Nauka Publishers, 2007, pp. 195.
- [Reznikov, 1990] Reznikov B.A. System analysis and methods of systems engineering [in Russian]. USSR Ministry of Defence, 1990, 522 pp.
- [Rostovtsev, 1992] Rostovtsev Yu.G. Foundations of construction of automated systems for gathering and processing information [in Russian] Saint-Petersburg: Mozhaiskii Military-Engineering Space Institute, 1992, 717 p.
- [Seletkov, Dneprovskaya, 2006] Seletkov S.R., Dneprovskaya N.V. Progress in theory of control of information [in Russian] Information Resources of Russia. Vol.94, No.6, 2006, pp. 12-14.
- [Sidorov, Yusupov, 1969] Sidorov V.N., Yusupov R.M. Algorithmic reliability of digital control systems [in Russian]. Leningrad: Mozhaiskii Military-Engineering Space Academy, 1969, 54 pp.
- [Sokolov, 1992] Sokolov B.V. Complex operations scheduling and structure control in ACS for active moving crafts [in Russian]. Moscow: USSR Ministry of Defence, 1992, 232 p.
- [Sokolov, Yusupov, 2002] Sokolov B.V., Yusupov R.M. Complex modeling of operation of automated control system for navigation spacecrafts [in Russian] Problems of Control and Computer Science, 2002, No. 5, pp. 24-41.
- [Sokolov, Yusupov, 2008a] Sokolov B.V., Yusupov R.M. Interdisciplinary approach to complex modeling of risks in management decision-making in complex technical-organizational systems [in Russian] International Workshop "Modeling and analysis of security and risks in complex systems" (MASR—2008). Russia, Saint-Petersburg, June 24–28, 2008, pp. 146–155.
- [Sokolov, Yusupov, 2008b] Sokolov B.V., Yusupov R.M. Neo-cybernetics: possibilities and perspectives of progress [in Russian] Report made at general plenary session of the 5th scientific conference "Control and information technologies" (CIT-2008), Russia, Saint-Petersburg, October 14–16, 2008. / CSRI "ELEKTROPRIBOR", Saint-Petersburg, 2008, pp. 1–15.
- [Sovetov, 2006] Sovetov B.Ya. Theory of automated control: textbook for institutes of higher education [in Russian] / B.Ya. Sovetov, V.V. Tsekhanovskii, V.D. Chertovskii. Moscow: Vys'shaya Shkola Publishers, 2006, 463 p.
- [Starodubov, 2006] Starodubov V.A. Control of life-cycle of production, from conception until realization [in Russian]. Saint-Petersburg, 2006, 120 p.
- [Tellin, 1996] Tellin S. Internet and Adaptive Innovations: transition from control to coordination in modern organizations [in Russian] // DBMS, No. 5-6, 1996, pp. 68-79.
- [Timofeev, Yusupov, 1994] Timofeev A.V., Yusupov R.M. Intellectualization of automated control systems Technical Cybernetics, № 5, 1994.
- [Vasiliev et al, 2000] Vasiliev S.N., Zherlov A.K., Fedosov E.A., and Fedunov B.E. Intellectual control of dynamic systems [in Russian]. Moscow: Fizmatlit, 2000.
- [Vershinskaya, 2007] Vershinskaya O.A. Information-and-communicative technologies and society / O.N. Vershinskaya: Institute for Socio-Economic Studies of Population, Russian Academy of Sciences [in Russian]. Moscow: Science Publishers, 2007, p. 203.
- [White, 2004] White T. What Business Really Wants from IT: A Collaborative Guide for Business Directors and CIOs. Elsevier, 2004.
- [Wiener, 1948] Wiener N. Cybernetics: Or the Control and Communication in the Animal and the Machine. MA: MIT Press, 1948, pp. 42-43.
- [Wiener, 1950] Wiener N. The Human Use of Human Beings: Cybernetics and Society. Da Capo Press, 1950, 30 p.

- [Yarushkina, 2004] Yarushkina N.G. Theory of fuzzy and hybrid systems: a tutorial [in Russian]. Moscow: Finance and Statistics, 2004, 320 p.
- [Yusupov, 2005] Yusupov R.M. Ninetieth anniversary of Academician E.P. Popov [in Russian] Management-information systems. No. 1, 2005, pp. 51-57.
- [Yusupov, Zabolotskii, 2000] Yusupov R.M., Zabolotskii V.P. Scientific and methodological foundations of informatisation [in Russian]. Saint-Petersburg: Nauka Publishers, 2000, 425 p.
- [Zatsarinnyi, Ionenkov, 2007] Zatsarinnyi A.A., Ionenkov Yu.S. Tendencies in the progress of information technologies with account for the conception of network-centered wars [in Russian] Systems and Tools of Computer Science, Issue 17. Moscow: Nauka Publishers, 2007, p. 47-64

Authors' Information

Boris Sokolov – Doctor of Sciences (Tech), Prof., Honored scientist of Russian Federation; Deputy-Director for Research of St. Petersburg Institute for Informatics and Automation of the Russian Academy of Sciences (SPIIRAS); SPIIRAS, 14th Line, 39, St.Petersburg, 199178, Russia; e-mail: sokol@iias.spb.su

Major Fields of Scientific Research: Development of research fundamentals for the control theory by structural dynamics of complex organizational-technical systems

Rafael Yusupov – Corresponding Member of the Russian Academy of Sciences (RAS), Doctor of Sciences (Tech), Professor, Director of Institution of RAS St.Petersburg Institute for Informatics and Automation of RAS (SPIIRAS), Honored scientist of Russian Federation; SPIIRAS, 14th Line, 39, St.Petersburg, 199178, Russia; e-mail: sokol@iias.spb.su

Major Fields of Scientific Research: Control theory, informatics, theoretic basics of informatization and information society, information security

Michael Okhtilev – Doctor of Sciences (Tech), Prof., Leading Researcher of St. Petersburg Institute for Informatics and Automation of the Russian Academy of Sciences (SPIIRAS); SPIIRAS, 14th Line, 39, St.Petersburg, 199178, Russia; e-mail: oxt@mail.ru

Major Fields of Scientific Research: Development of research fundamentals for the control theory by structural dynamics of complex organizational-technical systems

Oleg Maydanovich – PhD (Tech), Assistant professor; SPIIRAS, 14th Line, 39, St.Petersburg, 199178, Russia; e-mail: sid.sn@yandex.ru

Major Fields of Scientific Research: Development of complex military-technical systems

FLOOD RISK ASSESSMENT BASED ON GEOSPATIAL DATA

Nataliia Kussul, Sergii Skakun, Andrii Shelestov, Yarema Zyelyk

Abstract: *The problem statement of disaster risk assessment, based on heterogeneous information (from satellites and in-situ data, and modelling data) is proposed, the problem solving method is grounded and considers its practical use for risk assessment of flooding in Namibia. The basis of the method is the ensemble approach to the heterogeneous data analysis with the use of the data fusion techniques and evaluation the probability density function of a natural disaster using this method.*

Keywords: *risk assessment, natural disasters, geospatial data, remote sensing, data fusion, ensemble data processing, probability density function, parametric statistics, maximum likelihood classifier, neural network classifier*

ACM Classification Keywords: *I.5 PATTERN RECOGNITION - I.5.1 Models – Neural nets; G.1 NUMERICAL ANALYSIS - G.1.8 Partial Differential Equations - Inverse problems; F. Theory of Computation - F.1.1 Models of Computation - Probabilistic computation; G.4 MATHEMATICAL SOFTWARE - Parallel and vector implementations; H. Information Systems - H.3 INFORMATION STORAGE AND RETRIEVAL - H.3.5 Online Information Services; I.4 IMAGE PROCESSING AND COMPUTER VISION - I.4.6 Segmentation - Pixel classification; I.4.8 Scene Analysis - Sensor fusion; J. Computer Applications - J.2 PHYSICAL SCIENCES AND ENGINEERING - Earth and atmospheric sciences*

Introduction

Changes in climate are caused numerous natural disasters: floods, droughts, heavy snowfall, forest fires, etc., bringing great damage to the economy of individual countries and entire regions. In recent years, to monitor natural disasters are increasingly using the geospatial data of different nature: the satellite images and products (such as digital relief model, land use maps), as well as two-dimensional or three-dimensional modelling data (particularly meteorological or hydrological models). The monitoring result of such information use has become digital maps or multi-layered geospatial data, greatly facilitating the decision-making process relevant authorities. Such information can be used not only for the mapping of disaster areas during or after the event itself, but also at other stages of the disaster cycle - including for the construction of risk maps that illustrate the probability of occurrence and the damage that can be they caused.

To problem of operational services creating for natural disaster risk assessment in Europe the project SAFER (http://www.emergencyresponse.eu/site/FO/scripts/myFO_accueil.php?lang=EN) of GMES program (Global Monitoring for Environment and Security) is devoted. The French Space Agency CNES is actively developing an approach to risk assessment of infectious diseases caused by the spread of insect vectors caused by floods in Africa (<http://www.redgems.org/spip.php?rubrique4>). However the risk assessment techniques used today in operating systems are often too simplistic and are not based on a fairly well developed mathematical apparatus for the average risk assessment that was developed for the problem of the quality estimating of the functional dependencies recovery based on empirical data and used in the statistical learning theory [Vapnik, 1995; Vapnik, 1998; Haykin, 1999; Bishop, 2006].

The problem statement of disaster risk assessment, based on heterogeneous information (from satellites, in-situ data, and modelling data) is proposed in this paper, the problem solving method is grounded and considers its practical use for risk assessment of flooding in Namibia.

Existing approaches to assessment of the natural disaster risk on the basis of geospatial information

In a variety of subject areas (economics, public health, and financing activities) the general concept of risk is determined by roughly the same. "Risk is a combination of the likelihood of an occurrence of a hazardous event or exposure(s) and the severity of injury or ill health that can be caused by the event or exposure(s)" (OHSAS 18001:2007 — Occupational Health and Safety Management Systems Requirements Standard). In general mathematically risk R often simply defined as a function f of disaster probability p and expected loss l (http://www.wired.com/science/planetearth/magazine/17-01/ff_dutch_delta?currentPage=3):

$$R = f(p, l). \quad (1)$$

In the statistical decision theory risk function to estimate $\delta(x)$ of parameter θ (using the classifier or decision rule), calculated on the basis of observation x of parameter θ , is defined as the expected value of the loss function L [Christian, 2007]

$$R(\theta, \delta(x)) = \int L(\theta, \delta(x)) f(x/\theta) dx. \quad (2)$$

In [Jonkman et al, 2003] contains a detailed review of metrics for determining the risk of an individual, as well as social, economic and other risks associated with natural disasters. However, in general, the risk is described as a function of the probability of damage. For example, a simple measure of social risk is the expected number of victims per year, calculated by the formula

$$E(N) = \int_0^{\infty} x f_N(x) dx, \quad (3)$$

where $f_N(x)$ is probability density function of the number of victims per year.

Another example of the risk function [Piers, 1998] is a function of aggregated weighted risk (AWR), defined by the relation

$$AWR = \iint_A IR(x, y) h(x, y) dx dy, \quad (4)$$

where $IR(x, y)$ is the risk of a disaster (so-called individual risk) in the position with coordinates (x, y) , $h(x, y)$ is the number of houses on location (x, y) , and A is area, for which the AWR is determined.

The following sections of paper will be formalized concept of disaster risk on the basis of heterogeneous geospatial information, and will state the risk assessment method and will identify the data sources.

Problem statement of disaster risk assessment on the basis of heterogeneous geospatial information and method of its solution

The aggregated expected risk of disaster consequences (the aggregated expected losses) in the area A will be called the value

$$R_A = \iint_A r(x, y) dx dy, \quad (5)$$

where $r(x, y)$ is the individual expected risk of disaster consequences z (individual expected losses) at the point (x, y) calculated as the mathematical expectation of damage consequences function $h_{x,y}(z)$ in the location (x, y)

$$r(x,y) = \int_0^{\infty} h_{xy}(z) p_{xy}(z) dz, \quad (6)$$

where $p_{xy}(z)$ is probability density function of the disaster z at the point (x,y) being evaluated on the basis of joint analysis of heterogeneous geospatial data. One method of estimating of the probability density function $p_{xy}(z)$ and the damage consequences function is determined by the type of disaster and will be described below.

The probability density function $p_{xy}(z)$ of the disaster is determined by various environmental factors and weather conditions that may be directly or indirectly measured by in-situ and remote-sensing methods, or obtained through modelling.

For example, the likelihood of spring flooding is determined by snow watersupplies, snowmelt intensity, air temperature and rainfall in snowmelt period in the given area and upstream, as well as by soil structure, its degree of freezing and by other factors. Expected runoff volume and water level in the river can be estimated using hydrologic models with assimilation into model satellite data and in-situ measurements data.

To recover the probability density function can be used well-developed at the statistical learning theory [Vapnik, 1995; Vapnik, 1998; Haykin, 1999; Bishop, 2006] method of the empirical functional minimizing in the problem of the average risk minimizing. Constructing an empirical functional is based on an approximation of an unknown probability density that is included in the average risk functional type of (6) (in our case), some of the empirical density restored on the basis of measurement data and use it instead of the unknown density of the empirical functional. Next, we need not solve the minimizing problem of this empirical functional of average risk, as it is done in the classical theory of dependencies recovery from empirical data. In our case, interest is only the problem of the probability density $p_{xy}(z)$ reconstructing from sample data, which is classical problem of mathematical statistics. Having restored the probability density, we can estimate the individual risk according to (6) as the expected value of the disaster damage, we can estimate the aggregated expected risk of natural disaster consequences (the aggregated expected losses) in the area A in accordance with (5) and we can use the information about the risk to decide on measures, which reduce the damage of the disaster consequences.

The problem of reconstructing the probability density in the class of continuous functions is reduced to an ill-posed problem of numerical differentiation of probability distribution function [Vapnik, 1995; Vapnik, 1998]. It can be solved using the non-parametric methods (such as Parzen's method, the method of ordered risk minimization using the covariance matrix of correlated measurement errors), which take into account the ill-posed problem and rely on the statistical theory of regularization [Vapnik, 1995; Vapnik, 1998]. However, in cases where there is a priori information about the unknown probability density, we can avoid the ill-posed formulation of this problem. For example, if the restored probability density is known up to a finite number of parameters, the problem of its recovery from empirical data is correct, and for its solutions can be used effective methods of parametric statistics [Vapnik, 1995; Vapnik, 1998]. So in a class of average risk minimizing problems associated with the classification problem (pattern recognition learning), the recovery of the unknown parameters in the probability density $p_{xy}(z)$ may be performed, in particular in [Vapnik, 1995; Vapnik, 1998], using various methods of parametric statistics (depending from the problem context): Bayesian approximations method; the best unbiased approximations method, maximum likelihood method. In addition, as models for evaluating of the probability density function of a natural disaster can be used different regression models or model in the form of a black box (such as neural network, kernel methods, other methods of machine learning theory, etc.) [Vapnik, 1995; Vapnik, 1998; Haykin, 1999; Bishop, 2006].

To estimate disaster risk probability we should analyze (classify) information from different sources with different time and space resolution. For joint analysis of such information methods of data fusion are used [Mitchell, 2007]. Density of disaster probability is estimated via fusion not raw data but mainly information of higher levels of data processing [Das, 2008]. We propose following general scheme for estimation of disaster probability density $p_{xy}(z)$ (Fig. 1).

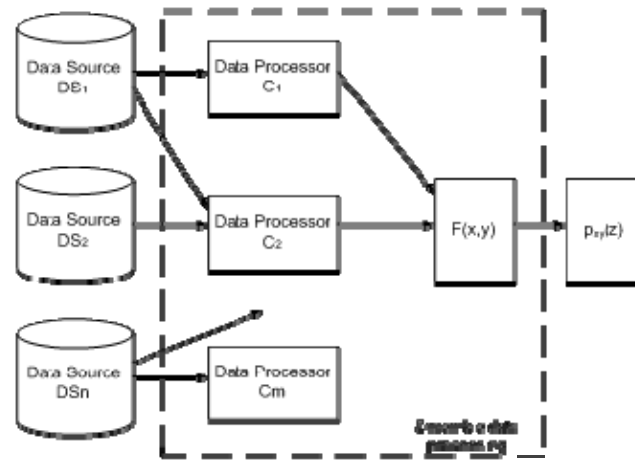


Figure 1. Estimation the density of disaster probability $p_{xy}(z)$ on the base of heterogeneous information

Blocks DS_i , $i=1, \dots, n$ in Fig. 1 represent different data sources — satellite data, in-situ observations and modelling data. Blocks C_i , $i=1, \dots, m$ provide data processing and higher level information acquisition. In general case $m \neq n$, because data from one source could be processed by several classifier and vice versa. So data fusion is provided already by classifiers. Concrete type of each classifier C_i , $i=1, \dots, m$ is determined by input data specifics. For example, satellite data processing includes several kinds of preprocessing (reprojection, geocoding, atmosphere and geometric correction and so on) before „thematic processing“. So each processor C_i , $i=1, \dots, m$ could provide several levels of data transformation, but for clearness we will not explicitly specify the levels.

Generally speaking, each processor C_i , $i=1, \dots, m$ is a special decision rule or classifier (so called weak or component classifier) analyzing data from one or several sources. Thus classifiers C_i , $i=1, \dots, m$ form an ensemble of experts (or „strong“ classifier) and their decisions are integrated within single decision F with correspondent weights α_i

$$F(x,y) = \sum_{i=1}^m \alpha_i C_i(x,y). \quad (7)$$

Ensemble of classifiers is shown in Fig. 2. Such complex method of data processing provides more accurate estimation of heterogeneous information then each of „weak“ classifiers [Jaakkola, 2006]. Note that the process of combining simple „weak“ classifiers into one „strong“ classifier is analogous to the use of kernels to go from a simple linear classifier to a non-linear classifier. The difference is that here we are learning a small number of highly non-linear features from the inputs rather than using a fixed process of generating a large number of features from the inputs as in the polynomial kernel. To improve the accuracy of classification we can use boosting technique [Kotsiantis and Pintelas, 2004], reduced to the estimation of a loss function and minimizing loss by adding new weak classifiers. Such an approach allows us to optimize the number of classifiers and complexity of the model.

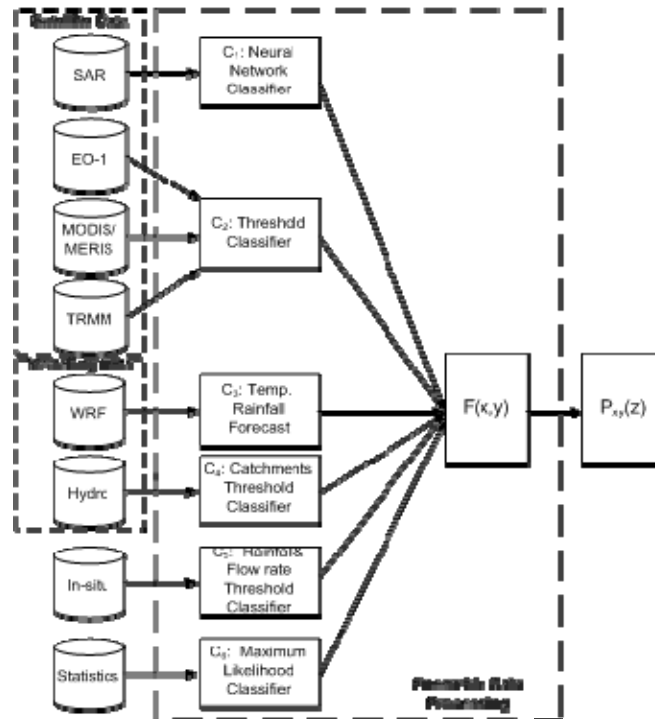


Figure 2. Estimation of $p_{x,y}(z)$ for Namibia

Case-study: flood risk assessment based on heterogeneous data

As a case study we consider the Namibia Sensor Web Pilot Project - A case study on Integrated Flood modeling, forecasting, mapping and Water-related disease management. The purpose of this project is to integrate remote sensing into a flood and water-related disease modeling, monitoring, and early warning and decision support system. This international project was initiated by Ministry of Agriculture, Water and Forestry (MAWF) and Ministry of Health and Social Services (MHSS) of Namibia; United Nations Platform for Space-based Information for Disaster and Emergency Response (UN-SPIDER); Ukraine Space Research Institute (USRI); NASA/GSFC; NOAA/National Environmental Satellite Data and Information Service (NESDIS); German Aerospace Center (DLR); and Committee on Earth Observing Satellites (CEOS) Working Group on Information Systems and Services (WGISS). The overall project framework is shown in Fig. 3.

The following data sets are used for flood risk assessment within a joint project of UN-SPIDER, NASA, DLR, NOAA and Space Research Institute NASU-NSAU:

- Satellite imagery:
 - synthetic-aperture radar: Envisat/ASAR
 - optical: EO-1, MODIS (Terra and Aqua)
 - TRMM
- Modelling data:
 - meteorological data (numerical weather prediction)
 - hydrological data (river catchments)

- In-situ observations and river gauges:
 - rainfall and river flowrateS
- Statistical data:
 - Statistical information on floods for previous years.

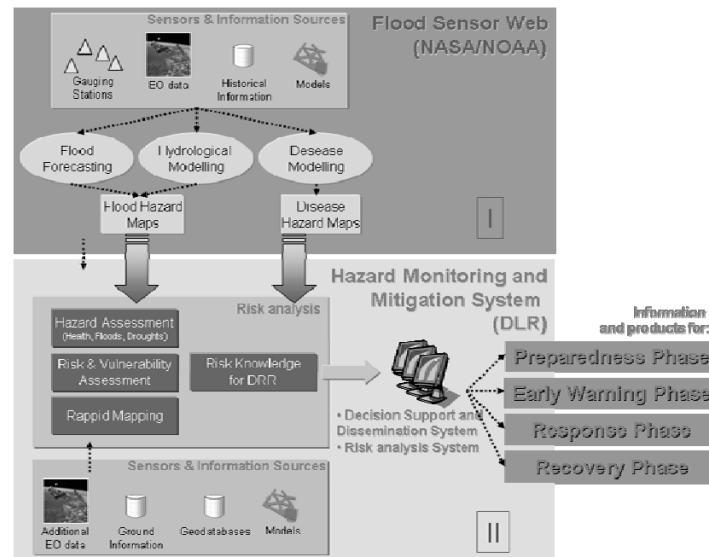


Figure 3. Overview of proposed framework

Flood mapping from satellite imagery. We use both microwave and optical satellite data to estimate the flood extent. We use intelligent computation techniques to derive flood mask from satellite imagery [Skakun, 2009; Kussul et al., 2008a; Kussul et al., 2008b]:

- Envisat/ASAR (within ESA Category-1 project): medium spatial resolution (150 m): products are delivered within 24h after image acquisition; high spatial resolution (30 m): products are delivered on demand.
- RADARSAT-2 (within the International Charter “Space and Major Disasters” and Disaster Working Group of GEO): high spatial resolution (3 to 30 m).

For cloud-free days we acquire data from optical sensors:

- Envisat/MERIS: medium spatial resolution (300 m);
- Terra and Aqua/MODIS: medium spatial resolution (250 m – 1 km);
- NASA EO-1: high spatial resolution (30 m).

Products are delivered in KML (for Google Earth), GeoTiff, WMS and others. An example is depicted in Fig. 4.

We use data from joint mission of NASA and JAXA Tropical Rainfall Measuring Mission (TRMM) to monitor rainfall rate (Fig. 5).

We have also setup a Website that shows all the flood products that were derived from Envisat/ASAR imagery (Fig. 6).

Meteorological data. We run Weather Research and Forecast (WRF) numerical prediction model to obtain forecast of meteorological parameters.

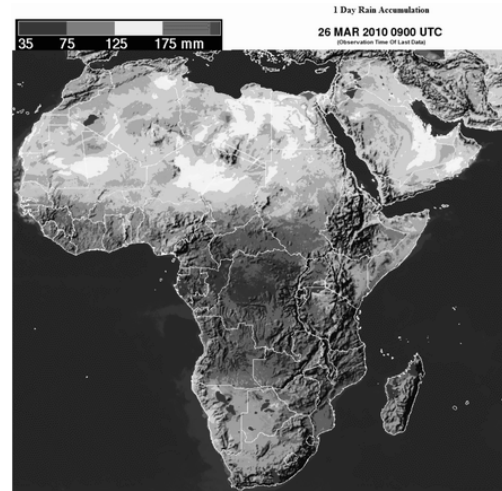
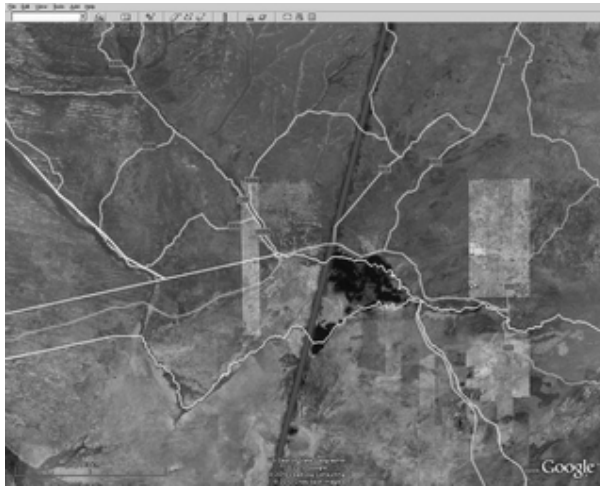


Figure 4. Flood mask for Katima-Mulilo region in Namibia Figure 5. TRMM observations derived from Envisat/ASAR, 03.03.2010

Firefox






http://namibia-project.ki.kiev.ua/

Namibian Project

Flood/water mask derived from SAR imagery

Image credit: Copyright ESA 2009, 2010

Image processing, map created by: Space Research Institute, National Academy of Sciences of Ukraine, National Space Agency of Ukraine.

Date	Flood/water Product	Product Quicklook
2010-05-24 (08:55 UTC)	KML link KML (archive) link GeoTiff link	 High-res Low-res
2010-04-25 (20:39 UTC)	KML link KML (archive) link GeoTiff link	 High-res Low-res
2010-03-28 (07:58 UTC)	KML link KML (archive) link GeoTiff link	 High-res Low-res
2010-03-27 (20:56 UTC)	KML link KML (archive) link GeoTiff link	 High-res Low-res
2010-03-03 (07:45 UTC)	KML link KML (archive) link GeoTiff link	 High-res Low-res

Firefox

Figure 6. The list of flood products derived from Envisat/ASAR data for Namibia.

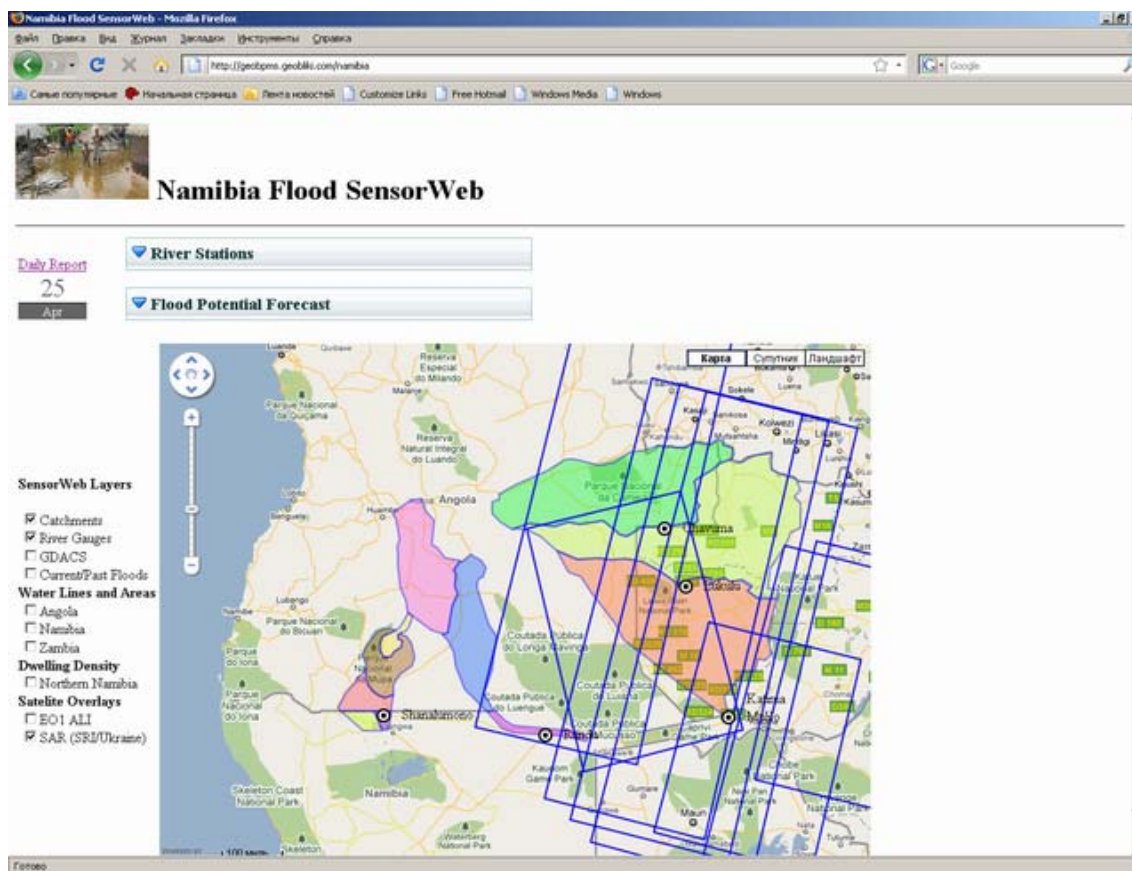


Figure 7. Web portal of the Namibia Sensor Web Pilot Project

River catchments. Data on river catchments are provided by the Ministry of Agriculture, Water and Forestry of Namibia. For each are, both archive and current data on rainfall and river flowrate are provided.

Statistical data. Statistical data of previous floods are derived from MODIS flood products that are provided by the Dartmouth Flood Observatory. These data are available from 1999.

For global flood detection we use data provided by the Joint Research Centre (JRC) of the European Commission [Groeve and Riva, 2009].

Web portal of the Namibia Sensor Web Pilot Project was established that integrates all the data and products derived by the participants (Fig. 7). These data are integrated using the ensemble approach proposed in this paper to provide flood risk assessment and are delivered to the end-users.

Conclusions

We proposed a unique approach to flood risk assessment using heterogeneous geospatial data acquired from multiple sources. This approach is based on statistical learning theory and incorporates an ensemble of classifiers for estimating probability density of the emergency. The advantage of the proposed approach is higher accuracy of risk assessment while using optimal model complexity. This approach is used within the Namibia Sensor Web Pilot Project for flood assessment using geospatial data of different nature. We plan to extend our approach to estimate the concrete risk category such as financial, social, economic, etc.

Acknowledgments

This work is partly supported by STCU project #4928 and Project № M/72-2008 of National Ministry of Education "Development of system for complex remote sensing data processing using Grid-technologies. The investigations are carried out within International Pilot Sensor Web Project for Flood Monitoring in Namibia.

Bibliography

- [Vapnik, 1995] Vapnik, V. The Nature of Statistical Learning Theory. - New York: Springer Verlag, 1995
- [Vapnik, 1998] Vapnik, V. - Statistical Learning Theory. New York: Wiley, 1998.
- [Haykin, 1999] Haykin, S. Neural Networks. A comprehensive Foundation. — New Jersey: Prentice Hall, 1999.
- [Bishop, 2006] Bishop C.M. Pattern Recognition and Machine Learning. - New York: Springer Science+Business Media, 2006. – 738 p.
- [Christian, 2007] Robert, Christian (2007). The Bayesian Choice (2nd ed.). New York: Springer. doi:10.1007/0-387-71599-1. MR1835885. ISBN 0-387-95231-4.
- [Jonkman et al, 2003] S.N. Jonkman et al. An overview of quantitative risk measures for loss of life and economic damage // Journal of Hazardous Materials A99 (2003). P. 1–30.
- [Piers, 1998] M. Piers, Methods and models for the assessment of third party risk due to aircraft accidents in the vicinity of airports and their implications for societal risk// In: R/E/ Jorissen, P.J.M.Stallen (Eds.), Quantified Societal Risk and Policy Making, Kluwer Academic Publishers, Dordrecht, 1998.
- [Mitchell, 2007] H. B. Mitchell, Multi-sensor Data Fusion – An Introduction (2007) Springer-Verlag, Berlin, ISBN 9783540714637.
- [Das, 2008] S. Das, High-Level Data Fusion (2008), Artech House Publishers, Norwood, MA, ISBN 9781596932814 and 1596932813.
- [Jaakkola, 2006] Tommi Jaakkola, course materials for 6.867 Machine Learning, Fall 2006. MIT OpenCourseWare(<http://ocw.mit.edu/>), Massachusetts Institute of Technology.
- [Kotsiantis and Pintelas, 2004] Kotsiantis S., Pintelas P. Combining Bagging and Boosting, International Journal of Computational Intelligence, Vol. 1, No. 4 (324-333), 2004.
- [Skakun, 2009] Skakun S. A neural network approach to flood mapping from satellite imagery // Scientific Papers of Donetsk National Technical University "Informatics, Cybernetics and Computer Science". - 2009. - Vol. 10(153). - P. 92-100. (in Ukrainian)
- [Kussul et al., 2008a] Kussul N., Shelestov A., Skakun S. Intelligent Computations for Flood Monitoring// International Book Series "Advanced Research in Artificial Intelligence" (ed. Markov K., Ivanova K., Mitov I.), 2008, number 2, pp.48-54.
- [Kussul et al., 2008b] Kussul N., Shelestov A., Skakun S. Grid System for Flood Extent Extraction from Satellite Images // Earth Science Informatics. - 2008. - 1(3-4). - P. 105-117.
- [Groeve and Riva, 2009] De Groeve, T., P. Riva, 2009. Global real-time detection of major floods using passive microwave remote sensing. Proceedings of the 33rd International Symposium on Remote Sensing of Environment, Stresa, Italy, May 2009.

Authors' Information



Natalia Kussul – Prof., Deputy Director, Space Research Institute NASU-NSAU, Glushkov Prospekt 40, build. 4/1, Kyiv 03680, Ukraine;
e-mail: inform@ikd.kiev.ua

Major Fields of Scientific Research: Grid computing, design of distributed software systems, parallel computations, intelligent data processing methods, Sensor Web, neural networks, satellite data processing, risk management and space weather.



Andrii Shelestov – Prof., Senior Scientist, Institute of Cybernetics NASU, Glushkov Prospekt 40, build. 4/1, Kyiv 03680, Ukraine;
e-mail: inform@ikd.kiev.ua

Major Fields of Scientific Research: Grid computing, intelligent methods of estimation and modeling, satellite data processing, Sensor Web, neural networks, multi-agent simulation and control, software engineering, distributed information systems design.



Sergii Skakun – Senior Scientist, Space Research Institute NASU-NSAU, Glushkov Prospekt 40, build. 4/1, Kyiv 03680, Ukraine;
e-mail: serhiy.skakun@ikd.kiev.ua

Major Fields of Scientific Research: Remote sensing data processing (optical and radar), image processing, Grid computing, Sensor Web, neural networks, analysis of computer system's users behavior.



Yarema Zyelyk – Prof., Leading Scientist, Space Research Institute NASU-NSAU, Glushkov Prospekt 40, build. 4/1, Kyiv 03680, Ukraine;
e-mail: adapt@ikd.kiev.ua

Major Fields of Scientific Research: estimation and control under uncertainty, dynamics of structural formations and magnetic fields on the Sun, the analysis and prediction of the temporal series and fields, information systems and technologies

LARGE VLSI ARRAYS – POWER AND ARCHITECTURAL PERSPECTIVES

Adam Teman, Orly Yadid-Pecht and Alexander Fish

Abstract: *A novel approach to power reduction in VLSI arrays is proposed. This approach includes recognition of the similarities in architectures and power profiles of different types of arrays, adaptation of methods developed for one on others and component sharing when several arrays are embedded in the same system and mutually operated. Two types of arrays are discussed: Image Sensor pixel arrays and SRAM bitcell arrays. For both types of arrays, architectures and major sources of power consumption are presented and several examples of power reduction techniques are discussed. Similarities between the architectures and power components of the two types of arrays are displayed. A number of peripheral sharing techniques for systems employing both Image Sensors and SRAM arrays are proposed and discussed. Finally, a practical example of a smart image sensor with an embedded memory is given, using an Adaptive Bulk Biasing Control scheme. The peripheral sharing and power saving techniques used in this system are discussed. This example was implemented in a standard 90nm CMOS process and showed a 26% leakage reduction as compared to standard systems.*

Keywords: VLSI Arrays, SRAM, Smart Image Sensors, Low Power, AB²C.

ACM Classification Keywords: B.3.1 Semiconductor Memories - SRAM, B.6 Logic Design – Memory Control and Access, B.7 Integrated Circuits – VLSI, E.1 Data Structures – Arrays, I.4.1 Digitization and Image Capture

Introduction

The continuing persistence of Moore's Law [Moore65] throughout recent years has led to great opportunities for embedding complex systems and extended functionality on a single die. The primary example of this trend is the modern day, high performance, multi-core microprocessor that employs large memory caches in order to achieve large bandwidth. Another popular example is the smart image sensor, which integrates additional capabilities of analog and digital signal processing into a conventional CMOS sensor array. Both microprocessors and image sensors are frequent components of various Systems-On-Chip (SOC) that also embed several additional SRAM arrays for various functionality. As a result of these trends, large VLSI arrays frequently cover a large area of various microelectronic systems, sometime well over half of the total silicon die.

One of the side effects of the integration of large VLSI arrays is, of course, power consumption. In the last decade, low-power design has ousted high-performance as the main focus of the VLSI industry. This is a result of the constant exponential rise in power density over the past three decades, coupled with the rise in popularity of mobile, battery powered devices. This power increase proved to be unacceptable in immobile, high performance systems, when the cost and complexity of heat dissipation became too high, and in mobile devices, where increased performance and functionality are required alongside the need for large spans between recharging. In today's systems, it is very common that the main source of power consumption is the large memory arrays. In digital camera systems, the pixel array along with its periphery are obviously the main consumers of power, and likewise, in other SOC's comprising smart imagers, they tend to be close to the top of the list. These facts lead us to realization that low power solutions for embedded arrays are a necessity in modern VLSI design.

In this paper, we have chosen the two types of arrays mentioned above, embedded SRAM bitcell arrays and image sensor pixel arrays, for discussion. Through these examples, we will show that there are several similarities in the architectures and power profiles of different types of arrays. Many techniques and solutions have been developed for power reduction in each type of array, but rarely has one technique been adapted to fit

another type of array. Through our discussion, we will show that such possibilities exist and provide an important direction for low power research.

The discussion will start with a review of the architectures of both types of subsystems (i.e. bitcell and pixel arrays), describing the components that compose each. We will then discuss the sources of power consumption and the related problems for each subsystem, as well as a number of existing low power solutions for each case. We will continue with a comparison of the two types of subsystems, highlighting similarities and discussing peripheral sharing opportunities. Finally, we will give an example of a system, recently developed by our group, that utilizes these similarities to achieve power reduction in a smart image sensor system with embedded memory.

SRAM Architecture and Power Considerations

Modern digital systems require the capability of storing and accessing large amounts of information at high speeds. Of the different types of memories, the Static Random Access Memory (SRAM) is the most common embedded memory, due to its high speeds and relatively high density in standard fabrication processes. SRAMs are widely used in microprocessors as caches, tag arrays, register files, branch table predictors, instruction windows, etc. and occupy a significant portion of the die area. In high-performance processors, L1 and L2 caches alone occupy over half of the die area [Mamidipaka, 2004]. Accordingly, SRAMs are one of the main sources of power dissipation in modern VLSI chips, especially high-end microprocessors and SOCs.

Figure 1 shows a typical block diagram of an SRAM, with emphasis on the main components and sub-blocks. The core of the SRAM is an array of identical bitcells, laid out in a very regular and repetitive structure, each bitcell storing either a '1' or '0' on a cross-coupled latch, and enabling read and write access. The bitcells are divided into rows and columns, allowing complete random access, through the use of X and Y addressing circuitry consisting of a row decoder and a column multiplexer. The addressing is propagated to the individual bitcell through a grid of horizontally wired *wordlines* and vertically wired *bitlines*. A particular bitcell is accessed (either read or written) when its row's wordline and its column's bitline are asserted simultaneously.

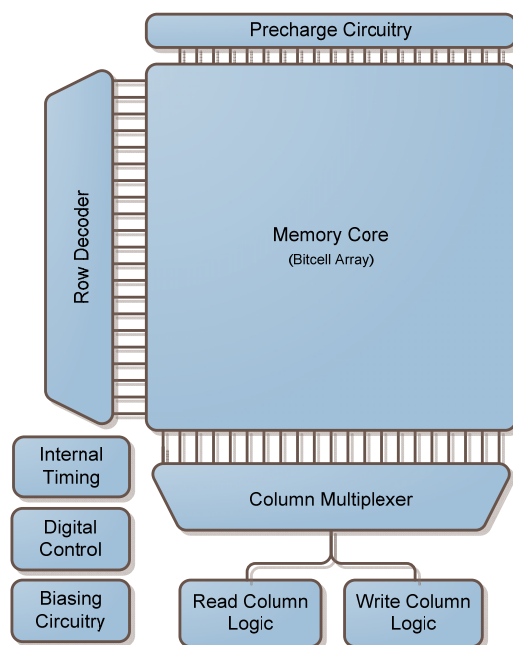


Figure 1: Typical SRAM Component Block Diagram

In order to initiate either a read or a write, the read column logic and write column logic blocks are required. The read column logic block typically consists of a low swing sense amplifier to enhance the performance and readout a digital signal from the asserted column. The write column logic block consists of a write driver that asserts the data to be written onto the relevant column. A read/write enable control signal selects which of the two blocks is activated, and the asserted wordline initiates the bitcell on the selected column to be read from or written to.

Additional blocks needed for SRAM operation include column precharge, internal timing, digital control blocks and biasing circuitry. The column precharge block prepares the read/write operation by setting the columns into a known state. The internal timing blocks sense various transitions in internal and external signals to initiate or terminate operation phases. The digital control blocks enable application of advanced error correcting, row/column redundancy, etc. The biasing circuitry is generally required for sense amplifier operation.

The power profile of SRAMs include both dynamic power, consumed during read and write operations, as well as static power, consumed during standby ("hold") periods. The dynamic power, similar to standard logic, is a function of the supply voltage and frequency, giving the standard tradeoff between power and performance/reliability. Static power in SRAMs, on the other hand, is mainly due to unwanted parasitic leakage currents. As technology scales, leakage currents become a more dominant factor, causing the static power of SRAMs to become a major issue and one of the primary static power components of many systems. A unified active power equation is given in Equation 1 [Rabaey2003] [Itoh2001]:

$$P = V_{DD} (I_{array} + I_{decode} + I_{periphery}) = V_{DD} \left\{ [mi_{act} + m(n-1)i_{hld}] + [(n+m)C_{DE}V_{int}f] + [C_{PT}V_{int}f + I_{DCP}] \right\} \quad (1)$$

where m and n are the number of columns and rows, respectively, f is the operating frequency, V_{DD} is the general supply voltage, V_{int} is the internal supply voltage, i_{act} is the effective current of the selected cells, i_{hld} is the data retention current of inactive cells, C_{DE} is the output node capacitance of each decoder, C_{PT} is the total capacitance of the digital logic and periphery circuits, and I_{DCP} is the static current of the periphery.

The dynamic power of an SRAM is mainly consumed in the following areas: address decoding, bitline charging/discharging, and readout sensing. During address decoding, power is consumed both by the switching of the decoders themselves, as well as by charging and discharging the selected wordlines, which can have high capacitances. During both read and write operations, the bitlines are precharged and subsequently discharged. This is especially power consuming during writes, when the bitline is fully discharged, or when a full discharge read scheme is chosen. Sense amplifiers typically depend on bias currents for operation, consuming constant power when they are activated.

The static power of an SRAM is primarily consumed through leakage currents inside the bitcells themselves during standby (hold) periods, i.e. when the particular cell (or the whole array) is not asserted. This includes subthreshold and gate leakages in both the inner cross coupled latch structure, as well as to/from the bitlines through the access transistors on unselected rows. Another large contributor to static power is from the precharge circuitry, when a constant charging scheme is used, i.e. a high-resistance supply or diode-connected transistor is placed on the bitlines to replenish lost precharge voltage. Other contributors to the static power are the leakage currents in the decoders and other blocks.

An in-depth analysis of the power dissipation by all SRAM components can be found in [Itoh2001].

Several standard methods have been developed over the years to reduce the power consumption of SRAMs. The standard methods are based on physical partitioning of the array in each of the axes. Banked organization of SRAMs divides the array both horizontally and vertically into sub-arrays. An external decoder raises the chip select of the selected bank, reducing the dynamic power consumption, as smaller decoders are needed, and less

wordline and bitline capacitances are charged/discharged. The Divided Word Line (DWL) approach divides the array horizontally, propagating the decoder output on a global wordline, and subsequently raising the local wordline of a partition of columns, reducing the overall capacitance charged, and requiring smaller wordline drivers. Partitioning the columns using the Divided Bitline scheme, with partial multiplexing inside the array, reduces the bitline capacitance and in certain sensing schemes, will reduce the power consumption. All of these solutions come at the expense of additional area overhead, but a good tradeoff can achieve a worthwhile reduction of power consumption as well as an improvement in performance.

Using advanced timing and sensing schemes is another standard method to achieve a substantial dynamic power reduction. Using pulsed word lines and/or reduced bitline voltage swings, results in less discharge during read cycles, but is accompanied with complex design considerations and higher sensitivity to process variations. Timing the activation of sense amplifiers limits biasing currents to be present only during the exact times that the sensing is carried out. Additional low static power sense amplifiers, such as a Differential Charge Amplifiers and Self Latching Sense Amplifiers, also achieve static power reduction.

Many schemes have been proposed to reduce the bitcell leakage power, such as Supply Voltage Gating [Powell2000] [Flautner2002], Reversed Body Biasing (RBB) [Nii1998] [Hanson2003], Dynamic Voltage Scaling [Kim2002] and Negative Word Line (NWL) application [Wang 2007]. Recently, many proposals have shown minimum energy point operation of SRAMs in the subthreshold or near-subthreshold region. Examples of these include various works by Chandrakasan and Calhoun et.al. [Chandrakasan2007] [Chandrakasan2008] [Calhoun2007].

CMOS Image Sensor Architecture and Power Considerations

Traditionally, digital image sensors were fabricated in Charge Coupled Device (CCD) technology, but the integration of image sensors into more and more products, made the Active Pixel Sensor (APS) an attractive solution. This image sensor architecture is implemented in standard CMOS technology processes, and provides significant advantages over the CCD imagers in terms of power consumption, low voltage operation, and monolithic integration. With the rising popularity of portable, battery operated devices that require high-density ultra low power image sensors, the CMOS alternative has become very widespread. In addition, the CMOS technology allow for the fabrication of so called "smart" image sensors that integrate analog and/or digital signal processing onto the same substrate as the imager and its digital interface. Low power smart image sensors are very useful in a variety of applications, such as space, automotive, medical, security, industrial and others [Fish2007].

CMOS image sensors generally operate in one of two modes: rolling shutter or global shutter (snapshot) mode. When rolling shutter mode is used, each row of pixels is initiated for image capture separately in a serial fashion. This creates a slight delay between adjacent rows, resulting in image distortion in cases of relative motion between the imager and the scene. With the global shutter technique, the image is captured simultaneously by all pixels, after which the exposure is stopped, and the data is stored in-pixel while the image is read out. The operation of both techniques can be divided into three stages: *Reset*, *Phototransduction* and *Readout*. During the *Reset* stage, an initial voltage is set on the photodiode capacitance that constitutes most of the pixel area. Subsequently, the pixel enters the *Phototransduction* stage, during which the incident illumination causes the capacitance to discharge throughout a constant integration time. Readout is commenced at the end of the integration time, and the final value of the pixel is read out and converted to a digital value.

Figure 2 shows a component block diagram of a generic smart CMOS APS based image sensor. The core of the image sensor is a pixel array, generally consisting of a photodiode, in-pixel amplification, a selection scheme and a reset scheme. A full description of the operation of this pixel is given by Yadid-Pecht, et.al. [Yadid-Pecht2004].

Some smart imagers employ more complex pixels, enabling them to perform analog image processing at the pixel level, such as A/D conversion.

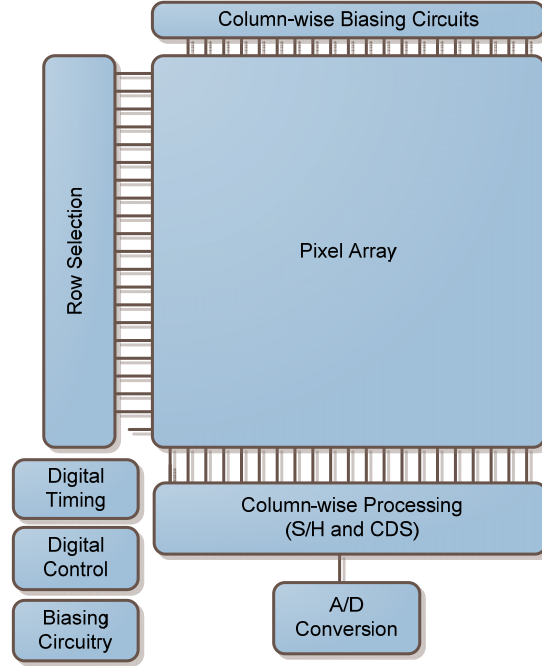


Figure 2: Generic Smart Image Sensor Component Block Diagram

Access to the pixels is carried out through the row selection block. This is usually made up of a shift register, as serial access is commonly employed, although in certain applications, a digital decoder is preferred. An entire row is generally accessed simultaneously for both reset and readout operations, except for in applications where random access is required, such as tracking window systems.

Several blocks are required at every column for the parallel operation of an entire pixel row. These include Sample and Hold (S/H) circuits, Correlated Double Sampling circuits and Analog to Digital Converters (ADC). The S/H circuitry generally measures the reset level of the pixels to enable the CDS to remove Fixed Pattern Noise (FPN). Column-wise ADCs are only one option; the others being in-pixel ADC or single ADC per imager. The selected scheme will be according to the tradeoffs of area, power, speed and precision.

Additional blocks that are required in the periphery of the imager include the general Biasing Circuitry and Bandgap References for creating biasing currents for the in-pixel signal amplifiers, usually implemented through a Source Follower (SF) scheme. and the ADCs; Digital Timing and Digital Control blocks for producing the proper sequencing of the addresses, ADC timing, etc.

The sensitivity of a digital image sensor is usually proportional to the area of the photodiode and the resolution is set by the number of pixels. This results in a relatively large area covered by the image sensor, compared to other on-chip circuits, and accordingly, a large percentage of the overall power consumption. The contribution of different image sensor components to overall power dissipation may vary significantly from system to system. For example, pixel array power dissipation can vary from a number of μ Watts for a small array employing 3 transistor APS architecture to hundreds of mWatts for large format "smart" imagers employing in-pixel analog or digital processing. The power dissipation of the pixel array of a generic "smart" image sensor can be given by Equation 2:

$$P_{Array} = F_R \times N \times M \times (E_{reset} + E_{read_out} + E_{analog} + E_{digital}) + N \times M \times P_{leakage} \quad (2)$$

where F_R is the imager's frame rate, N and M are the number of rows and columns, respectively, E_{reset} is the energy required for pixel reset, $E_{\text{read_out}}$ is the energy dissipated during signal readout during one frame, E_{analog} and E_{digital} are energy dissipation components dissipated by in-pixel analog and/or digital processing during one frame and P_{leakage} is the in-pixel leakage power.

The dynamic power in the above equation is proportional to the frame rate and is composed of the energy required to refill the photodiode capacitance during reset; the power dissipated through column-wise biasing currents during readout; and additional energy consumed by (optional) in-pixel functionality. The static power is due to leakage through the reset and row selection switches during integration and standby periods. These leakages also degrade the performance and precision of the imager.

The row selection block can be another major source of power dissipation, depending on the size of the array and the method of operation. In both global and rolling shutter modes, the row reset and row selection capacitances are periodically charged, proportional to the frame rate. In window tracking applications, on the other hand, the power of the row (and column) selection blocks can be dominated by leakage power, as the majority of the rows/columns may not be asserted for long periods.

The other primary source of power dissipation is the analog circuitry, including the ADCs, S/H, CDS and biasing circuitry. Optimally, these are timed to consume power only during their precise periods of operation, but they generally have a high power profile. The analog peripheral blocks present a constant tradeoff between speed, noise immunity, and precision versus power consumption and area, and for low power systems, the choice of these blocks needs to be made cautiously.

Additional power is dissipated in the digital timing and control blocks; however, the complexity and frequency of these tend to be lower than standard digital circuits, and so most common power reduction techniques can be implemented on these blocks.

An in depth description of all the contributions to power dissipation in a smart image sensor is given by Fish, et.al. [Fish2008].

Image sensors provide power reduction opportunities at all the design levels, starting with the technology and device levels, through the circuit level and all the way to the architecture and algorithm level. Standard power reduction techniques, such as supply voltage reduction and technology scaling, aren't always applicable to CMOS image sensors, as they are frequently accompanied by unacceptable tradeoffs. Supply voltage reduction reduces both the precision and the noise immunity of image sensors, while technology scaling generally includes side effects, such as increased leakage current and dark current, as well as reduced photoresponsivity. However, at the technology level, processes can be modified for low power image sensor fabrication albeit, at an increased cost. An example of such a process is the Silicon-on-Sapphire (SOS) process that provides a very low power figure and enables backside illumination [Culuriciello2004].

The device and circuit level provide several opportunities for limiting power dissipation, depending on the options and layers provided by the chosen technology. The presence of separate wells for both nMOS and pMOS transistors enables the application of body biasing on inactive rows for leakage reduction. This technique loses its effect with scaling, as the effect on a devices threshold voltage is reduced, but image sensors are generally fabricated in technologies up to 90nm, where it is still efficient. Additional devices, such as high-VT transistors and thick oxide transistors can also be used for leakage reduction on slow busses. Another technique commonly used for leakage reduction is serial connection of "off" transistors for "stack effect" utilization [Narendra2001].

Smart image sensors provide many interesting opportunities for power reduction at the architectural and algorithm levels. Depending on the functionality of the sensor, these systems can be equipped with designated blocks for eliminating unnecessary power consumption. An example of this is the tracking sensor we proposed

[Teman2008] that used row and column shift registers for window definition and an analog winner-take-all circuit for motion tracking. In this system, the pixels outside the window of interest were deactivated and ADCs were used only for initial detection. The switching activity of the shift registers was very low, as well, further reducing the system power consumption.

Similarities between SRAMs and Image Sensors

In the previous two sections, the architectures and power profiles of two types of VLSI arrays, SRAMs and Image Sensors, were presented along with a number of examples of methods for reducing the power consumption of each. This section will deal with the similarities between these two architectures and their sources of power dissipation, arguing that low power approaches and methods developed for one type of array should be researched and adapted for the other.

Clearly, the first similarity between the two architectures is the two-dimensional array based structure of m rows and n columns of identical unit cells. SRAM bitcells have been optimized over the years to produce a dense layout to fit as much memory as possible onto a given area. This is possible due to the regular patterning of the cells, allowing many exceptions in design rules. The dense layout results in reduced capacitances, provides benefits in power and performance, as well as smaller peripheral circuits for a given memory size. In the case of pixel arrays, dense layouts provide similar benefits; however, the reduction in pixel size has a negative effect on pixel sensitivity, quantum efficiency, noise figures, etc. Various approaches for an optimal pixel layout have been proposed, such as the hexagonal shaped pixel [Staples2009].

For both SRAM bitcell and imager pixel design, leakage current during idle cycles (“hold” cycles for bitcells and “integration” cycles for pixels) has become a major focus. In some cases, smart image sensors contain memory circuits in-pixel, which further deepens the similarity. Utilization of leakage reduction methods, such as multiple threshold transistors and body biasing have been presented for both types of arrays. Modern CMOS processes include designated transistors for use in SRAM bitcells, optimized for leakage reduction. Several groups are researching low voltage operation of SRAMs in the subthreshold or near-subthreshold regions of operation. This approach could be used for image sensors, especially for operation of in-pixel or peripheral logic, due to their reduced frequency requirements.

Random or pseudo-random access to the unit cells in both types of arrays is achieved through row and column addressing. In SRAM design, the row addressing for wordline assertion is generally achieved through a row decoder, while standard image sensors, operating in the global or rolling shutter modes, use shift registers for reset and row selection. Both types of circuits are fitted to the pitch of the rows for layout and have been deeply researched for optimal operation in terms of power, area and performance, especially due to the fact that they drive large capacitances. Certain image sensors employ decoders for row addressing (such as the tracking window example, given above), while serially accessed SRAMs benefit from using a shift register. Other architectures have also been proposed, such as daisy chaining bitcells for robust digital column-wise readout [Chandrakasan2006]. SRAMs often save power and improve performance by sub-dividing the arrays into banks, local wordlines, etc. Image sensors could partially adopt similar techniques at opposing sides of the array or fully adopt them at the expense of losing several pixels that could be compensated for through signal processing.

Column addressing, which is inherent to most SRAM designs, is used in some image sensors, when random access is necessary. Column-wise operations are performed on the data of both types of arrays; SRAMs perform write-driving, precharging and readout in this fashion, while imagers perform column-wise CDS and readout. The primary noise cancellation mechanism for imagers is the CDS function, while SRAMs often employ dummy columns for timing and level comparison. Both concepts provide opportunities to be adapted to the other field.

Both fields employ analog blocks, necessary for performance, accuracy and functionality. SRAMs use sense amplifiers to speed up readout and reduce bitline swing, while imagers use ADCs to create a digital readout from the analog signal measured by the pixel. Both are done either column-wise or one-per array (or bank). Both require biasing currents for proper operation. Both are major power consumers and should be timed carefully to operate only when necessary. Smart image sensors sometimes use alternative readout blocks, such as Winner Take All (WTA) circuits, when binary decisions are required rather than precise level readout. Similar uses could be applied to SRAMs used by specific applications.

Finally, both architectures employ digital control and timing blocks to administer their respective operating modes. Image sensors require precise timing of their reset and integration signals, as well as for CDS and ADC operation. SRAMs are often asynchronously self-timed, employing Address Transition Detectors (ATD) and other circuits to initiate precharge, read and write phases. Digital control logic maneuvers the components between operation modes, and often registers are used to latch read out signals. Careful design of these timing and control blocks can provide substantial power savings.

Table 1. summarizes the architectural similarities between SRAMs and Image Sensors:

Designation	SRAM Component	Imager Component
$m \times n$ Array	Bitcells	Pixels, in-pixel memory, in-pixel ADC
Row Addressing	Decoder, Row Drivers, wordline	Shift Register, Row Drivers, Row Selection lines, Reset lines
Column Addressing	Column Multiplexer, Bitlines	Readout columns, optional column decoder/multiplexer
Column-wise Operation	Precharge circuits, Write Drivers, Bitlines, Column Sense Amplifier	Sample and Hold, CDS, Column ADC
Analog circuitry	Sense Amplifiers, Biasing Circuitry	ADC, Biasing Circuitry, Bandgap Reference
Timing and Control	Digital Control, Self Timing logic, ATD, Dummy Column, Error Correction	Digital Control, Digital Timing

Table 1: Summary of Architectural Similarities between SRAM and Image Sensors

Peripheral Sharing

In the previous section, we discussed the similarities between SRAM arrays and Image Sensors. The correlation between the two types of arrays is even more inherent in systems that employ both units, working in cohesion to achieve certain functionalities. This is often the case in smart image sensor systems that use SRAM arrays to temporarily store previously read out data or results of image processing. In such cases, the similarities provide several architectural opportunities for sharing peripherals, thus resulting in a reduction of both power and area, and often a performance improvement due to the inherent synchronization between the units.

Figure 3. shows two examples of peripheral sharing. In the Figure 3(a), a column-wise shared architecture is shown. In this case, the readout columns of the pixel array are directly connected to the vertical writing and/or reading logic of the SRAM. A possible application is a smart image sensor that periodically stores spatial data in an embedded SRAM for further use or processing. In this case, the parallel readout of the image is directly routed

(through the column-wise processing, such as CDS, S/H and possibly ADC circuits) to the SRAM write drivers. SRAM operation is simplified to a one-dimensional (row) access scheme, as an entire row of data is read out from the image sensor and written in parallel. This architecture saves power and area, by simplifying or even eliminating the column addressing circuitry, integrating the timing and control signals of the two arrays, and even providing opportunity for replacing the SRAM's row decoder with a much smaller and less power hungry shift register. Careful design can further reduce the digital and analog blocks by creating control signals and biases appropriate for both arrays.

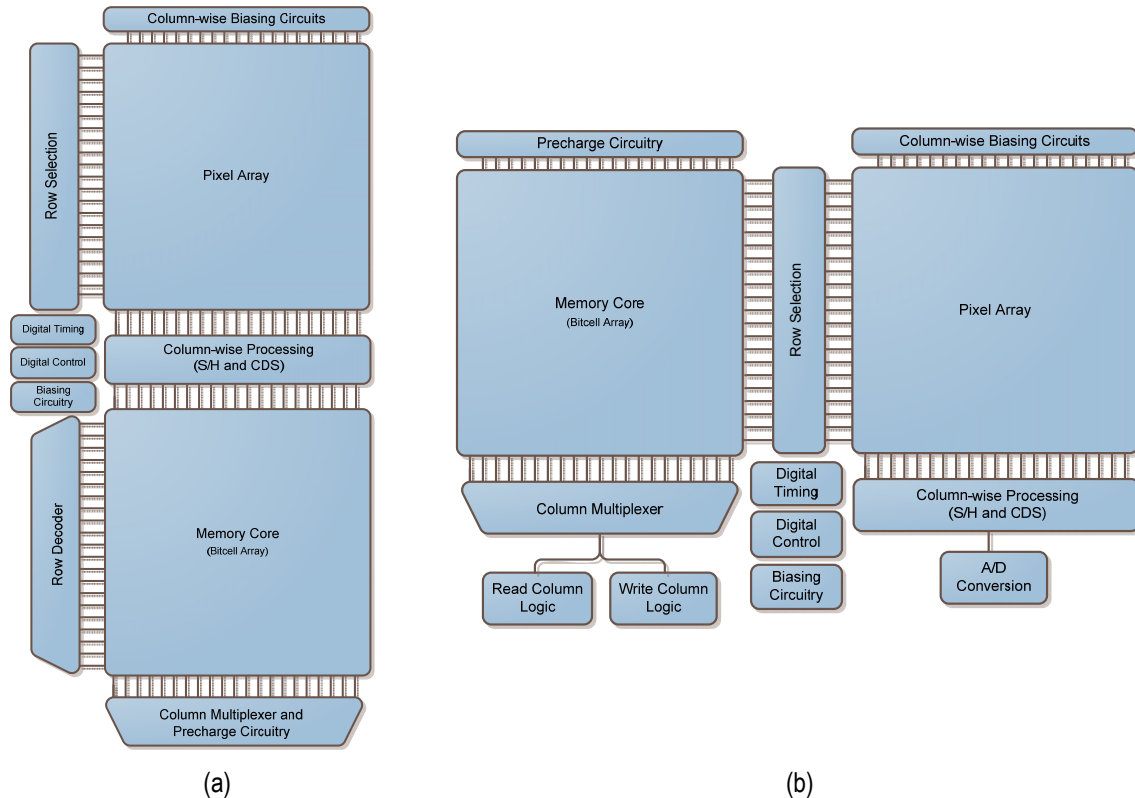


Figure 3. Two examples of peripheral sharing between SRAM and Imager arrays.
 (a) Column-wise peripheral sharing. (b) Row-wise peripheral sharing

Figure 3(b) shows a possible row-wise approach to peripheral sharing. In this architecture, the two arrays are placed on a horizontal axis, enabling the distribution of row addressing signals via a mutual row selection block. One example would place a shift register in between the two arrays, asserting the reset and row selection lines of the imager in coordination with the wordlines of the SRAM, according to a predefined timing scheme. This method of SRAM addressing could be used in a serially accessed memory working in coordination with the adjacent imager. Certain applications would allow a further reduction in peripherals (saving both power and area), by integrating the column addressing blocks of the two arrays. Digital timing and control blocks could again produce common signals and analog blocks could be designed to use similar biasing levels, further integrating the two systems.

Several other peripheral sharing architectures and techniques could be proposed, depending on the application, the relationship between the smart imager and the SRAM and the operating profile of the system. Such peripheral sharing doesn't necessarily have to include complete integration between the two arrays. For example, a

significant reduction in area could be achieved by using a single bandgap reference block for a standard imager and an SRAM array on the same die, even if they are independent of each other.

These architectural opportunities should be taken into consideration when developing a system that uses both types of blocks, as the saving in power and area, as well as the prospect of performance enhancement, can be considerable. The following section gives a practical implementation example of such a system.

Implementation Example: An Improved Adaptive Bulk Biasing Control (AB²C) System

In the previous sections, we argued that image sensors and memory arrays have many common features and that power reduction techniques, developed for one field, could be adapted for the other. In addition, we proposed opportunities for peripheral sharing in systems, such as smart image sensors, that include both pixel arrays and embedded SRAM arrays. In this section, we will present an example of a system, developed by our group, that utilizes both approaches for power and area reduction, as well as performance optimization.

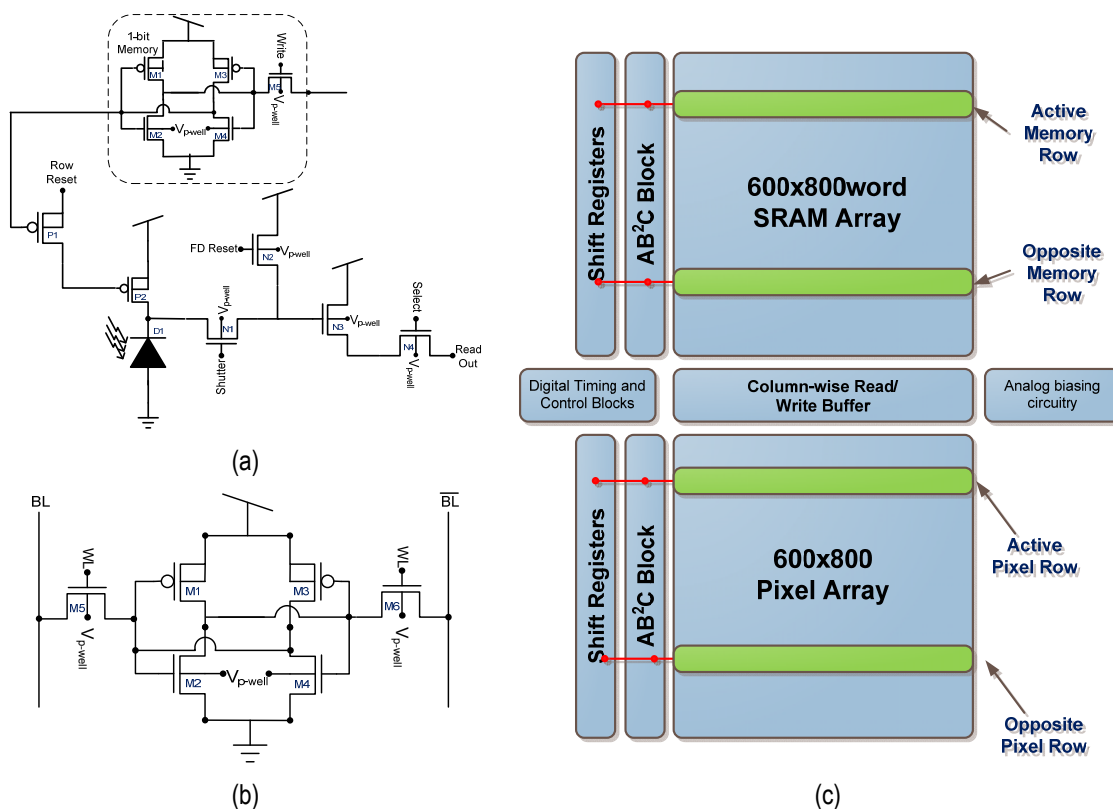


Figure 4. Architecture and basic circuits for Improved AB²C System.

(a) Schematic of Smart Pixel (b) Schematic of SRAM Bitcell (c) Full system architecture

The Adaptive Bulk Biasing Control (AB²C) approach to leakage reduction in image sensors was originally proposed by Fish, et. al. [Fish2007]. This system took advantage of the serial row access scheme, inherent to the majority of image sensors, for the application of a gradually changing body biasing to reduce the leakage in image sensors during the long integration periods. It can be shown that the slow voltage gradient applied to the bulk of a given row requires less power and causes less spatial noise than a standard pulsed approach. The

system applies the full Reversed Body Bias (RBB) to the rows farthest away from the selected (i.e. reset or readout) row, and no RBB (or potentially a performance enhancing Forward Body Bias) to the selected row.

An improved AB²C system [Teman2009] implements the original concept on a smart image sensor employing an embedded memory. Figure 4(a) shows the pixel circuit implemented in the smart image sensor employing an in-pixel memory bit. The serial access scheme of the smart imager includes a periodic partial readout of the pixel level, and according to the illumination level, data is written to both the in-pixel memory bit and the embedded SRAM array. After the full integration time, the final pixel level is read out along with the data stored at the associated SRAM address. This system provides opportunities for both row-wise and column-wise peripheral sharing, due to the synchronized serial operation of the image sensor with its associated SRAM addresses. A column-wise approach was chosen, as the parallel propagation of the column data to and from the SRAM proved to be more dense.

Implementation of the adaptive bulk biasing approach for leakage reduction in the SRAM was enabled by the serial access operation, inherent to the system. A twin-well was used to separately bias the bulks of each row of nMOS transistors, as the pMOS body biasing in deep submicron technologies (a standard 90nm TSMC process was used) is inefficient. The SRAM bitcell schematic is shown in Figure 4(b). The body nodes of the nMOS transistors was connected to the AB²C circuit, driven by the row addressing shift register, used to serially access the array.

The full architecture for the Improved AB²C system is shown in Figure 4(c). The column-wise setup enabled parallel writing of the image sensor readout values directly into the selected SRAM word below it, and subsequent readout of the SRAM value along with the final pixel value after integration. Similar row addressing blocks were used for both arrays, comprising shift registers for horizontally wired signals (reset, row select, wordlines) and AB²C circuits. These circuits include a network of resistors with connections to the bulks of rows between them. The resistor network is biased with a voltage running between the active row and the opposite row (i.e. the row farthest away from the active row), thus creating a gradual voltage drop on the bulks of adjacent rows. The bias point is switched along with the row selection shift register, causing a small charge/discharge of the row bulk capacitance, with a minimal energy penalty. The row selection blocks (including AB²C circuitry) could be shared between the two arrays, pending routing options, further saving area and power.

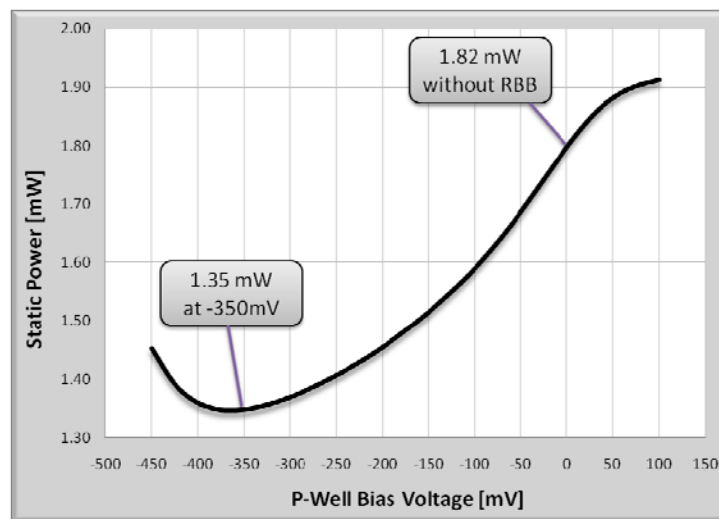


Figure 5: Static power consumption at various body biasing levels for the presented smart image sensor with embedded SRAM employing an AB²C biasing scheme.

The static power reduction achieved with the application of the AB²C architecture is plotted in Figure 5 for the presented system implemented in a standard TSMC 90nm CMOS process. The minimum energy point was achieved with a reverse biasing voltage of 350mV. A higher RBB results in higher power dissipation of the AB²C blocks, while a lower RBB results in more pixel/bitcell leakage power. This results in a 26% power reduction as compared to the same system without the biasing voltage or the AB²C power, as seen at the 0V point on the figure. This reduction improves with array sizes, and is even more effective at older technologies with a higher supply voltage, often used for image sensor implementation.

Conclusions and Further Research

A novel approach to power reduction in VLSI arrays was presented. The architectures of two types of arrays, image sensors and SRAM arrays were described. Sources of power consumption were noted for each array type, and some common techniques for power reduction were shown. It was contended that the similarities between the array types provide many opportunities for adaptation of methods and techniques for power reduction and optimization between the two. A number of architectural concepts based on peripheral sharing were suggested for systems employing both types of arrays. Finally, an example of a system that implements both approaches (method adaptation and peripheral sharing) was presented. The example showed an AB²C scheme that was originally developed for image sensors and was implemented on an SRAM array, as well, providing a substantial static power reduction for the entire system. The image sensor and SRAM array were connected in a column-wise scheme, further saving both area and power, while optimizing the operation process.

Bibliography

- [Rabaey2003] J.M. Rabaey, A. Chandrakasan, B. Nikolic, Digital Integrated Circuits: A Design Perspective, 2nd Edition, Prentice Hall, 2003
- [Itoh2001] K.Itoh, VLSI Memory Chip Design, Springer-Verlag, 2001
- [Mamidipaka, 2004] M. Mamidipaka, K.I Khouri, N. Dutt , M. Abadir, Analytical models for leakage power estimation of memory array structures, In: Proceedings of the 2nd IEEE/ACM/IFIP international conference on Hardware/software codesign and system synthesis, Stockholm, Sweden , Sep. 2004.
- [Flautner2002] K. Flautner, N.S. Kim, S. Martin, D. Blaauw, T. Mudge, Drowsy caches: simple techniques for reducing leakage power, In: Proceedings of the 29th annual international symposium on Computer architecture, Anchorage, Alaska, 2002.
- [Powell 2000] M. Powell, S.H. Yang, B. Falsafi, K. Roy, T.N. Vijaykumar, Gated-Vdd: a circuit technique to reduce leakage in deep-submicron cache memories, Proceedings of the 2000 international symposium on Low power electronics and design, pg. 90-95, Rapallo, Italy, 2000
- [Nii1998] K. Nii, H. Makino, Y.Tujihashi, C. Morishima, Y. Hayakawa, H. Nunogami, T. Arakawa, H. Hamano, A low power SRAM using auto-backgate-controlled MT-CMOS, Proceedings of the 1998 international symposium on Low power electronics and design, pg. 293-298, Monterey, California, 1998
- [Wang2007] C.C. Wang, C.L. Lee, W.J. Lin, A 4-kb Low-Power SRAM Design with Negative Word-Line Scheme, IEEE Trans. on Circuits and Systems I: Fundamental Theory and Applications, Vol. 54, No. 5, pp. 1069-1076, May 2007
- [Hanson2003] H. Hanson, M.S. Hrishkesh, V. Agarwal, S.W.Keckler, D. Burger, Static energy reduction techniques for microprocessor caches, IEEE Transactions on Very Large Scale Integration (VLSI) Systems, vol. 11, Issue 3, pg. 303-313, June 2003
- [Kim2002] N.S. Kim, K. Flautner, D. Blaauw, T. Mudge, Drowsy instruction caches: leakage power reduction using dynamic voltage scaling and cache sub-bank prediction, Proceedings of the 35th annual ACM/IEEE international symposium on Microarchitecture, pg. 219-130, Istanbul, Turkey, 2002

- [Chandrakasan 2007] B.H. Calhoun, A.P. Chandrakasan, A 256-kb 65-nm Sub-threshold SRAM Design for Ultra-Low-Voltage Operation, IEEE Journal of Solid-State Circuits, vol. 42, no. 3, pp. 680-688, March 2007
- [Chandrakasan 2008] N. Verma, A.P. Chandrakasan, A 256 kb 65 nm 8T Subthreshold SRAM Employing Sense-Amplifier Redundancy, IEEE Journal of Solid-State Circuits, pp. 141-149, January 2008
- [Calhoun2008] J.Wang, B.H. Calhoun, Techniques to Extend Canary-based Standby VDD Scaling for SRAMs to 45nm and Beyond, IEEE Journal of Solid-State Circuits, Vol. 43, No. 11, pages 2514-2523, November 2008
- [Fish2008] A. Fish, O. Yadid-Pecht, "Considerations for Power Reduction in "Smart" CMOS Image Sensors", by Kris Iniewski, CRC Press, 2008
- [Yadid-Pecht2004] O. Yadid-Pecht and R. Etienne-Cummings, CMOS imagers: from phototransduction to image processing, Kluwer Academic Publishers, 2004
- [Culurciello2004] E. Culurciello and A. G. Andreou, A 16x16 Silicon on Sapphire CMS Photosensor Array With a Digital Interface For Adaptive Wavefront Correction, Proc. ISCAS, Vancouver, May, 2004.
- [Teman2008] A. Teman, S. Fisher, L. Sudakov, A. Fish, O. Yadid-Pecht, Autonomous CMOS image sensor for real time target detection and tracking. Proc. of ISCAS 2008, pg. 2138-2141, 2008
- [Narendra2001] S. Narendra, et. al., "Scaling of Stack Effect and its Application for Leakage Reduction," Proc. of ISLPED 2001, pp. 195-200, 2001
- [Staples2009] C. J.Stapels, P. Barton, E. B. Johnson, D. K. Wehe, P. Dokhale, K. Shah, F. L. Augustine, J. F. Christian, Recent developments with CMOS SSPM photodetectors, Proceedings of the Fifth International Conference on New Developments in Photodetection, Pg. 145-149, October 2009
- [Chandrakasan2006] A. Wang, B.H. Calhoun, A.B. Chandrakasan, Sub Threshold Design For Ultra Low Power Systems, Springer 2006
- [Fish2007] A. Fish, T. Rothschild, A. Hodes, Y. Shoshan and O. Yadid-Pecht, Low Power CMOS Image Sensors Employing Adaptive Bulk Biasing Control (AB2C) Approach, Proc. IEEE International Symposium on Circuits and Systems, pp. 2834-2837, New-Orleans, USA, May 2007.

Authors' Information



Adam Teman – Masters Student, The VLSI Systems Center, Ben Gurion University of the Negev, P.O. Box 653 Be'er Sheva 84105, Israel; e-mail: teman@ee.bgu.ac.il

Major Fields of Scientific Research: VLSI, Digital circuit design, Low power memories and CMOS image sensors.



Dr. Orly Yadid-Pecht – iCore Professor, Director of Integrated Sensors, Intelligent Systems (ISIS), University of Calgary, 2500 University Drive N.W. Calgary, Alberta, Canada T2N1N4, e-mail: orly@atips.ca

Major Fields of Scientific Research: VLSI, CMOS Image Sensors, Neural Networks and Image Processing.



Dr. Alexander Fish – Senior Lecturer, Head of Ultra Low Power Circuits and Systems Lab, The VLSI Systems Center, Ben Gurion University of the Negev, P.O. Box 653 Be'er Sheva 84105, Israel; e-mail: afish@ee.bgu.ac.il

Major Fields of Scientific Research: Ultra-low power VLSI, Low-power image sensors, Mixed signal design.

SYSTEM APPROACH TO ESTIMATION OF GUARANTEED SAFE OPERATION OF COMPLEX ENGINEERING SYSTEMS

Nataliya Pankratova

Abstract: *A system approach to estimation of guaranteed safe operation of complex engineering systems is proposed. The approach is based on timely and reliable detection, estimation, and forecast of risk factors and, on this basis, on timely elimination of the causes of abnormal situations before failures and other undesirable consequences occur. The principles that underlie the strategy of the guaranteed safety of CES operation provide a flexible approach to timely detection, recognition, forecast, and system diagnostic of risk factors and situations, to formulation and implementation of a rational decision in a practicable time within an unremovable time constraint.*

Keywords: *system approach, time constraints, multiple-factor risks, abnormal mode, diagnostics, survivability, serviceability, safety*

ACM Classification Keywords: *H.4.2. INFORMATION SYSTEM APPLICATION: type of system strategy*

Conference topic: *Applied Program Systems*

Introduction

The analysis of failures and accidents reveals their most important causes and the shortcomings of the existing principles of control of serviceability and safety of modern plant and machinery [1, 2]. One of such causes is the specific operation of diagnostic systems intended to detect failures and malfunctions. Such an approach to safety excludes a priori prevention of abnormal mode, which thus may become an abnormality, accident, or a catastrophe. Therefore, there is a practical need to qualitatively change the diagnostic system to account for the principles and structure of control of serviceability and safety of modern complex engineering systems (CESs) under the real conditions of multiple-factor risks.

We believe it expedient to consider a safety control problem as a system problem involving the detection of risk factors whose influence may result in abnormal situations. At the same time it is taken into consideration that safety and functioning control is implemented under conditions of incompleteness and uncertainty of dynamics of the transition of a normal to abnormal mode. The normal mode is not permanent and changes considerably at different stages of a system operation cycle. These conditions are met by the operation of many CES, the modes of which during operation differ fundamentally.

System approach is based on the suggested conceptual foundations of system analysis, multicriterion estimation, and forecast of risk situations. The main idea of the suggested concept consists in the replacement of the typical principle of detection of the operability state turning into the inoperability state based on detection of failures, malfunctioning, faults, and forecast of reliability of an object by a qualitatively new principle [3]. The essence of this principle is the timely detection and elimination of the causes of a possible changeover of an operability state to an inoperability state based on the system analysis of multifactor risks of abnormal situations, a credible estimation of margin of permissible risk for different modes of operation of a complex engineering object, and a forecast of the main operability indicators of an object during the assigned operating period.

The purpose of the paper is to propose a system strategy for the guaranteed safety operation of a CES as a unified complex of a methodology of serviceability and safety and its implementation as a toolkit of technical diagnostics of the CES while in service.

1. Main principles of problem solution and realization strategy

First, note the principal differences between the given problem of guaranteed safe operation of CESs and typical control problems. The main difference is that the initial information about a complex object contains only a small part of information about its state, properties, functioning processes, and operational capability characteristics. This information represents only the state and work characteristics of such objects in normal mode. Undoubtedly, this information is enough for decision making during the complex object control only on the condition that the normal mode continue for a long time. However, in real objects in view of existing technical diagnosis systems, oriented toward failure and malfunction detection, it is impossible to ensure that a malfunction or a failure will not appear within the next 5–10min. It is a priori unknown how much time it will take to repair a malfunction. It may take from a few minutes up to several hours or even days and months. And, consequently, the possible damage is a priori unknown, and thus the safety control system is, essentially, a recorder of information about facts and damage. A fundamentally different system approach can be realized on the basis of the proposed principle of timely detection of causes of abnormal situations, prompt prevention of regular situations from becoming abnormal or emergency, revealing risk factors, predicting basic survivability parameters of an object during a specified period of its operation as fundamentals of the guaranteed safety in the CES dynamics, eliminating the causes of possible nonserviceability based on systems analysis of multiple-factor risks of abnormal situations [3].

The key idea of the strategy is to provide (under actual conditions of the operation of a complex system) timely and reliable detection and estimation of risk factors, prediction of their development during a certain period of operation, and timely elimination of the causes of abnormal situations before failures and other undesirable consequences occur.

Strategy realization is based on the following principles:

- system consistency in purposes, problems, resources, and expected results as to the measures of providing safety operation of the complex system;
- timely detection, guaranteed recognition, and system diagnostics of risk factors and situations;
- prompt forecasting, reliable assessment of abnormal situations;
- formation and realization of a rational solution in a practicable time within an unremovable time constraints.

Let us formulate the mathematical problem of recognizing an abnormal situation in the dynamics of a dangerous industrial object.

For each situation $S_k^\tau \in S_\tau$, the set $M_k^\tau \in M_\tau$ of risk factors is known to be formed as $M_k^\tau = \left\{ \rho_{q_k}^\tau \mid q_k = \overline{1, n_k}^\tau \right\}$. For each risk factor $\rho_{q_k}^\tau \in M_k^\tau$, given are a fuzzy information vector $I_{q_k}^\tau = \left\{ I_{q_k}^\tau \mid q_k = \overline{1, n_k}^\tau; k = \overline{1, K_\tau} \right\}$ and its components:

$$I_{q_k}^\tau = \left\{ \tilde{x}_{q_k j_k p_k}^\tau \mid q_k = \overline{1, n_k}^\tau; j_k = \overline{1, n_{q_k}^\tau}; p_k = \overline{1, n_{q_k j_k}^\tau} \right\},$$

$$\tilde{x}_{q_k j_k p_k}^\tau = \left\langle x_{q_k j_k p_k}^\tau, \mu_{H_{q_k j_k p_k}} \left(x_{q_k j_k p_k}^\tau \right); x_{q_k j_k p_k}^\tau \in H_{q_k j_k p_k}^\tau; \mu_{H_{q_k j_k p_k}} \in [0, 1] \right\rangle,$$

$$H_{q_k j_k p_k}^\tau = \left\langle x_{q_k j_k p_k}^\tau \mid x_{q_k j_k p_k}^- \leq x_{q_k j_k p_k}^\tau \leq x_{q_k j_k p_k}^+ \right\rangle.$$

For each situation $S_k^\tau \in S_\tau$ and each risk factor $\rho_{q_k}^\tau \in M_k^\tau$, $M_k^\tau \in M_\tau$, it is necessary to recognize an abnormal situation in the dynamics of a dangerous industrial object and ensure the survivability of the complex system during its operation.

We will consider the safety of a system as its ability to timely prevent the normal operation mode from becoming an abnormal one, an accident, or a catastrophe based on prompt detection of significant risk and to prevent it from becoming a catastrophic risk. The safety is characterized by the following parameters: the degree of risk η_i which is the probability of undesirable consequences; risk level W_i , which is the damage because of undesirable consequences of any risk factors at any instant of time $T_i \in T^\pm$ during the operation of the complex system; resource of the admissible risk of abnormal mode T_{pr} , which is the period of operation of the complex system in a certain mode during which the degree and level of risk do not exceed a priori specified admissible values. The safety parameters are evaluated by solving the general problem of analysis of multiple-factor risks [4, 5].

The system consistency of the rates of diagnostics and the rates of processes in various operation modes of CES is provided by a unified algorithm for safety control in abnormal situations [6]. This algorithm implements procedures of diagnostics and assessment of abnormal situations during the transition from the normal mode into a sequence of abnormal situations. Based on this, a database and a scenario of a sequence of abnormal situations are formed, and the possibility whereby the complex object passes from abnormal situations to the normal mode is determined.

The strategy of system control of CES serviceability and safety is realized based of a diagnostic unit as an information platform of engineering diagnostics of the CES.

2. Information platform for engineering diagnostics of CES operation

The diagnostic unit, which is the basis of a safety control algorithm for complex objects in abnormal situations, is developed as an information platform (fig. 1).

Let us detail some of these modules of the information platform of engineering diagnostics (IPED).

Data accessing of the Initial Information during CES Operation. By a CES we mean an engineering object consisting of several multi-type subsystems that are system-consistent in tasks, problems, resources, and expected results. Each subsystem has functionally interdependent parameters measured with sensors. To this end, groups of sensors are connected to each subsystem, each having different parameters (time sampling, resolution, etc.), depending on what is its nature.

The engineering diagnostics during the CES operation requires samples of size N_{01} and N_{02} , where N_{01} ($N_{01} \gg 200$) is the total sample size during the CES real-mode operation; N_{02} ($N_{02} \ll N_{01}$; $N_{02} = 40 \div 70$) is the size of the basic sample required to estimate the FDs. The initial information is reduced to a standard form, which makes it possible to form FDs from discrete samples. In view of the proposed methodology, biased Chebyshev polynomials are taken as basic approximating functions, which normalizes all the initial information to the interval [0, 1].

Recovery of Functional Dependences based on Discrete Samples. In the general case, the initial information is specified as a discrete array [7]

$$M_0 = \langle Y_0, X_1, X_2, X_3 \rangle,$$

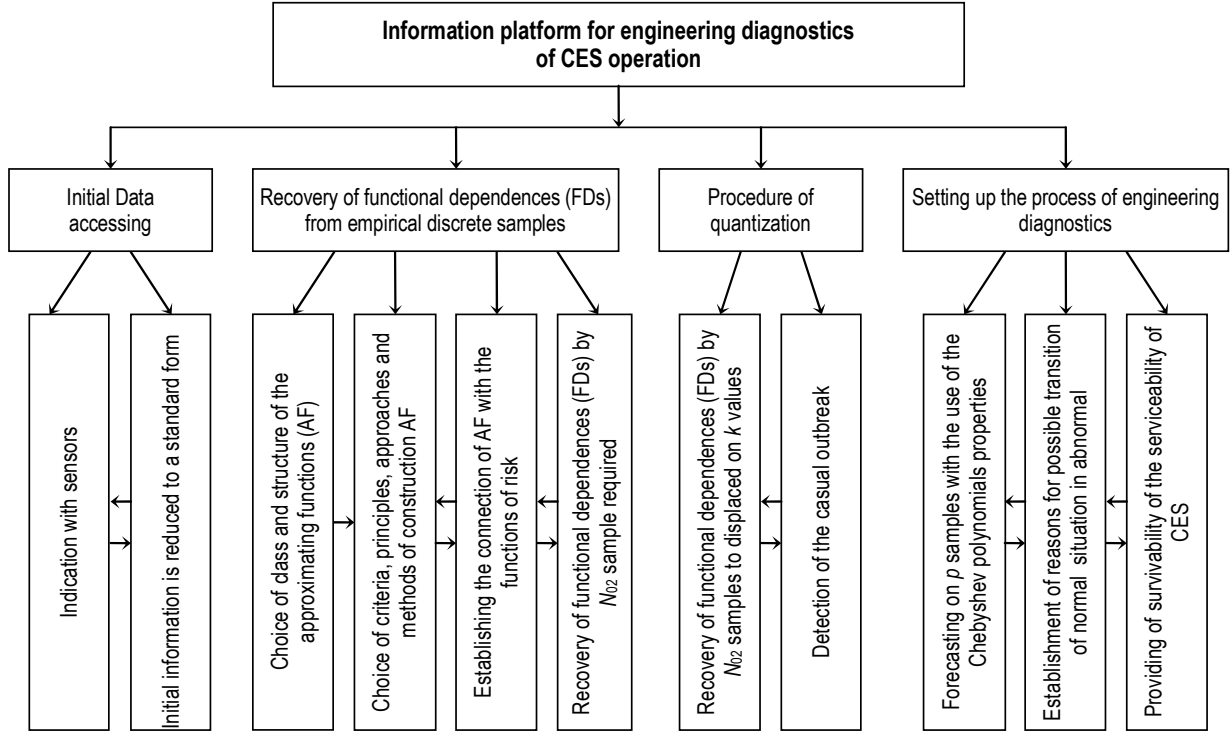


Fig. 1 Structural diagram of information platform for engineering diagnostics

$$Y_0 = (Y_i | i = \overline{1, m}), Y_i = (Y_i[q_0] | q_0 = \overline{1, k_0}), X_1 = (X_{1j_1} | j_1 = \overline{1, n_1}), X_{1j_1} = (X_{1j_1}[q_1] | q_1 = \overline{1, k_1}),$$

$$X_2 = (X_{2j_2} | j_2 = \overline{1, n_2}), X_{2j_2} = (X_{2j_2}[q_2] | q_2 = \overline{1, k_2}), X_3 = (X_{3j_3} | j_3 = \overline{1, n_3}),$$

$$X_{3j_3} = (X_{3j_3}[q_3] | q_3 = \overline{1, k_3}),$$

where the set Y_0 determines the numerical values $Y_i[q_0] \Rightarrow \langle X_{1j_1}[q_1], X_{2j_2}[q_2], X_{3j_3}[q_3] \rangle$ of the unknown continuous functions $y_i = f_i(x_1, x_2, x_3)$, $i = \overline{1, m}$; $x_1 = (x_{1j_1} | j_1 = \overline{1, n_1})$, $x_2 = (x_{2j_2} | j_2 = \overline{1, n_2})$, $x_3 = (x_{3j_3} | j_3 = \overline{1, n_3})$. To each value of $q_0 \in [1, k_0]$ there corresponds a certain set $q_0 \Leftrightarrow \langle q_1, q_2, q_3 \rangle$ of values $q_1 \in [1, k_1]$, $q_2 \in [1, k_2]$, $q_3 \in [1, k_3]$. The set Y_0 consists of k_0 different values $Y_i[q_0]$. In the sets X_1, X_2, X_3 a certain part of values $X_{1j_1}[q_1], X_{2j_2}[q_2], X_{3j_3}[q_3]$ for some values $q_1 = \hat{q}_1 \in \hat{Q}_1 \subset [1, k_1]$, $q_2 = \hat{q}_2 \in \hat{Q}_2 \subset [1, k_2]$, $q_3 = \hat{q}_3 \in \hat{Q}_3 \subset [1, k_3]$ repeats each, but there are no completely coinciding sets $\langle X_{1j_1}[q_1], X_{2j_2}[q_2], X_{3j_3}[q_3] \rangle$ for different $q_0 \in [1, k_0]$. We have also $n_1 + n_2 + n_3 = n_0$, $n_0 \leq k_0$.

It is known that $x_1 \in D_1, x_2 \in D_2, x_3 \in D_3$, $X_1 \in \hat{D}_1, X_2 \in \hat{D}_2, X_3 \in \hat{D}_3$, where

$$D_s = \langle x_{sj_s} | d_{sj_s}^- \leq x_{sj_s} \leq d_{sj_s}^+, j_s = \overline{1, n_s} \rangle, \quad s = \overline{1, 3};$$

$$\hat{D}_s = \langle X_{sj_s} | \hat{d}_{sj_s}^- \leq X_{sj_s} \leq \hat{d}_{sj_s}^+, j_s = \overline{1, n_s} \rangle, \quad s = \overline{1, 3};$$

$$d_{s j_s}^- \leq \hat{d}_{s j_s}^-, \quad d_{s j_s}^+ \geq \hat{d}_{s j_s}^+.$$

It is required to find approximating functions $\Phi_i(x_1, x_2, x_3)$, $i = \overline{1, m}$, that characterize the true functional dependences $y_i = f_i(x_1, x_2, x_3)$, $i = \overline{1, m}$, on the set D_s with a practicable error.

Since the initial information is heterogeneous as well as the properties of the groups of factors under study, which are determined, respectively, by the vectors x_1, x_2, x_3 , the degree of the influence of each group of factors on the properties of approximating functions should be evaluated independently. To this end, the approximating functions are formed as a hierarchical multilevel system of models. At the upper level, the model determining the dependence of the approximating functions on the variables x_1, x_2, x_3 is realized. Such a model in the class of additive functions, where the vectors x_1, x_2, x_3 are independent, is represented as the superposition of functions of the variables x_1, x_2, x_3 :

$$\Phi_i(x_1, x_2, x_3) = c_{i1}\Phi_{i1}(x_1) + c_{i2}\Phi_{i2}(x_2) + c_{i3}\Phi_{i3}(x_3), i = \overline{1, m}. \quad (1)$$

At the second hierarchical level, models that determine the dependence Φ_{is} ($s = 1, 2, 3$) on the components of the variables x_1, x_2, x_3 , respectively, and represented as

$$\begin{aligned} \Phi_{i1}(x_1) &= \sum_{j_1=1}^{n_1} a_{ij_1}^{(1)} \Psi_{1j_1}(x_{1j_1}), \quad \Phi_{i2}(x_2) = \sum_{j_2=1}^{n_2} a_{ij_2}^{(2)} \Psi_{2j_2}(x_{2j_2}), \\ \Phi_{i3}(x_3) &= \sum_{j_3=1}^{n_3} a_{ij_3}^{(3)} \Psi_{3j_3}(x_{3j_3}). \end{aligned} \quad (2)$$

are formed.

At the third hierarchical level, models that determine the functions $\Psi_{1j_1}, \Psi_{2j_2}, \Psi_{3j_3}$ are formed, choosing the structure and components of the functions $\Psi_{1j_1}, \Psi_{2j_2}, \Psi_{3j_3}$ being the major problem. The structures of these functions are similar to (2) and can be represented as the following generalized polynomials:

$$\Psi_{sj_s}(x_{j_s}) = \sum_{p=0}^{P_{j_s}} \lambda_{j_s p} \varphi_{j_s p}(x_{sj_s}), s = 1, 2, 3. \quad (3)$$

In some cases, in forming the structure of the models, one should take into account that the properties of the unknown functions $\Phi_i(x_1, x_2, x_3)$, $i = \overline{1, m}$, are influenced not only by a group of components of each vector x_1, x_2, x_3 but also by the interaction of their components. In such a case, it is expedient to form the dependence of the approximating functions on the variables x_1, x_2, x_3 in a class of multiplicative functions, where the approximating functions are formed by analogy with (1)-(3) as a hierarchical multilevel system of models

$$\begin{aligned}
[1 + \Phi_i(x)] &= \prod_{s=1}^{S_0} [1 + \Phi_{is}(x_s)]^{c_{is}} ; [1 + \Phi_{is}(x_s)] = \prod_{j_s=1}^{n_s} [1 + \Psi_{sj_s}(x_{sj_s})]^{a_{ij_s}^s} ; \\
[1 + \Psi_{sj_s}(x_{sj_s})] &= \prod_{p=1}^{P_{j_s}} [1 + \varphi_{j_s p}(x_{sj_s})]^{\lambda_{j_s p}} .
\end{aligned} \tag{4}$$

We will use the Chebyshev criterion and for the functions $\varphi_{j_s p}$, we will use biased Chebyshev polynomials $T_{j_s p}(x_{j_s p}) \in [0, 1]$. Then the approximating functions are found based on the sequence $\Psi_1, \Psi_2, \Psi_3 \rightarrow \Phi_{i1}, \Phi_{i2}, \Phi_{i3} \rightarrow \Phi_i$ which will allow obtaining the final result by aggregating the corresponding solutions. Such an approach reduces the procedure of forming the approximating functions to a sequence of Chebyshev approximation problems for inconsistent systems of linear equations [8, 9].

Due to the properties of Chebyshev polynomials, the approach to forming the functional dependences makes it possible to extrapolate the approximating functions set up for the intervals $[\hat{d}_{j_s}^-, \hat{d}_{j_s}^+]$ to wider intervals $[d_{j_s}^-, d_{j_s}^+]$, which allows forecasting the analyzed properties of a product outside the test intervals.

Quantization of Discrete Numerical Values. The quantization is applied in order to reduce the influence of the measurement error of various parameters on the reliability of the solution being formed. The procedure of quantization of discrete numerical values is implemented as follows.

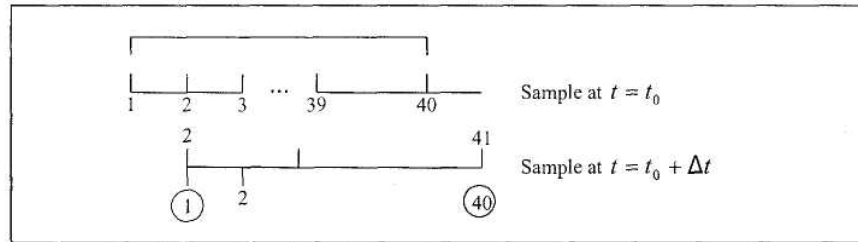


Fig. 2. Sample at $t = t_0$ and $t = t_0 + \Delta t$

As the base reference statistic for each variable $x_1, \dots, x_n, y_1, \dots, y_m$, the statistic of random samples in these variables of size $N_{01} \geq 200$ is taken.

As the base dynamic statistic in the same variables, the statistic of the sample of the dynamics of the object for the last N_{02} measurements is taken. Therefore, the very first measurement of the original sample should be rejected and measurements should be renumbered in the next measurement $N_{02} + N_2$. Figure 2 schematizes the sample for the instant of time $t = t_0, N_{02} = 40$ and $t = t_0 + \Delta t (t = 1, 2, 3, \dots, t_k, \dots, T)$.

For the current dynamic parameters, we take the statistics of samples of size $N_{02} + N_2$ biased by N_2 with respect to the statistics of samples of size N_{02} .

Processing of each sample in each variable should involve the following procedures:

— evaluating $\Phi_i^k(x_1^k, \dots, x_j^k, \dots, x_n^k), i = \overline{1, m}, j = \overline{1, n}$, in the form (2) or (4) for each k -th measurement for $t = t_k$;

— setting up a step function of the first level

$$Z_{1j} = \sum_{p=1}^{M_1} d_{jp} U(\hat{x}); U(\hat{x}) = \begin{cases} 0, & \hat{x} < 0 \\ 1, & \hat{x} \geq 0, \end{cases} \hat{x} = x_j - x_p; d_{jp} = \begin{cases} 0,15 & \text{if } p = 1, \\ 0,1 & \text{if } p = \overline{2,9}, \\ 1 & \text{if } p = 10, \end{cases}$$

$$x = 0,1 \cdot p; p = \overline{1,10}; M_1 = 10;$$

— setting up a step function of the second level

$$Z_{2j} = \sum_{p=1}^{M_2} d_{jp} U(\hat{x}); M_2 = 10; x_p^0 = \begin{cases} 0,5 & \text{if } p = 1, \\ 0,1 & \text{if } p = \overline{2,9}, \\ 1,15 & \text{if } p = 10, \end{cases} d_{jp} = \begin{cases} 0,1 & \text{if } p = \overline{1,9}, \\ 1 & \text{if } p = 10. \end{cases}$$

Forecasting Nonstationary Processes. The models for predicting nonstationary processes are based on the original sample of the time series for the initial interval D_0 and base dynamic model of processes (1)-(3). To this end, we will use the well-known property of Chebyshev polynomials that functions are uniformly approximated on the interval $[0, 1]$. The essence of the approach is as follows. The initial data are normalized for the interval $D = \{t | t_0^- \leq t \leq t^+\}$, $D = D_0 \cup D_0^+$, which includes the initial observation interval $D_0 = \{t | t_0^- \leq t \leq t^+\}$ and the prediction interval $D_0^+ = \{t | t_0^+ < t \leq t^+\}$. Then, to determine the dynamic model of the processes as the estimated approximating functions (1) or (4) based on the initial data, the system of equations is formed for the interval D_0 as follows:

$$0,5a_0 + \sum_{n=n_1}^N a_n T_n^*(\tau_{k_1}) - \hat{y}_{k_1} = 0; k_1 = \overline{1, K_0}; \quad (5)$$

$$\hat{y}_{k_1} = y(\tau_{k_1}), \tau_{k_1} \in [0, \tau_{k_1}^+], \tau_{k_1}^+ = \frac{t_k^+ - t_0^-}{t^+ - t_0^-} < 1, t_k \in D_{K_0}, D_{K_0} \subset D_0.$$

The dynamic model of the process within the observation interval D_0 is determined by solving system (5) and is described by

$$\Phi_1(\tau_1) = 0,5a_0^0 + \sum_{n=n_1}^N a_n^0 T_n^*(\tau_1); \tau_1 \in D_0 \quad (6)$$

The dynamic forecasting model is based on the extrapolation of function (6) to the interval D_0^+ and is expressed by the formula

$$\Phi_2(\tau_2) = 0,5a_0^0 + \sum_{n=n_1}^N a_n^0 T_n^*(\tau_2); \tau_2 \in D_0^+. \quad (7)$$

The dynamic model of the process within the given interval $D = D_0 \cup D_0^+$ based on (6) and (7) is described by the

$$\Phi_0(\tau) = \begin{cases} \Phi_1(\tau_1) & \text{if } \tau_1 \in D_0, \\ \Phi_2(\tau_2) & \text{if } \tau_2 \in D_0^+, \end{cases}$$

Models for long-term and short-term forecast differ by both the ratio of observation and forecast intervals and the order of the Chebyshev polynomials used in the model.

Setting up the Process of Engineering Diagnostics. We will use the system of CES operation models to describe the normal operation mode of the object under the following assumptions and statements.

Each stage of CES operation is characterized by the duration and by the initial and final values of each parameter y_i determined at the beginning and the end of the stage, respectively. The variations of y_i within the stage are determined by the corresponding model.

All the parameters y_i are dynamically synchronous and inphase in the sense that they simultaneously (without a time delay) increase or decrease under risk factors.

The control $U = (U_j | j = \overline{1, m})$ is inertialess, i.e., there is no time delay between the control action and the object's response.

The risk factors $\rho_{q_k}^\tau | q_k = \overline{1, n_k^\tau}$ change the effect on the object in time; the risk increases or decreases with time.

The control can slow down the influences of risk factors or stop their negative influence on the controlled object if the rate of control exceeds the rate of increase in the influence of risk factors. The negative influence of risk factors is terminated provided that the decision is made and is implemented prior to the critical time T_{cr} . At this moment the risk factors cause negative consequences such as an accident or a catastrophe.

To analyze an abnormal mode, let us introduce additional assumptions as to the formation of the model and conditions of recognition of an abnormal situation.

The risk factors $\rho_{q_k}^\tau | q_k = \overline{1, n_k^\tau}$ are independent and randomly vary in time with a priori unknown distribution.

The risk factors can influence several or all of the parameters y_i simultaneously. A situation of the influence of risk factors is abnormal if at least two parameters y_i simultaneously change, without a control, their values synchronously and in phase during several measurements (in time).

The influence of risk factors will be described as a relative change of the level of control. The values of each risk factor vary discretely and randomly.

Based on acceptable assumptions, let us present additional models and conditions to detect an abnormal situation. Denote by \tilde{y}_i the value of the parameter y_i influenced by the risk factors; $F_i(\rho_{q_k})$ is the function that takes into account the level of influence of the risk factors on the i th parameter y_i ; ρ_{q_k} is the value of the q th risk factor at the instant of time t_k .

According to item 8, we assume that the value of $\tilde{y}_i[t_k]$ at the instant of time t_k is determined by

$$\tilde{y}_i[t_k] = \frac{1}{m} \sum_{j=1}^m \tilde{b}_{ij} \sum_{r=0}^{R_j} a_{jr} T_r^*(U_j); \tilde{b}_{ij} = b_{ij} \cdot F_i(\rho_{q_k}) \quad (8)$$

where the function $F_i(\rho_{q_k})$ should correspond to the condition whereby $\tilde{y}_i = y_i$ in the absence of the influence of risk factors (i.e., for $\rho_{q_k} = 0$). Therefore, one of the elementary forms of the function $F_i(\rho_{q_k})$ is

$$F_i(\rho_{q_k}) = 1 - \prod_{q_k=1}^{n_{qk}} (1 - c_{iq_k} \rho_{q_k}) \quad (9)$$

Note that risk factors can vary in time continuously (for example, pressure continuously changes as an aircraft lifts) or abruptly (for example, during cruise flight at a certain height, pressure may change abruptly at the cyclone-anticyclone interface). The most complex is the case where one risk factors vary continuously and others abruptly.

We will recognize risk situations by successively comparing $\tilde{y}_i[t_k]$ for $\tilde{y}_i[t_k]$ for several successive values of $t_k, k = \overline{1, k_0}$, where $k_0 = 3 \div 7$. As follows from item 2 of the assumptions, the condition of a normal situation is synchronous and inphase changes of \tilde{y}_i for several (in the general case, for all) parameters, whence follows a formula for different instants of time t_k for all of the values of i and for the same instants of time t_k for different values of i (different parameters):

$$\text{sign}\Delta\tilde{y}_i[t_1, t_2] = \dots = \text{sign}\Delta\tilde{y}_i[t_k, t_{k+1}] = \dots = \text{sign}\Delta\tilde{y}_i[t_{k_0-1}, t_{k_0}], \quad (10)$$

$$\text{sign}\Delta\tilde{y}_1[t_k, t_{k+1}] = \dots = \text{sign}\Delta\tilde{y}_i[t_k, t_{k+1}] = \dots = \text{sign}\Delta\tilde{y}_n[t_k, t_{k+1}], i = \overline{1, n}. \quad (11)$$

As follows from (10) and (11), given an abnormal situation on the interval $[t_1, t_{k_0}]$, the following inequalities hold simultaneously:

- the inequality of the signs of increment $\Delta\tilde{y}_i$ for all the adjacent intervals $[t_k, t_{k+1}]$ for $k = \overline{1, k_0}$ for each parameter $\tilde{y}_i, i = \overline{1, n}$;
- the inequality of the signs of increment $\tilde{y}_i, i = \overline{1, n}$, for all of the parameters \tilde{y}_i for each interval $[t_k, t_{k+1}], k = \overline{1, k_0}$.

Conditions (10) and (11) are rigid; for practical purposes, it will suffice to satisfy the conditions for the representative number (3-5), which determine the parameters \tilde{y}_i but not for all parameters i . The corresponding quantities in (10) and (11) are defined by

$$\Delta\tilde{y}_i[t_k, t_{k+1}] = \tilde{y}_i[t_{k+1}] - \tilde{y}_i[t_k], \quad (12)$$

where $\tilde{y}_i[t_k]$ are defined by (8); we assume that $\rho_{q_k}[t_{k+1}] > \rho_{q_k}[t_k]$ i.e., the dependence of each risk factor is a function of time, which increases, or $\rho_{q_k}[t_{k+1}] < \rho_{q_k}[t_k]$ i.e., the dependence is a decreasing function.

The practical importance of recognizing an abnormal situation based on (10) and (11) is in the minor alteration of $\tilde{y}_i[t_k]$ subject to risk factors since the “indicator” of the change is the sign of the difference in (10) and (11) rather than the value defined by (12). In other words, such an approach is much more sensitive than typical approaches used in diagnostics. Moreover, it allows “filtering” random changes and random measurement errors \tilde{y}_i for separate i according to (10) or for individual $[t_k, t_{k+1}]$ according to (11).

3. Diagnostic of the circulating water system

A real circulating water system (with functional circuit shown on Fig. 3) is examined as an example of system strategy implementation for the guaranteed safety of the CTS functioning. The main purpose of the system: ensuring given water consumption $Q_1 \leq 0.045 \text{ m}^3/\text{s}$ for cooling of the technical plant (TP) (priority object) and additional non-critical consumption $Q_2 \leq 0.045 \text{ m}^3/\text{s}$ used by the support equipment. The full regular consumption $Q_n = Q_1 + Q_2$.

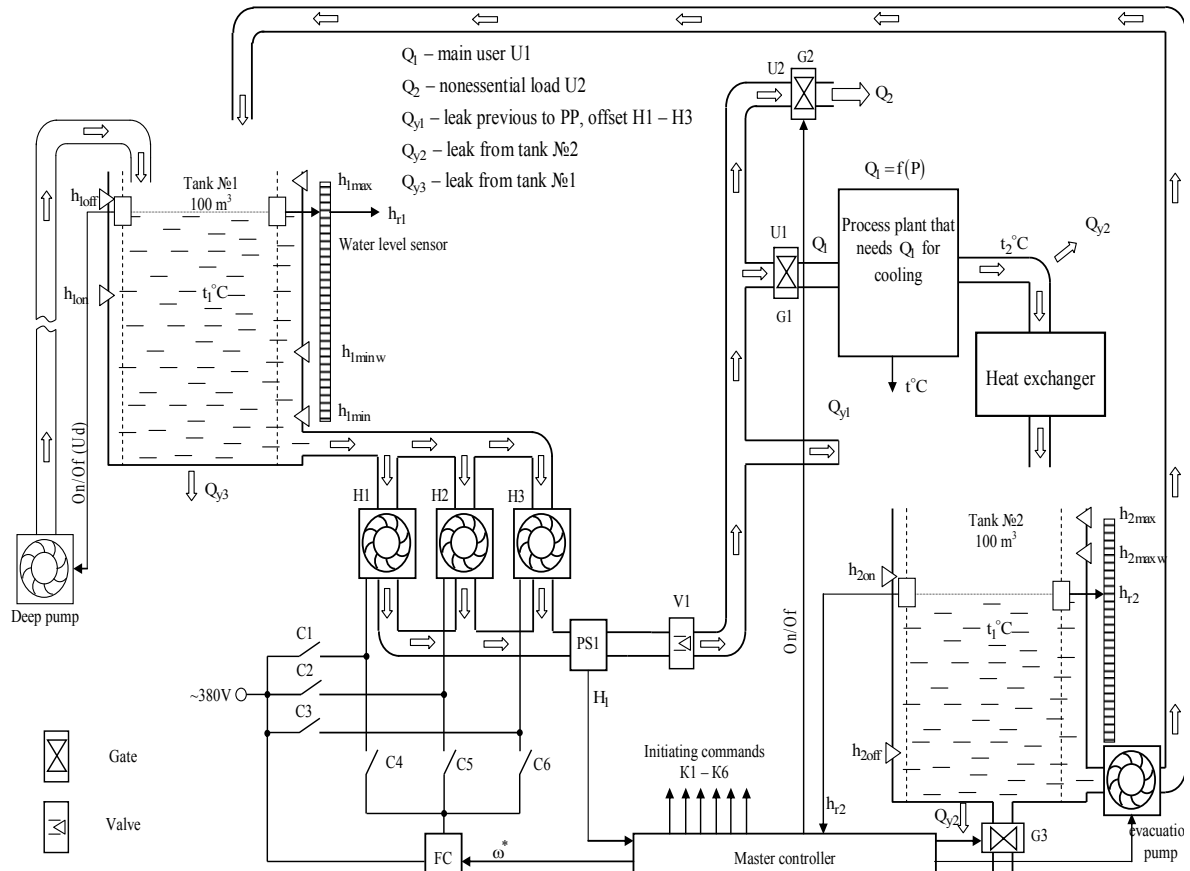


Fig. 3. A function chart of a deep water supply system

The circulating water system consists of:

- deep-well replenishment pump;
- pump installation of three pumps, two of which are unregulated;
- pressure sensor (PS1) at the pump installation output (for the feedback of the pressure stabilization system);
- regulated gates G1, G2, G3;
- valve V1;
- water cooling system of the technological process;

- heat exchanger intended for cooling of the discharge water (e.g. water-cooling tower)
- output tank №2 for the discharge water collection;
- pump for return of the discharge water to the tank №1;
- operation controller (K1-K6);
- frequency converter (FC) for the regulated pump rotating velocity control;
- switching equipment for engaging and disengaging of unregulated pumps.

Engaging the replenishment pump occurs: if the water level in the tank №1 is lower than $h_{1on} = 50 \text{ m}^3$, disengaging given $h_{1off} = 80 \text{ m}^3$, if the return pump is off; if the water level is lower than $h_{1on} = 40 \text{ m}^3$, disengaging given $h_{1off} = 60 \text{ m}^3$, if the return pump is on. When engaged, the deep-well pump provides pump delivery $Q_{dn} = 0.05 \text{ m}^3/\text{s}$. A water level sensor is installed in the tank №1, providing the h_{r1} signal. Lowering of the water level to less than 40 m^3 with engaged return pump means an abnormal mode (leak).

The water return pump is engaged when the water level in the tank №2 is higher than $h_{2on} = 80 \text{ m}^3$, disengaging when lower than $h_{2off} = 20 \text{ m}^3$. The return pump provides pump delivery $Q_{dn2} = 0.05 \text{ m}^3/\text{s}$. A water level sensor is installed in the tank, providing the h_{r2} signal.

If the water level in the tank №1 is higher than 97 m^3 and the return pump is on, it is forced to disengage to avoid an overflow of the tank №1. Its operation is recommenced when the h_{r1} level decreases to 95 m^3 .

Only the water run through the technical plant and the cooling system is put into the tank №2. After the second usage water is discharged into a sewer system.

The pump installation of the three forcing pumps (P1 – P3) is mounted to supply water to the consumers: two pumps work in regular mode, one is the emergency pump. Each pump productivity is $Q_{nn} = 0.05 \text{ m}^3/\text{s}$. The pump installation works in pressure stabilization mode, which is ensured by the pressure sensor PS1. Water comes through the valve V1 into the mains system.

Overall water consumption over a small measurement time interval T_s can be determined from the tank water level change by the formula (under the hypothesis that the deep-well pump condition is invariable during the measurement time interval):

$$Q = 3600 \left(h_{r(k-1)} - h_{r(k)} \right) / T_s + U_{d1} Q_{dn1} + U_{d2} Q_{dn2}, k = 1, 2 \dots n.$$

Two consumer groups (U1,U2) with maximum regular consumption levels $Q_1 \leq 0.045 \text{ m}^3/\text{s}$, $Q_2 \leq 0.045 \text{ m}^3/\text{s}$ are connected to the mains system. The water supply to the consumer U2 is not a critical factor and can be cut off by closing the regulated valve G2. The water supply to the technical plant (consumer U1) is compulsory, its failure leads to a accident.

The TP water supply needs in a regular mode are fully satisfied, i.e. necessary heat removing is assured and the TP temperature is in the permissible limit.

Temperatures up to $75 \text{ }^\circ\text{C}$ are considered the permissible limit. A situation is abnormal under temperatures in range from $75 \text{ }^\circ\text{C}$ до $85 \text{ }^\circ\text{C}$ and emergency after the $85 \text{ }^\circ\text{C}$ threshold.

The temperature is regulated by a separate PI controller, outputting the hydraulic resistance value of the gate G1 at the TP input, which, in turn, controls the water consumption in TP.

In compliance with the requirements of the developed IPED instrument, throughout the bottom of the tanks and in a number of reference points of the water system sensors were installed, providing measures every 20 seconds, modeling time – 10000 s. The measures of the water level sensors h_{r1} и h_{r2} in the tanks №1 and №2, respectively, the head H_1 at the input of the technical plant, the temperature T of the technical plant and their arguments are provided during 10000 seconds (500 samples).

Real-time monitoring of the technical diagnostics is conducted in the water system operation process with the purpose of timely exposure of potentially possible abnormal situations and guaranteeing the serviceability of the system's functioning. In compliance with the developed methodology of the guaranteed CTS functioning safety at the starting phase $t = t_0$, functional recovery $y_i = f_i(x_1, \dots, x_j, \dots)$ is performed using $N_{02} = 50$ given discrete samples of values h_{r1} , H_1 , T , h_{r2} and their arguments. Here $y_1 = h_{r1}(x_{11}, x_{12}, x_{13})$, $y_2 = H_1(x_{21}, x_{22}, x_{23})$, $y_3 = T(x_{31}, x_{32}, x_{33}, x_{34}, x_{35})$ and $y_4 = h_{r2}(x_{41}, x_{42}, x_{43}, x_{44})$, where x_{11} is the overall water consumption; x_{12} - deep-well pump productivity; x_{13} - return pump productivity; x_{21} - overall water consumption; x_{22} - number of engaged pumps (1 to 3); x_{23} - regulated pump velocity; x_{31} - cooling water temperature; x_{32} - pressure H_1 ; x_{33} - heat loss ΔP ; x_{34} - given TP temperature; x_{35} - hydraulic resistance value at the TP input; x_{42} - return pump productivity; x_{43} - valve G3 state (open valve provides water leak at 0.1 m³/s speed); x_{44} - water pressure H_1 .

Abnormal mode is caused by the heat exchanger malfunction, i.e. heated in TP water is not cooled to the necessary temperature but is returned to the tank №1. This leads to the cooling water temperature of 57 °C at 3000 s time, significantly reducing its cooling quality. As a result, the TP temperature increases. At some moments the gate G1 is open (minimal hydraulic resistance value), providing maximum possible water flow into TP. As these factors could cause TP overheat, a decision is made at 3000 s time to open the water drain gate G3 and to disengage the return pump. The deep-well replenishment pump injects cold water at 20 °C temperature that gradually lowers the cooling water temperature and ceases abnormal temperature situation. At the same time the water level in the tank №1 is close to 50 m³, which could potentially lead to an emergency situation caused by the water level in the tank №1 in case of an additional leak or a deep-well pump shutdown. Consumer U2 can be cut off if necessary. The complete escape from abnormal situation is possible only after recommencing the heat exchanger's work and engaging the return pump. The heat exchanger is assumed to be repaired at 5000 second, closing the gate G3 and permitting to engage the return pump when the water level in the tank №2 reaches h_{2on} . This happens at the time 8060 s, when the return pump starts to pump warm water into the tank №1, causing gradual cooling water temperature increase to 28 degrees.

Some results of the monitoring are presented as a time distribution of the estimated functional dependences $Y1$ and $Y4$ of the water level in the tanks №1 and №2 $Y2$, of the water head $H1$ and $Y3$ of the temperature of TP.

During the operation of the circulating water system, a data panel displays the diagnostic process quantitatively and qualitatively and the operator obtains the timely preliminary information on the possible transition of the water level h_{r1} , h_{r2} in the tanks №1 and №2, respectively, the head H_1 at the input of the technical plant, the temperature T of the technical plant to an abnormal mode. This allows detecting the cause in due time and making a decision on eliminating the abnormal situation, accident or catastrophe.

The analysis shows that the monitoring of the circulating water system operation by using the developed methodology of the guaranteed safety of CES operation allows making a real-time decision to ensure the survivability of the serviceability of the deep water supply system.

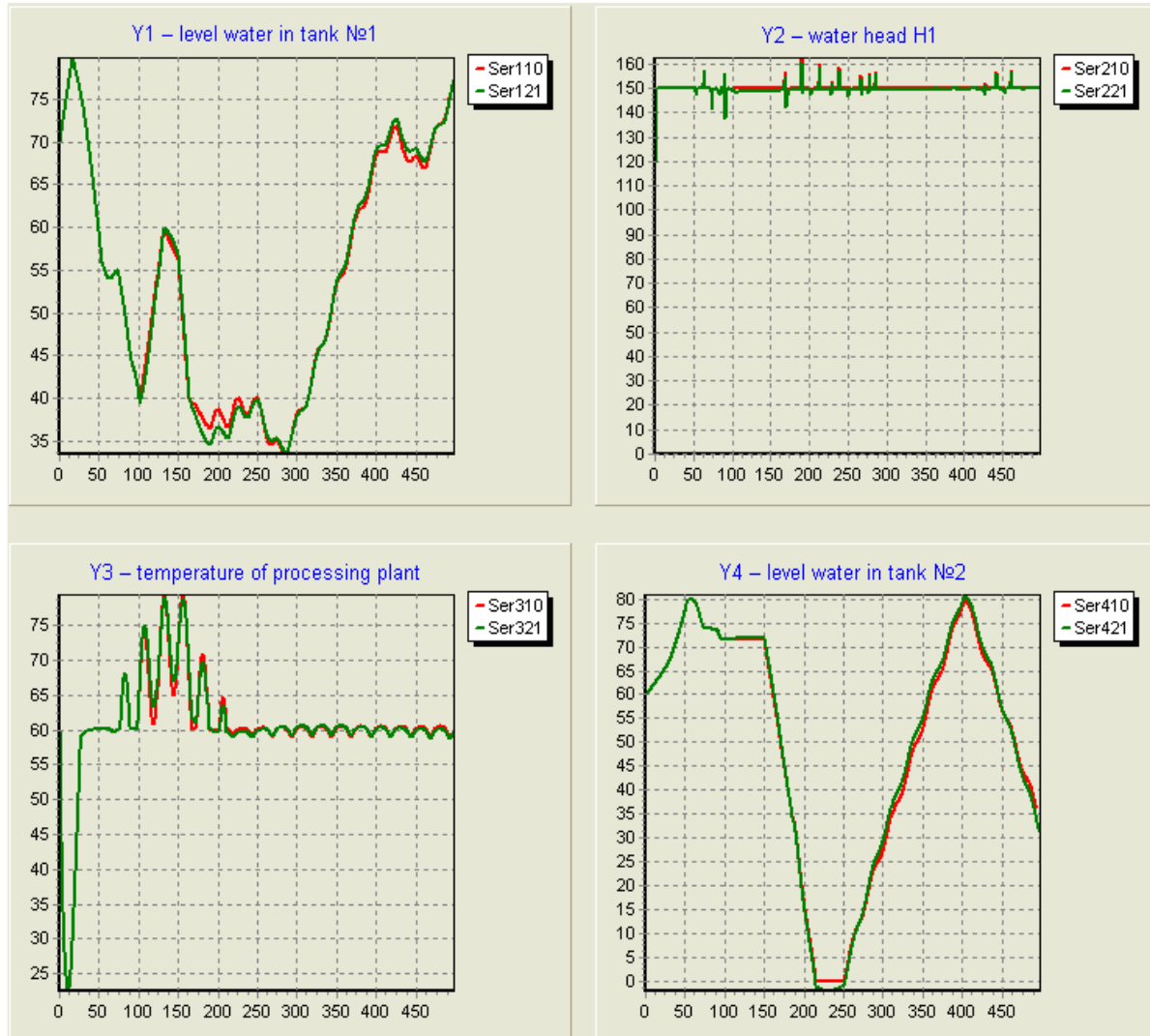


Fig.4. Distribution of the functional dependences Y1 and Y4 of the water level in the tanks №1 and №2; Y2 of the water head H1 and Y3 of the temperature of TP

Conclusion

The proposed approach to estimation of guaranteed safe operation of complex engineering systems implemented as an IPED toolkit prevents the inoperativeness and abnormal situations. The real-time complex, system, and continuous estimation of the parameters of object operation detects situations that can bring the object out of the normal-mode operation. The simultaneous monitoring and integrated estimation of the parameters of a finite number of functionally dynamic parameters allow detailing the processes of object operation of any order of complexity. For situations that may cause deviations of the parameters from the normal mode of object operation,

a timely decision can be made to change the mode of operation or to artificially correct some parameters to make the operation survivable. The principles that underlie the strategy of the guaranteed safety of CES operation provide a flexible approach to timely detection, recognition, prediction, and system diagnostic of risk factors and situations, to formulation and implementation of a rational decision in a practicable time within an unremovable time constraint.

Bibliography

- [1] Frolov K. V. (gen. ed.), Catastrophe Mechanics [in Russian], Intern. Inst. for Safety of Complex Eng. Syst., Moscow — 1995. —389 p.
- [2] Troshchenko V. T. (exec. ed.), Resistance of Materials to Deformation and Fracture: A Reference Book, Pts. 1, 2 [in Russian], Naukova Dumka, Kyiv. —1993, 1994. —702 p.
- [3] Zgurovsky M. Z., Pankratova N. D., System Analysis: Theory and Applications, Springer, Berlin. —2007. — 475 p.
- [4] Pankratova N. and Kurilin B., Conceptual foundations of the system analysis of risks in dynamics of control of complex system safety. P. 1: Basic statements and substantiation of approach // J. Autom. Inform. Sci. —2001. —33, №. 2. —P. 15-31.
- [5] Pankratova N. and Kurilin B., Conceptual foundations of the system analysis of risks in dynamics of control of complex system safety. P. 2: The general problem of the system analysis of risks and the strategy of its solving //J. Autom. Inform. Sci. — 2001. —33, No. 2. —P1-14.
- [6] Pankratova N. D., System analysis in the dynamics of the diagnostic of complex engineering systems //Syst. Doslidzh. Informats. Tekhnol. — 2008. —No. 1. —P 33-49.
- [7] Pankratova N. D. A rational compromise in the system problem of disclosure of conceptual uncertainty //Cybern. Syst. Analysis. — 2002. —38, No. 4. —P. 618-631.
- [8] Lanczos C., Applied Analysis, Prentice-Hall, Englewood Cliffs, N. J. . —1956. —524 p.
- [9] Remez E. Ya., Foundations of Numerical Methods of Chebyshev Approximation, Naukova Dumka, Kyiv. - 1969. — 624 p.

Authors' Information



Nataliya Pankratova – *Depute director of Institute for applied system analysis, National Technical University of Ukraine “KPI”, Av. Pobedy 37, Kiev 03056, Ukraine; e-mail: natalidmp@gmail.com*

Major Fields of Scientific Research: System analysis, Mechanics of solid body, Applied mechanics, Applied mathematics, System information technology in education.

SOA PROTOCOL WITH MULTIRESLTING

Michał Plewka, Roman Podraza

Abstract: *This paper presents a framework for distributed results in SOA environment. These distributed results are partially produced by a service and passed to a client program on demand. This approach is a novelty in SOA technology. Motivations for this new design is figured out and a sketch of implementation is outlined. A configuration for the new tool is suggested and finally its efficiency is compared to a conventional web service implementation.*

Keywords: *SOA, multiresulting, client-server, web services.*

ACM Classification Keywords: *D.2.11 Software Architectures*

Introduction

By multiresulting we understand a case when a remote service is returning a collection of data. Usually the client program has to wait until the service is completed to obtain the complete result. In a number of applications the server processes a huge amount of data what makes the waiting time very long. However in many cases it would be possible to start processing if the client program received only part of results returned by a remote method. So it will be desirable if the remote service could return collection of data in number of steps. For example an attempt to migrate a knowledge discovering system to Service Oriented Architecture focused our attention on the issue of dealing with processing of large collections of data in a distributed environment. A client program usually organized a processing as a sequence of calls to remote services and in many cases the services could be used in a pipeline style if the data could be processed in a continuous or quasi-continuous manner. So we decided to it create such a framework which will allow to return partial results on demand of the consumer of remote services. We named these partial results as multiresults (and the whole approach as multiresulting) to underline their usefulness when they are available separately. Currently remote services are obtainable mainly with technology of web services. They are easy to develop, to maintain, and what is very important, easy to reuse on many different machines thanks to SOAP (Simple Object Access Protocol). Our framework proposal should not be inferior to the known and accepted SOA concepts, but it should provide a satisfactory remedy to the presented drawbacks of existing approach.

Architecture

The prepared mechanism has to work asynchronously and should return data in such a way which allows to split it to many parts. The platform should be used as an independent component which works in a transparent way. The solution obviously has to contain two main layers - the first one is the server side and the second is concerned with the client layer. How the communication between these layers should work? Protocol HTTP was chosen, due to its simplicity. This kind of protocol is often associated with a servlet technology [B. Basham, 2004]. A client sends information about full qualified class and its method name which should be initiated on the server side. Arguments are sent in serialized form as an array of various elements. Then on the server side the request is being intercepted by the specified servlet which invokes method on an Enterprise Java Bean (EJB) [P. Debu, 2007]. In fact, the EJB deserializes arguments and invokes method in specified class using the reflection mechanism. Then a unique identifier is returned to the client. In next iterations the client layer must use this identifier to retrieve the data from the remote method. In the mean time the remote method is being executed on

the server and the results are being collected by a queue. When the client asks the server for a part of data available results are being retrieved from the appropriate queue recognized with the help of the identifier which was generated in the first client-server interaction. Finally, this identifier should be used to identify the correct process. When the remote method is completed then the EJB is notified that the service has been accomplished no more partial results will be returned. The rest of results in the queue has to be returned to the client and with the last portion a special flag is being set to indicate that no more active pooling is required from the client side. It is very important to set a time interval for client interactions. If this value is too small the system performance may be degraded by an inefficient network traffic. Moreover, the time interval for the client-server interactions influences how often the remote method delivers the partial results. If this interval is too short it should be adjusted to a higher value to get more data per interaction.

Implementation

This paragraph presents a framework, which was designed and implemented to support multiresulting as a new way of client-server communication. Some procedures how it should be used are proposed and explained.

A developer should prepare a remote method (or service) firstly. The class which contains it must extend class *MultiResultProcess*, which is provided in the framework. The remote methods producing multipart results have to be public, void and annotated with predefined annotation *@MultiResultMethod*. The annotation has a single parameter *maxElementsPerInteraction*. It defines a maximum number of collection elements which could be returned in a single interaction.

When a partial result is ready to be passed to the client a special method must be invoked by the remote service itself. This method (*newDataArrived(Serializable[])*) takes an array of serializable elements as its parameter. The method is inherited from parent class and it performs all actions required to process the available partial results. The function cannot use just the *return* statement, because it still has to produce further results. Finally, when the whole operation is finished another inherited method (*endProcess()*) should be invoked. It notifies the container controlling the remote call, that the remote method has been finished successfully. Each remote method supporting multiresulting mode of communication has to be completed in this way. If method *endProcess()* is invoked for a given remote method more than once its successive call will be futile and an appropriate exception is thrown.

When a remote method has been prepared it can be installed in Java Enterprise Application context. This could be done by creating dynamic web project and modifying *application.xml* to use the specified multiresult EJB and map the servlet to a required path. File *web.xml* should contain entries which map class *ResponseServlet* to a specified path. It could be done using tags *<servlet>* and *<servlet-mapping>*. The full enterprise application configuration is described elsewhere.

Finally when correct paths are set and libraries are imported, the client side may be prepared. To invoke a multiresult method deployed on a remote server, class *RequestProcessorFactory* has to be used. This class enables producing objects supporting remote method calls. Class *RequestProcessorFactory* has method *initiateProcessing(String)*, which returns object of type *RequestProcessor* enabling method initiation. As a parameter the method gets a service name which is defined in a configuration file presented in the next paragraph. This object of *RequestProcessor* has a method called *Process(Serializable...)*, which handles the whole communication with the server side. To track information about availability of partial results produced by a multiresult remote method a special listener (*DataListener*) must be set. It has two methods, first one receives portions of the remote results and the second one is invoked when the remote method is terminated.

Server side has to be deployed in JEE Application Server environment, so if one decides to use it, the server must be also provided. The installation is trivial and does not differ from any normal application installation.

Configuration

Configuration of the framework is very simple. File *multiResultConfig.xml* should be created in the package resources. To configure a service firstly it is necessary to define the root element `<configuration>`. It contains number of `<service>` sections. Each of them contains the following tags:

- `<name>` - defining name of the remote service,
- `<targetServletEndpoint>` - specifying location of the remote servlet,
- `<processor>` - fully qualified name of class handling the communication between a client and the remote service working in multiresult mode; standard implementation provides class *SimpleRequestProcessor*, however some elaborated versions may be developed by extending class *RequestProcessor* and applied,
- `<qualifiedClassName>` - fully qualified name of class containing the remote method,
- `<methodName>` - name of the remote method,
- `<delay>` - integer number specifying minimal delay (in milliseconds) between successive interactions of the client and the remote method.

The client program can have configured many remote methods working in the multiresult mode. They can be deployed in many different locations and the client program may organize their parallel execution receiving partial results as quickly as they are ready. It is worth mentioning that by receiving partial results the client program is notified that the remote methods are in a continuous progress. This may be a very important feature in organizing distributed processing with some backup capabilities.

Efficiency

The presented framework of multiresulting was compared to classics web services. A test benchmark included searching data using some specified criteria. Amount of data to be searched was unknown as well as amount of results satisfying the criteria, so it was impossible to assess expected time of execution. This benchmark was quite suitable for applying multiresulting, because the data matching the criteria could be returned partialy to the client. The efficiency of multiresulting appeared to be very good. The overall time of benchmark execution was smaller for the multiresulting mode. It was not obvious result because more communication overhead is required in this way of processing. What is very important the first response from the remote service came just after the request and then the client could retrieved the first set of data. In the web service case a large amount of data was returned in a single block. In general for a number of tests the multiresulting mode was approximately 10% faster than the web service. Probably it was caused by using xml marshalling in web services while our multiresulting mode employs native Java serialization mechanism.

This solution was also assessed by number of users. They were expected to subjectively evaluate both approaches and express their estimation in scale from 0 to 10 points. Most of the users noticed and underlined higher speed of processing in the multiresulting mode. The speed-up was more evident for the larger result set. However the most striking was effect of immediate response from the program run in the multiresulting mode. In contrast, a long delay until the program exploiting web services responded in any way was in general unacceptable by the users. And it was the basic psychological advantage of the new proposed methodology. All users were convinced that the multiresulting mode is significantly better than classic web service, while dealing with results of large dimensions. When software developer proposes a service returning collection of data, the new option supporting the multiresulting approach should be concerned.

Applications of the Multiresulting Mode

When should we use the multiresulting framework instead of standard approach? The general answer to this question is not as simple as it may look. The first obvious answer could be as follows: "Use it when a huge collection of data is sent to client". This is of course true, but the multiresulting mode could be used in many other situations. Look at possibility of returning Data Transfer Object to(DTO) the client. DTO is a snapshot of database or any other stateful resource. DTO object contains attributes which are set by back-end of application and when all of them are collected the object is sent to client. In the multiresulting mode the attributes could be returned partially, what could improve performance.

More interesting situation happens when a remote service uses some other remote methods itself and then their answers are put together and returned to the client. In a standard solution if one of the remote results comes very late or is not available at all, the method has to wait for it (maybe forever) to complete the full set of partial results or throw some kind of timeout exception. In this situation client cannot get any results. In this situation the multiresulting mode could be used and is real remedy for the problem. If n-th remote system will not answer, the client will retrieve at least n-1 first answers or all except the failed one. This will make client side less vulnerable to back-end malfunctions and in general any distributed processing may be more efficient and more reliable. This second feature can be achieved by repeating calls to remote services and/or their backups if they were unable to provide their results. In the multiprocessing mode the repeated requests may be much better targeted than in atomic, coarse-grained web services.

Not negligible characteristics of the multiprocessing mode is user's feeling that whole process is responsive and progressive. So the multiresulting mode should replace the solutions when processing takes a long time. Lets imagine situation when user does some action and then waits few minutes to complete it. He may think that abnormal situation occurred and the process should be initiated once again. To avoid such unwise behaviour some kinds of progress bar are applied. But in remote processing environment such progress bars cannot present the level of advancement of the initiated process. Applying our framework it is possible to implement reasonable indicators of progress in processing. Very often some parts of final results may be presented by a client program in parallel to continuous remote processing.

On the other hand there are situations where standard solutions are better. When a small amount of data is retrieved from the remote processing there is no gain from applying the multiresulting mode. If the communication traffic is a bottleneck of the distributed system then applications of the multiresulting should be restricted. Otherwise the multiresulting using active pooling may increase the traffic and as result the response from remote services may be delayed even more than when applying web service technology.

Conclusion

Using the new kind of remote services could be a good alternative to the standard solutions. The multiresulting mode is a bit faster than web services and could return data on demand. It is configurable and implementing a client is not a hard task thanks to the support of the implemented framework.

The development time in which the multiresulting solution could be launched in real product environment is very important. It looks that this time can be short, because after implementing business methods it is only necessary to annotate them and add few entries to application descriptors. For now most of internet application are deployed in full application servers like IBM Websphere, JBoss, etc. so the multiresulting may be used easily in such environments. This kind of service is a new one, it requires much more testing in a real development process and this experience may show some more advantages or disadvantages of this solution.

Acknowledgment

The paper is published with financial support by the project ITHEA XXI of the Institute of Information Theories and Applications FOI ITHEA (www.ithea.org) and the Association of Developers and Users of Intelligent Systems ADUIS Ukraine www.aduis.com.ua.

Bibliography

[B. Basham, 2004] B. Basham, K. Sierra, B. Bates. Head First Servlets & JSP. Ed. O'Reilly, 2004.

[P. Debu, 2007] P. Debu, R. Reza, L. Derek. EJB3 in action, Manning Co. 2007.

[E. Freeman, 2007] E. Freeman, E. Freeman. Head First Design Patterns, O'Reilly 2007.

[JEE] Java Enterprise Edition documentation, <http://java.sun.com/javaee>.

Authors' Information



Michał Plewka – Institute of Computer Science, Warsaw University of Technology, Nowowiejska 15/19, 00-665 Warsaw, Poland; e-mail: plewka.michal@gmail.com
Major Fields of Professional Activities: Java Developer, SOA technologies.



Roman Podraza – Institute of Computer Science, Warsaw University of Technology, Nowowiejska 15/19, 00-665 Warsaw, Poland; e-mail: R.Podraza@ii.pw.edu.pl
Major Fields of Scientific Research: Artificial Intelligence, Knowledge Acquisition.

CALCULATING OF RELIABILITY PARAMETERS OF MICROELECTRONIC COMPONENTS AND DEVICES BY MEANS OF VIRTUAL LABORATORY

Oleksandr Palagin, Peter Stanchev, Volodymyr Romanov, Krassimir Markov,
Igor Galelyuka, Vitalii Velychko, Oleksandra Kovyriova,
Oksana Galelyuka, Iliya Mitov, Krassimira Ivanova

Abstract: Today together with actual designing and tests it is often used virtual methods of designing, which support prior calculations of parameters of the developed devices. Such parameters include reliability. For this purpose the virtual laboratory for computer-aided design, which is developed during joint project by V.M. Glushkov Institute of Cybernetics of NAS of Ukraine and Institute of Mathematics and Informatics of BAS, contains program unit for calculating reliability parameters of separate microelectronic components and whole devices. The program unit work is based on two computing method: the first one uses exponential distribution of failure probability and the second one – DN-distribution of failure probability. Presence of theoretical materials, computing methods description and other background materials lets to use this program unit not only for designing and scientific researches, but also in education process.

Keywords: Virtual Laboratory; Computer-Aided Design; Reliability Calculation; Distributed System.

ACM Classification Keywords: J.6 Computer-Aided Engineering – Computer-Aided Design (CAD); K.4.3 Organizational Impacts – Computer-Supported Collaborative Work.

Introduction

Modern microelectronic component base lets to develop portable devices for wide and everyday using on the base of effects and phenomena, which exist in medicine, biology, biochemistry etc. Actual design of new devices and systems, which is often used, needs a lot of time, material and human resources. These expenses may be reduced with help of virtual methods of designing, which are realized by means of virtual laboratories of computer-aided design (VLCAD) [Palagin, 2009].

VLCAD is worth to be used on the stage of the requirements specification or EFT-stage, because it gives the possibility enough fast to estimate the project realization, certain characteristics and, as a result, expected benefit of its practical realization. Using of VLCAD doesn't need expensive actual tests and complicated equipments.

Calculating of parameters

Devices, what are developed, are, generally, data acquisition and processing channels. During designing and modeling such systems (channels) it is very important to make prior parameters calculating and evaluating. These parameters include reliability, precision, performance, cost etc.

Having only model of future device it is possible enough quickly to make prior calculating of device parameters. Also we have possibility to calculate several alternative variants of project and choose the optimal one according to user' criteria (or predetermined criteria). For prior calculating we developed and filled databases, which contain information about a large amount of microelectronic components and units. Databases contain next main parameters: 1) microelectronic component name; 2) manufacturer of microelectronic component; 3) group, to which microelectronic component belongs; 4) parameters, which are intrinsic to some group of microelectronic components, e.g. nominal currents, voltages and powers, interface types; 5) manufacturing technique;

6) reliability parameters; 7) work temperature range; 8) price; 9) body type and conditions of installing; 10) unique properties of microelectronic component.

For prior calculating of each parameter it is developed program model. Under program model we mean separate program, group of programs or program complex, which let by means of calculating sequence and graphical display of result to reproduce processes of object functioning under influence of, as a rule, random (or predetermined) factors. The model purpose is to obtain quantitative or qualitative results. Quantitative result is intended to predict some future values or explain some past values, which characterized whole systems. Qualitative results, on the base of analysis, let to detect some earlier unknown system characteristics.

Generalized graphical display of stages sequence of prior calculating is shown on the fig. 1 (portable device "Floratest" [Romanov, 2007] is as example).

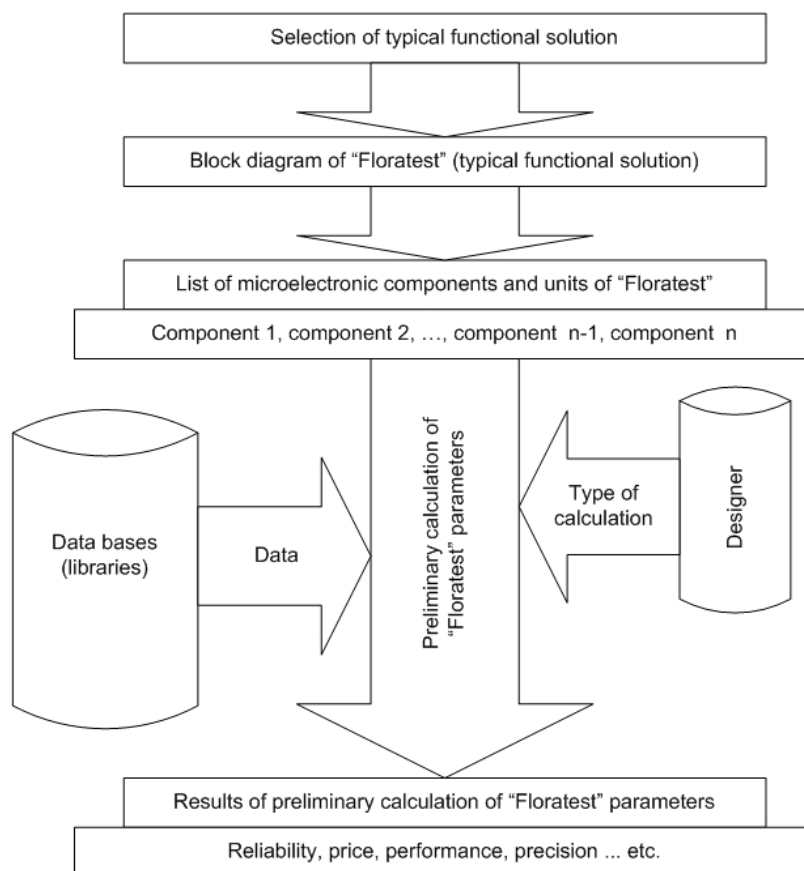


Fig. 1. Sequence of stages of prior parameter calculating of developed devices

Calculating of reliability

Most devices are generally data acquisition and processing channels. Such channels are, per se, systems without recovery, elements of which are connected serially. But meaning "series connection" doesn't always agree with physical series connection [Belyaev, 1985]. In this case we mean, that failure of any system element causes failure of whole device or channel.

Because of complexity of physical processes, which cause failures, impossibility to take account of all start conditions and random influences, in present days it is accepted to consider a failure as random event in meaning, that if we even know structure type of system and application conditions it is impossible to detect time and place of failure.

As a rule, prior evaluation of reliability parameters of separate microelectronic components is made by manufacturers on the base of results of highly accelerated stress tests [MIL-STD-883]. Reliability parameters are shown in general form and grouped by types of manufacturing technique or large classes of functional-similar components (amplifiers, converters, processors, controllers, memory etc.). Reliability level of modern microelectronic components is estimated by next characteristics:

- 1) failure rate λ , which measured in units FIT (failure in 10^9 hours of work);
- 2) mean time to failure T_0 .

Mean time to failure is estimated as function of failure rate in consideration of distribution law of failure probability. Today there are many computing methods of reliability, which are based on different distribution laws of failure probability [Azarskov, 2004]. We analyzed methods, which are oriented on computer and instrumentation tools, and adapted them for realization as program models.

For prior calculating of reliability parameters of components, devices and systems it is created program unit for calculating of reliability parameters. This program unit has next features:

- 1) calculating reliability parameters of separate microelectronic component or whole device;
- 2) using different computing methods, which are based different distribution laws of failure probability;
- 3) selecting for calculating reliability parameters of device both component experimental data, which are got by manufacturers with help of accelerated tests, and component data, which calculated by program unit;
- 4) comparing of reliability parameters calculating results, which were obtained by means of different computing methods;
- 5) optimizing reliability index on the base of microelectronic components from VLCAD databases;
- 6) containing fundamentals about reliability theory, including computing sequence of reliability parameters by means of every computing method.

The work of program unit for calculating reliability parameters of microelectronic components and whole devices is based on using two computing methods, which are adapted by us for realization as program models:

- 1) by using exponential law of distribution of failure probability (probabilistic model of failure distribution), which is used by many manufacturers of microelectronic components;
- 2) by using DN-distribution of failure probability (probabilistic-physical model of failure distribution), which was proposed by Ukrainian scientists V. Strelnikov and O. Feduhin [Azarskov, 2004, Streljnikov, 2004].

Selected models of distribution have very important difference. Probabilistic model allows only two state of system elements – operable and faulty. Probabilistic-physical model considers continuous set of system element and system states with continuous time. First model uses exponential law of distribution of failure probability, the second one – DN-distribution.

As was written above the exponential law of distribution of failure probability is used by many manufacturers of microelectronic components, such as: Analog Devices Inc. [ADI, 2004], Motorola [Motorola, 1996] etc. This law of distribution is recommended by Department of Defense of USA [MIL-STD-883]. Exponential law is one-parametric function.

According to this method failure rate λ is got from experimental data of manufacturer or calculated by formula

$$\lambda = \frac{\chi^2}{2 \cdot N \cdot H \cdot At}, \quad (1)$$

where χ^2 – function, values of which depend on component failure quantity and confidence bounds of interval and are received from [Romanov, 2003, Motorola, 1996]; N – quantity of components under accelerated tests; H –

duration of accelerated tests, hours; At – coefficient of failure rate acceleration, which is calculated by the formula [Streljnikov, 2002] (Arrhenius law)

$$At = e^{-\frac{Ea}{k} \left(\frac{1}{T_v} - \frac{1}{T_r} \right)}, \quad (2)$$

where Ea – activation energy (= 0,7 eV); k – Boltzmann constant (= $8,617 \cdot 10^{-5}$); T_v – temperature of accelerated tests, K; T_r – work temperature of microelectronic component, K.

Mean time to failure T_0 , in accordance with exponential law, is calculated by formula

$$T_0 = \frac{1}{\lambda}. \quad (3)$$

For calculating mean time to failure T_{cp}^e (exponential law) of whole device it is used the next formula

$$T_{cp}^e = \left(\sum_{j=1}^n m_j \lambda_j \right)^{-1}, \quad (4)$$

where m_j – quantity of units of j type ($j = 1, 2, \dots, n$); λ_j – failure rate of units of j type.

For calculating reliability parameters by means of second adapted method it is used DN-distribution of failure probability, which is two-parametric function as opposed to exponential law. In [Azarskov, 2004, Streljnikov, 2004] it is shown, that by using DN-distribution it is possible to get result, which are more close to real data.

For calculating reliability parameters of microelectronic components by means of DN-distribution it was updated and adapted algorithm, which is shown in [Azarskov, 2004]. It will be shown on the example the features of every method. For this it is necessary to calculate mean time of test of every sample by formula

$$t_n = \frac{EDH}{N}, \quad (5)$$

where EDH – parameter "equivalent device hours", which can be found in the manufacturer documentation, manufacturer web-site or calculated by formula

$$EDH = N \cdot H \cdot At. \quad (6)$$

Mean time to failure T_0 is calculated by substitution of values λ and t_n in formula (λ can be calculated by (1) or found in the manufacturer documentation)

$$\lambda = \frac{f(t_n)}{R(t_n)}, \quad (7)$$

$$\text{where } f(t_n) = \frac{\sqrt{T_0}}{t_n \sqrt{2\pi t_n}} \exp \left[-\frac{(t_n - T_0)^2}{2T_0 t_n} \right]; \quad (8)$$

$$R(t_n) = \Phi \left(\frac{T_0 - t_n}{\sqrt{T_0 t_n}} \right) - \exp(2) \cdot \Phi \left(-\frac{t_n - T_0}{\sqrt{T_0 t_n}} \right). \quad (9)$$

Since during experimental evaluation of failure rate of microelectronic components the rate of microelectronic components with failures is 1...5 % [Streljnikov, 2004], so it is possible to consider $\lambda \approx f(t_n)$. So, the formula (7) can be written as

$$\lambda \cong \frac{\sqrt{T_0}}{t_n \sqrt{2\pi t_n}} \exp \left[-\frac{(t_n - T_0)^2}{2T_0 t_n} \right]. \quad (10)$$

For calculating mean time to failure T_{cp}^D (DN-distribution) of whole device it is used next formula

$$T_{cp}^D = \left(\sum_{j=1}^n m_j T_j^{-2} \right)^{-\frac{1}{2}}. \quad (11)$$

Computing algorithms of reliability parameters by means of these two methods are shown on fig. 2. Program unit, which is developed by authors, is a part of VLCAD and operates on the base of these computing algorithms.

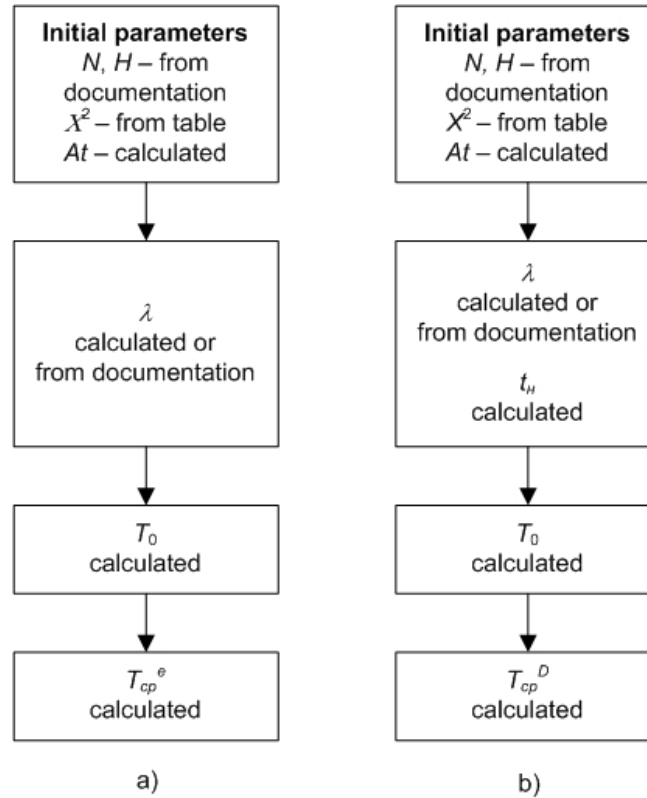


Fig. 2. Computing algorithms of reliability parameters with using of exponential law of distribution of failure probability (a) and DN-distribution of failure probability (b)

Using developed program unit it were obtained dependences of system operating reliability versus time. For simplifying it is considered, that system consists of same microelectronic components, that are manufactured by means of same manufacturing technique (e.g. "Bipolar $2.5 \mu\text{m}^2$"). Should note, that program unit allows to calculate reliability parameters of systems, which consist of different quantities of microelectronic components, manufactured by means of different manufacturing technique. Dependences were obtained for 4 systems, which consist of 1 thousand, 10 thousand, 20 thousand and 100 thousand microelectronic components respectively. Confidence bounds equal 60 %. Dependences, obtained by means of method on the base of exponential law of distribution of failure probability, are shown on fig. 3 and fig. 4. Dependences, obtained by means of method on the base of DN-distribution of failure probability, are shown on fig. 5.

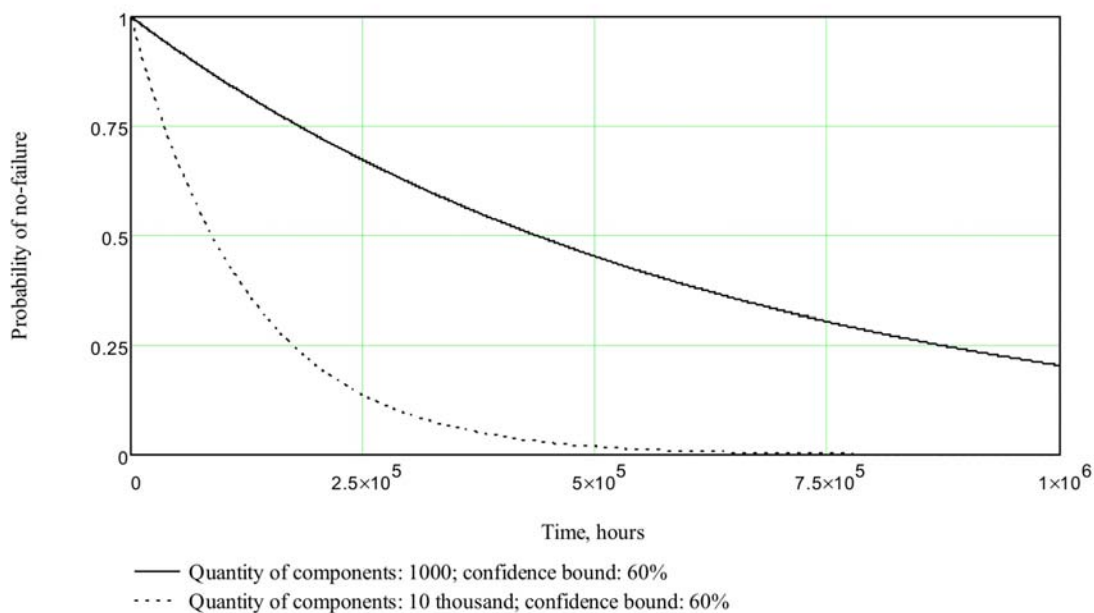


Fig. 3. Dependence of probability of system operating reliability (1 000 and 10 000 components) versus time, exponential law of distribution of failure probability

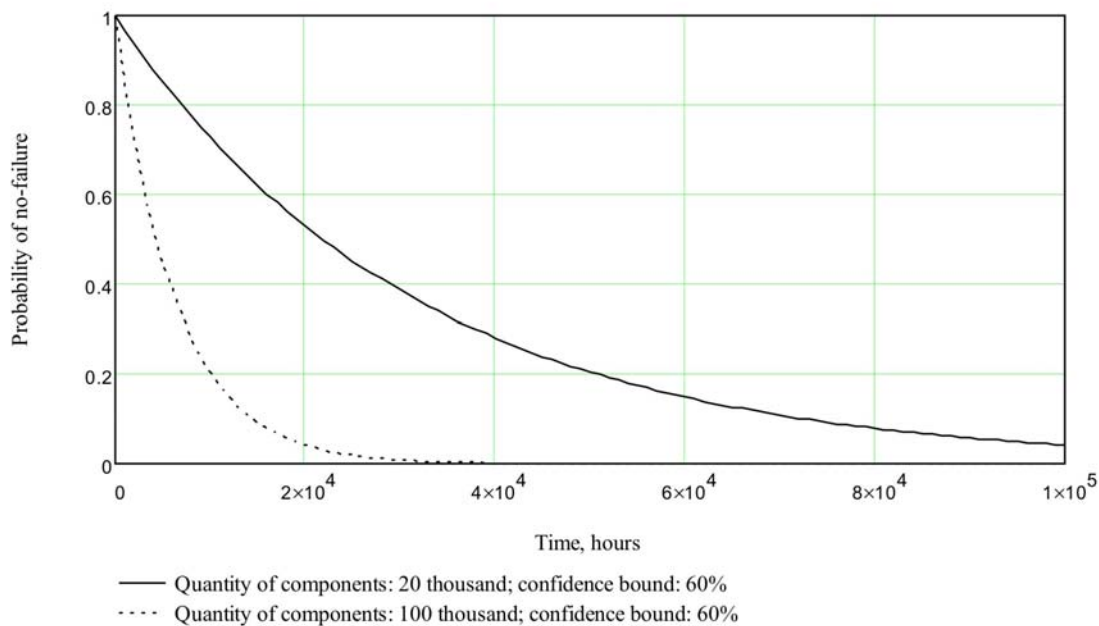


Fig. 4. Dependence of probability of system operating reliability (20 000 and 100 000 components) versus time, exponential law of distribution of failure probability

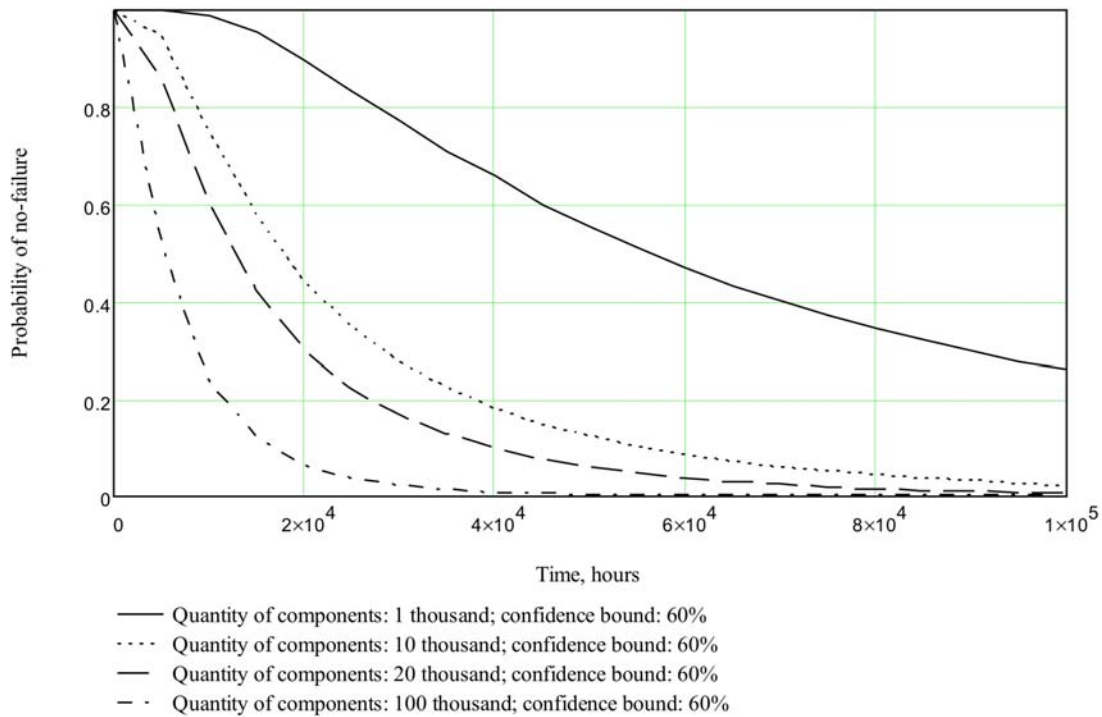


Fig. 5. Dependence of probability of system operating reliability (1 000, 10 000, 20 000 and 100 000 components) versus time, DN-distribution of failure probability

Combine some calculated values to the table for more detail analysis of obtained dependences (table 1). For more visualization the table data are used for plotting graphic (fig. 6). Should note, that, as in previous cases, system consists of same microelectronic components, that are manufactured by means of same manufacturing technique "Bipolar $2.5 \mu\text{m}^2$".

Table 1. Duration of no-failure operation of systems, which consist of different quantities of components

Quantities of micro-electronic components	T_0	Exponential law of distribution		DN-distribution	
		CB* = 60 %	CB = 90 %	CB = 60 %	CB = 90 %
1 000	hours	629723	320256	82298	76647
	years	71,886	36,559	9,395	8,749
20 000	hours	31486	16372	18402	17139
	years	3,594	1,869	2,101	1,956
50 000	hours	12594	6549	11639	10840
	years	1,438	0,748	1,329	1,237
100 000	hours	6297	3274	8230	7665
	years	0,719	0,373	0,939	0,875
200 000	hours	3149	1637	5819	5420
	years	0,359	0,187	0,664	0,619

* CB = Confidence bounds

It was analyzed systems, which consist of from 1 thousand to 200 thousand microelectronic components. Such range of components wasn't selected by accident. On this range according to our analysis it is happened some changes in ratio of reliability parameters, which were calculated by two methods.

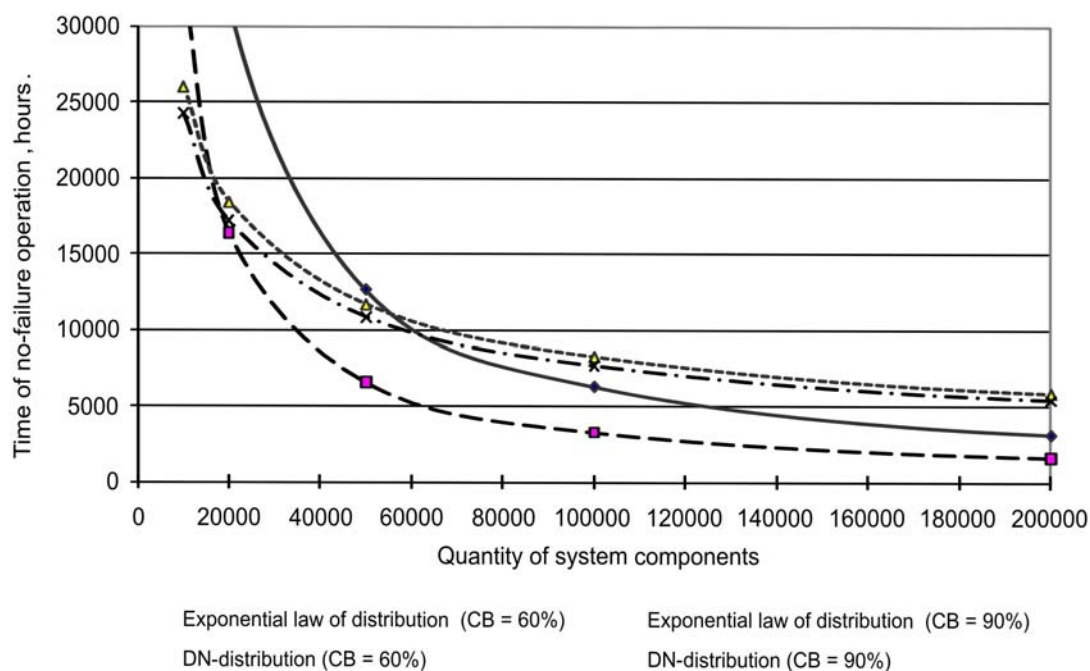


Fig. 6. Dependence of duration of no-failure operation of systems, which consist of different quantities of components, versus quantities of system components

Having analyzed graphics (fig. 3–6) and table 1, it is possible to make next conclusions. For every confidence bound (in our case there are 60 % and 90 %) there is such system components quantity, below of which the method on basis of exponential law of distribution overstates the reliability parameters, and above of which – puts too low them. For confidence bound 60 % this quantity equals approximately 60 thousand components, and for confidence bound 90 % – respectively 20 thousand. The difference of results, that are obtained by means of these two computing methods, can be explained by method error [[Streljnikov, 2004]. The method error becomes apparent, because exponential law of distribution is one-parametric function, while DN-distribution is two-parametric function or, in other words, diffusive distribution.

Our conclusion conforms very well with test and computing results, which were fulfilled by Ukrainian scientists in the field of reliability theory [Streljnikov, 2005]. Should note, that results in article were obtained for the first time thanks to using developed virtual (program) models of computer tools.

Program models were used for calculating reliability parameters (e.g. time of no-failure operation) of portable device for express-diagnostic of plant state "Floratest", which is developed and created in the V.M. Glushkov Institute of Cybernetics of NAS of Ukraine.

Conclusion

Presence of program unit for reliability parameters calculating as a part of VLCAD allows to fulfill prior calculating of reliability parameters of designed device and, on the basis of obtained results, make conclusion about correspondence of calculated parameters to beforehand specified ones. Positive features of program unit is availability of two methods of reliability parameters calculating of separate microelectronic components and whole devices, which are based on different models of distribution of failure probability (one- and two-parametric functions).

Program unit can be used not only for calculating of reliability parameters, but in education process for gaining theoretical information from reliability theory.

Program unit now is used in practical tasks for calculating reliability parameters of portable devices.

Acknowledgements

This work is partially financed by Bulgarian National Science Fund under the joint Bulgarian-Ukrainian project **D 002-331 / 19.12.2008** "Developing of Distributed Virtual Laboratories Based on Advanced Access Methods for Smart Sensor System Design" as well as Ukrainian Ministry of Education under the joint Ukrainian-Bulgarian project No: **145 / 23.02.2009** with the same name.

Bibliography

- [Palagin, 2009] Palagin O., Romanov V., Markov K., Velychko V., Stanchev P., Galelyuka I., Ivanova K., Mitov I. Developing of distributed virtual laboratories for smart sensor system design based on multi-dimensional access method // Classification, forecasting, data mining: International book series "Information Science and Computing". Number 8: Supplement to International Journal "Information Technologies and Knowledge". Volume 3/2009.– 2009. – P. 155–161.
- [Romanov, 2007] V. Romanov, V. Fedak, I. Galelyuka, Ye. Sarakhan, O. Skrypnyk. Portable Fluorometer for Express-Diagnostics of Photosynthesis: Principles of Operation and Results of Experimental Researches // Proceeding of the 4th IEEE Workshop on "Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications", IDAACS'2007. – Dortmund, Germany. – 2007, September 6–8. – P. 570–573.
- [Belyaev, 1985] Belyaev Yu., Bogatryov V., Bolotin V. et al. Reliability of technical systems: reference book / Editor: Ushakov I. – Moskow, 1985. – 608 p. (in Russian).
- [MIL-STD-883] MIL-STD-883. Test Methods and Procedures for Microcircuits.
- [Azarskov, 2004] Azarskov V., Streljnikov V. Reliability of control and automation systems: tutorial. – Kiev, 2004. – 164 p.
- [Streljnikov, 2004] Streljnikov V. Estimating of life of products of electronic techniques // Mathematical machines and systems. – 2004. – № 2. – P. 186–195.
- [ADI, 2004] ADI reliability handbook. – Norwood : Analog Devices, Inc., 2004. – 86 c.
- [Motorola, 1996] Reliability and quality report. Fourth quarter 1996. – Motorola Inc., 1996
- [Romanov, 2003] Romanov V. quantitative estimation of reliability of integrated circuits by results of accelerated tests / Electronic components and systems. – 2003. – № 10. – P. 3–6.
- [Streljnikov, 2002] Streljnikov V., Feduhin A. Estimation and prognostication of reliability of electronic elements and systems. – Kiev.: Logos, 2002. – 486 p.
- [Streljnikov, 2005] Streljnikov V. Method errors of calculating of reliably of systems // Systemic problems of reliability, quality, information and electronic technologies: 10-th international conference / Conference works. – October, 2005. – Sochi, Russia. – Part 6.– P. 136–143.

Authors' Information



Oleksandr Palagin – Academician of National Academy of Sciences of Ukraine, Depute-director 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: palagin_a@ukr.net



Peter Stanchev – Professor, Kettering University, Flint, MI, 48504, USA
Institute of Mathematics and Informatics – BAS;
Acad. G.Bontchev St., bl.8, Sofia-1113, Bulgaria; e-mail: pstanche@kettering.edu



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, VRomanov@i.ua



Krassimir Markov – Institute of Mathematics and Informatics, BAS,
Acad. G.Bonthev St., bl.8, Sofia-1113, Bulgaria; e-mail: markov@foibg.com



Igor Galelyuka – Senior research fellow of V.M. Glushkov's Institute of Cybernetics of National Academy of Sciences of Ukraine; Candidate of technical science; Prospect Akademika Glushkova 40, Kiev–187, 03680, Ukraine; e-mail: gilib@gala.net



Vitalii Velychko – Doctoral Candidate; V.M.Glushkov Institute of Cybernetics of NAS of Ukraine, Prosp. Akad. Glushkov, 40, Kiev-03680, Ukraine; e-mail: velychko@aduis.com.ua



Oleksandra Kovyriova – 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: alexandara.skripka@gmail.com



Oksana Galelyuka – Research fellow of Institute of encyclopedic researches of National Academy of Sciences of Ukraine; Tereschenkivska str., 3, Kiev, 01004, Ukraine



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



Krassimira Ivanova – Researcher; Institute of Mathematics and Informatics, BAS,
Acad. G.Bonthev St., bl.8, Sofia-1113, Bulgaria; e-mail: ivanova@foibg.com

OPTIMIZING ROUTING PROCESS WITH A KINETIC METHOD

Olexandr Kuzomin, Ievgen Kozlov

Abstract: *The article describes a "native" algorithm for getting packet forwarding decision, based on considering a network to be an ionic gas. Therefore necessity for the information becomes a field intensity, with distribution function gradient showing the best forwarding way to bring the system to a minimal energy state and all the nodes having an equal information set. The proposed algorithm can be optimized on any of function parameters versus metrics only in classic routing algorithm.*

Keywords: *Packet forwarding, Boltzmann equation, Vlasov equation, Thermal grid, Kinetic algorithm*

ACM Classification Keywords: *C.2.2 Network Protocols - Routing protocols*

Introduction

The aim of the proposed research is consideration and optimization of packet forwarding direction decision making as one aspect of a "quality of service" problem in point-to-point networks. A task to increase a "quality of service" level while decreasing transit nodes load does appear in modern networks because of frequent reaching top load level (even causing packet loss) in some network parts with non-equivalent network load distribution in other parts. It is clear that for the task mentioned above we need to build some algorithm, which allows optimizing node interaction, taking into consideration different requirements caused by many kinds of information, e.g. we must optimize routing (packet forwarding) algorithm. It is clear that the algorithm we are looking for may fit the following requirements:

- an RSVP-method should not be used;
- flooding network traffic should not be generated or the traffic should be minimal, if it is not possible to do without it;
- the algorithm should be fast enough to find a solution to forward the next packet while the current packet is being transmitted; in doing so it is desirable to reach the solution with a low CPU utilization;

At present, a Bellman-Ford routing algorithm is a mainly used one, for example, in a 'routed' daemon. This algorithm is rather simple and fast, but its realization needs sending up routing tables across the network (thus flooding it with a service-special traffic) every 30 seconds. Moreover, this method has no capabilities of adjusting itself for optimizing data transfer process for some special conditions of networking tasks and different parameters of network parts, because one and the only parameter it uses is a "metric", which can be referred to as a statistical value, implicitly depending on a number of network characteristics; the function used to describe this dependency being very complicated. Adjusting capabilities of packet routing algorithm make up a base for "quality of service" question when adopting a network for the type of information being transferred by it, because this will allow taking into account this information type and its requirements at the very moment of taking a decision regarding packet forwarding. Recently some research attempts have been made with the aim of finding a routing algorithm with a capability to optimize packet forwarding direction decision making with dependence on the information type and network characteristics. Among such algorithms one can find the "Thermal grid" algorithm (provided by H. Unger, [Unger, 2009]), the "Ant pheromone" algorithm (given in particular in [Singh, 2010]). Note the idea of describing a networking process through classical physics analogies. The idea itself is not new. An example of an earlier work using a "temperature state of a system" appears in [Yong Cui, 2003].

Techniques of research

A proposed "Kinetics" algorithm is based on "Thermal grid" algorithm and is aimed at increasing its efficiency, because the latter can give improvement in delivery time only when it is used in about 50% of decision making [Unger, 2009]. In their later works the authors of a "Thermal grid" algorithm added two parameters [Lertsuwanakul, 2009], having influence on each method "weight" in a decision making (considering not only "temperature", i.e. load of a buffer, but also transmitting time), taking a partial step backward towards classical routing methods based on metrics.

As the "Thermal algorithm" is based on associating a buffer load of a node with a temperature, the "Kinetics algorithm" basis is associating networking characteristics (e.g. connection speed, buffer size, switching time etc.) with the identical physical quantity (mass, temperature, electric charge or so). In such a conception, every node in a network (or grid) can be described by

a) in case of a network

a system of nodes:

- each node has some information and demands some information;
- each node has some direct connections;
- each node has a limited channel width and each node has a limited cache;
- each node has a stability rate;

b) in a case of electromagnetic field:

a system of cells with particles on it, mainly one particle per cell;

- any particle (ion) has its electric charge;
- each particle can interact with any number of particles only around it;
- each particle has some limited velocity (or temperature. Here a meaning of a "temperature" acquires a new meaning and describes not only a cache size, but a connection speed too);
- each particle has a mass-property.

It is clear that these parameters can be associated with each other. Thinking in such a way, we can associate networking nodes with particles and describe our system (remembering it is a P2P-network really) with Boltzmann kinetic equation:

$$\frac{\partial f}{\partial t} + \frac{\partial f}{\partial x} \cdot \frac{p}{m} + \frac{\partial f}{\partial p} \cdot F = \frac{\partial f}{\partial t} \Big|_{st}, \quad (1)$$

Here f is a particle distribution function;

t is time;

x is a coordinate;

p is a particle impulse;

m is a particle mass;

F is a field of external(*) forces wherein a particle exists;

$\frac{\partial f}{\partial t} \Big|_{st}$ is a so called "collision integral" showing an impact interaction of particles in classical mechanics.

For vacuum or collisionless plasma this integral is considered to be a zero. In our case it would be non-zero only when we want to take overload of some network parts and impossibility for transferring data through that path into

account. This equation contains a particle distribution function – unsteady in time parameter, describing a probability density distribution of containing a particle in a infinitely-small volume dV . Solving this equation to find this function is rather a hard task in general [Vlasov, 1938], though this task is well solvable by iteration schemes in many cases and even analytically for some particular cases. The task was solved, for example, for plasma (by A. A. Vlasov himself). It is known that it was solved for colloids in a two-dimensional approach etc. While one should take a look at the mentioned equation, one may note, that if we associate, say, a necessity for information with an electric charge, a field F (mind an asterisk near it) would be a self-consistent one. This field can be easily found particularly with a "clouds-in-cell" method. So, we can model our system by describing it with Vlasov-Maxwells equation system [Vlasov, 1938] and solving it numerically.

$$\frac{\partial f_i}{\partial t} + \text{div}_v v f_i + \frac{q_i}{m_i} \left(E + \frac{1}{c} [v, B] \right) \text{grad}_v f = \sum_j \left(\frac{\partial f}{\partial t} \right)_{st}^{i,j},$$

$$\text{div} E = 4\pi\rho, \quad \text{rot} E = -\frac{1}{c} \frac{\partial B}{\partial t},$$

$$\text{div} B = 0, \quad \text{rot} B = \frac{1}{c} \frac{\partial E}{\partial t} + \frac{4\pi}{c} j,$$
(2)

Here we came across with parameter q_i - a charge of particles of i type. Remembering that the data transferred by the network can be of many types, each of those types gives us a new type of particles (i), but to avoid an interaction of different information types, we should mark the nodes (particles) not only with a charge but with a color.

The second interesting parameter in this equation is j - a "current density". This parameter indicates a channel load (but not a buffer load!). A field tension is defined in the same way as defined in a classical physics. A total number of variable type parameters in the equation amounts to twelve, so the algorithm can be easily adapted to many various requirements of numerous data types even without using any other forwarding decision making algorithms.

The above mentioned system (2), [Vlasov, 1938], allows finding a distribution function state in any time moment if its state is known at the starting (zero) time. Our aim consists in not just solving the system in terms of finding a distribution function, but finding its changes in small time moments. Thus the information about preferred forwarding way should be found in the spatial gradient of derivate of the distribution function with respect to time. After network points are positioned in the cells of some grid and known system characteristics, associated with physical quantities in a mentioned way, are substituted into the system (2), we can start iterations, which will give us some shift of the distribution function's extreme points. This shift will give us the preferred way of packet forwarding. The difference between a network node and a real particle is that the latter one has many degrees of freedom, in opposite to the network node, which can forward a single packet only one way at any moment (indeed, we do not take a parallel data transfer and duplexes into account here, because these parameters are certain to occur in a modeled system, but we must exclude a broadcasting as mentioned in the task before). So, we have to choose one and the only way for packet forwarding each time, and we will assume it to be a direction of maximum extreme point shift. It is clear, we have to correct our model after the forwarding has been done before the next iteration takes place, because the model assumes a distribution function and a space to be continuous, while the forwarding is made in a discrete space of network nodes. Thus the model needs correction after each iteration takes place. But while correcting a model we can quickly adjust it to the changes in a network topology (e.g. some new points can appear in a net, while some links go down etc.) so the subsequent iteration will use up-to-date data both about the structure and state of the network.

For the equation system (2) we must build a finite-difference scheme to start iterations on. After all we have said above, we can just use a simple scheme of a first order regardless of its instability, because of data correction after each iteration. Moreover, we consider all the variables to be normalized. A consequence of actions for each iteration looks like:

- 1) gather the data about the current network state;
- 2) dispose the nodes over the cells;
- 3) calculate a field E (after calculating charge and current densities);
- 4) calculate a field B (no networking equivalent this time);
- 5) calculate a distribution function f;
- 6) find a $\frac{\partial f}{\partial t}$ – derivate of the distribution function with respect to time and its spatial gradient;
- 7) perform the forwarding and correct a model etc.

The scheme can be found to look like

$$\frac{f - f_{-1}}{t - t_{-1}} + v \cdot \text{div}_r f + \frac{q}{m} \left(E + \frac{1}{c} [v, B] \right) \text{grad}_v f = \frac{\partial f}{\partial t} \Big|_{st}, \quad (3)$$

Now we can see the external field component is not present, but there can be a charge source in the net (i.e. a system is not a complete and closed-up in a general occasion). Today we discard this fact, because one can easily notice that this will not have a significant affect on a result. The equation (3) transfers into a working formula for iterations:

$$w = \text{grad}_r \frac{\partial f}{\partial t}, \quad (4)$$

$$\text{where } \frac{\partial f}{\partial t} = f - f_{-1} = + \frac{\partial f}{\partial t} \Big|_{st} - v \cdot \text{div}_r f - \frac{q}{m} \left(E + \frac{1}{c} [v, B] \right) \text{grad}_v f.$$

In the system (4) the variable we are looking for is the projection of the vector W on our coordinate grid, where the node (point) is positioned on and described by, and (again) we choose a direction where this projection is maximal as a primary forwarding one. Note the time is excluded in the last equation. It is implicitly present as we are working with the system within one technically infinitely small time period.

A modeling itself runs a rather simple algorithm (though now it has many limitations). The algorithm below can be described by these five steps:

- 1) get a matrix of connectedness;
- 2) mark the nodes having some information;
- 3) mark the nodes needed the information (targets);
- 4) start iterations:
 - a) find routes ;
 - b) look if the forwarding target node is already busy on this step;
 - c) if this node is busy, skip one iteration, otherwise perform data transfer;
 - d) correct the model according to the transfers which have just taken place;
- 5) if the information is delivered to the targets, the job is done; otherwise continue from the step 4-a.

The step 4-a includes making a decision on a packet forwarding route. The model is to run two times: first the routing will be done with a classical algorithm (excluding a step 4-d), then with an alternative algorithm.

The open question is currently a transit node transfer. We have mentioned above that all the points will be marked with a color (but, in opposite to quarks our particles-nodes can carry many colors at the same time). This is figured out by a necessity for simultaneous transferring many different types of data, while not mixing these data. In a field of each color a node may be charged positively (it has certain information), negatively (it does need that information), or neutral. The problem is that in many cases the dominant part of transit nodes will be neutral in one particular color, but at the same time it can carry a heavy charge of other color. This might be interpreted as the node does not need and is not loaded with the information of current type, but is heavily loaded with an active transfer of other information. This is partially described by the node temperature, but for more accurate modeling this fact needs taking a right part of a Boltzmann equation, i.e. "collision integral" into consideration, which rises a complexity of a model together with making the solving of equation more difficult at every step, slightly decreasing an efficiency.

Experiment and results

For the experiment a simulating program was written using GNU Octave. No more special simulation software was used. The simulating program consists of 3 big modules. The main one is used for simulating a network itself, and the other two are simulating routing algorithms. Within the main module the information about connections (in a matrix form, $N \times N$ nodes, where N is the number of points in the network, now 256) and about data presence is stored. A network topology is available on a fig. 1.

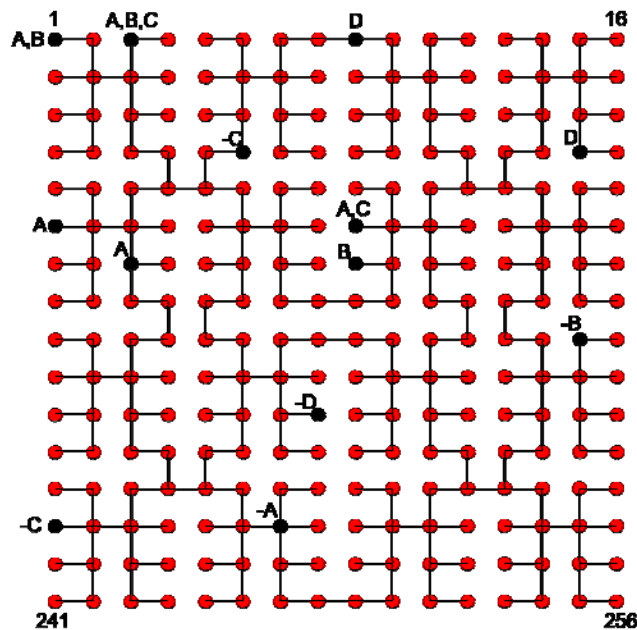


Fig. 1 – a simulated network topology with the color marks

The only characteristic of the connection was "metric" and no node parameters were used except for data presence. The buffers of nodes were considered to be infinite and the nodes stored a copy of all the packets for infinite time.

Each information type was marked with a color by a letter (A to L), and the letter "N" found on a node represents that the node is empty. The nodes were queried in a cycle, and when the node is queried, it looks up if it has the data of each color and, if so, forwards these data to one neighboring node. A destination node was chosen by a classical Dijkstra algorithm (program module 2) on the first run and with a "kinetic" algorithm (program module 3),

we switched to on the second model run. A node that gets a forwarded packet becomes marked with a packet color, and then the next node is queried. After the last one node is queried, a model looks up if all the nodes which needed the information have received it, and if they have not, the nodes are queried for transferring the information packets once more. The map of nodes marked by color is printed after each modeling cycle. The number of nodes marked colors A and B is shown with the diagram on a fig.2., on first, second and last steps.

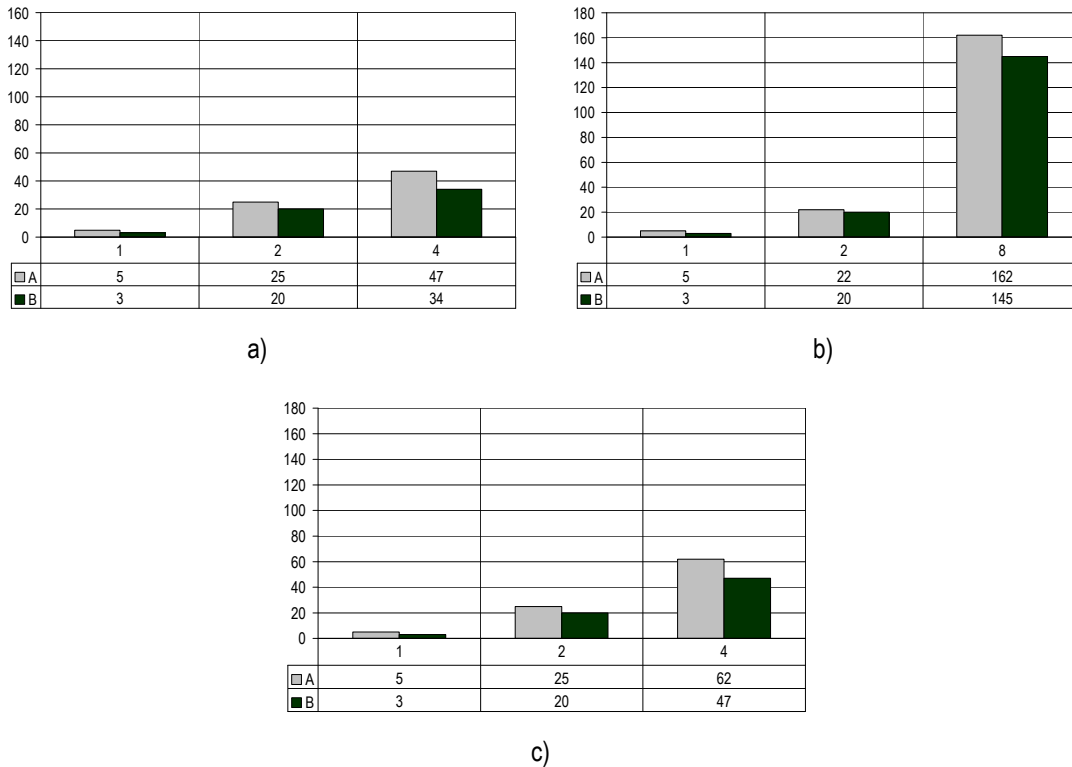


Fig. 2 – nodes of the colors A and B on first, second and last steps of modeling.
 a) the classic algorithm used, b) the “kinetic” algorithm with two (no electric field) and
 c) the “kinetic” algorithm with three parameters were used

An experiment was provided twice with the kinetic forwarding: the first time only the information presence on nearest nodes and the speed of connections were taken into consideration. The result was that the information was spread equally over the network (see fig.2), regardless of the fact whether it was needed by the node or not. This looks like particles of paint that are spread equally over the volume when the paint is dropped into a glass of water. The network seems to be flooded with the packets. On the second run the field was calculated and taken into consideration. The nodes having information were marked positive, the nodes needed that information were marked negative, and the electric field was found as it is defined from the nodes potentials (by Kirchoff's circuit laws).

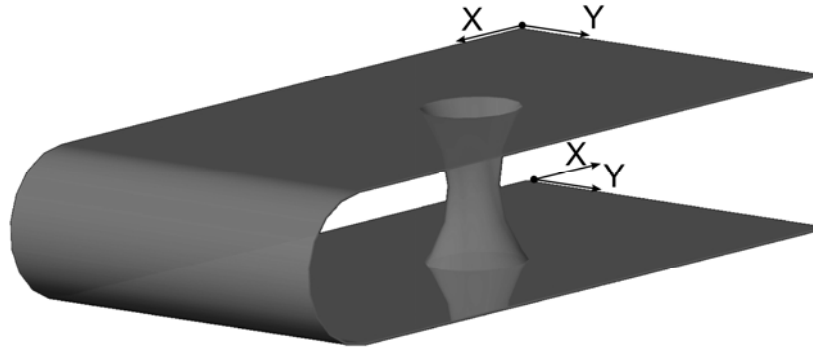


Fig. 3 – The “wormhole” - a ground of using multidimensional system when different parts of the network are connected directly. The hole is in Z-coordinate, X and Y are shown on a bent surface.

While programming, it was chosen finally to use multidimensional system. The connections then were separated over the dimensions so that the node could have only one connection in each dimension. An explanation for associating the connections with the multidimensional structure can be seen on a fig. 3: each direct connection jumps over the layer of a subnet and can shortly connect nodes like a direct wormhole. The longitude part of Vlasov equation (5) (see [Vlasov, 1938] for explanations) was solved then on a multi-dimensional field:

$$\frac{\partial f}{\partial t} + \text{div}_r v f + \frac{q}{m} E \cdot \text{grad}_v f = 0. \quad (5)$$

The result was that the performance of forwarding process had been dramatically improved: the information was forwarded almost the same way it was forwarded by a classical algorithm, so the kinetic method proved to be rather effective even on the system, which was initially built to model transfers by Dijkstra algorithm. A table showing the number of forwards by each color with classic algorithm, the “kinetic” algorithm without field and the “kinetic” algorithm including field is shown below (Table 1). The Dijkstra algorithm was in ideal conditions for its performance, while the conditions for the “kinetic” algorithm were very poor: only few of possible optimization parameters could be taken into account. There were no overloaded nodes; the connections were infinitely stable and so on. The “kinetic” algorithm proved to be flexible enough to be optimized on each network parameter (since the performance of the algorithm had been dramatically improved after taking the “electric” field into account).

Table 1. An average number of forwards for 4 colors with a classic and “kinetic” algorithm.

Color	Classic algorithm	“Kinetic” algorithm	
		without field component	including field component
A	46	161	61
B	33	144	46
C	92	255	121
D	26	115	37

Conclusion and future work.

The described method looks rather complex and easily optimizable. At the current step of modeling we achieved efficiency over 70% of the speed of Dijkstra algorithm (efficiency is measured as time elapsed on transmission itself, not including time elapsed on taking forwarding decision). Any way we should consider an algorithm to be

quite good if it gives efficiency slightly less versus classical algorithms when dealing with a system within the same parameter set. This can be understood easily if we turn to the fact that any iteration algorithm gives less accurate result within a longer period of time versus analytical solutions. But we must remember that in some cases analytical solution does not exist or can hardly be found, thus on occasion when we need taking not only metrics and even more than metrics and buffer, alternative algorithms would show greater performance. Within our experiment we have put Dijkstra's algorithm into ideal conditions for its performance while our "kinetic" algorithm was used with only small parts of its parameters. Even in such a poor occasion for its performance the algorithm proved to be a) rather effective versus classic algorithm and b) this algorithm gives good adjustment possibilities, as we can see by adding just one parameter to the equation. In the future it is desired to build a more complex model with 64K nodes, including unstable nodes, small-buffered nodes and overloaded nodes. In such a model it is expected to find performance of this algorithm much more perfect than classic algorithm, and to avoid packet losing on overloaded networks (this cannot be done by a classical algorithm). In order to achieve maximum efficiency ([Unger, 2009] and [Lertsuwanakul, 2009]) it is still recommended to apply many algorithms with different "weights" to decision making. In the nearest future, the model will be optimized for running on non-linear grids and more accurate physical analogy will be found for networking.

Acknowledgment

The paper is published with financial support by the project ITHEA XXI of the Institute of Information Theories and Applications FOI ITHEA (www.ithea.org) and the Association of Developers and Users of Intelligent Systems ADUIS Ukraine www.aduis.com.ua.

Bibliography

- [Unger, 2009]) L.-O. Lertsuwanakul and H. Unger. A Thermal Field Approach in a Mesh Overlay Network. In: 5th National Conference on Computing and Information Technology (NCCIT'09), Bangkok, Thailand, 2009
- [Singh, 2010] R. Singh et al. Ants Pheromone for Quality of Service Provisioning In Mobile Adhoc Networks. R. Singh, D. K. Singh and L. Kumar. In: International Journal of Electronic Engineering Research, Volume 2, Number 1 (2010), pp. 101–109, 2010
- [Yong Cui, 2003] Yong Cui et al. Multi-constrained Routing Based on Simulated Annealing. Yong Cui, Ke Xu, Jianping Wu, Zhongchao Yu, Youjian Zhao. In: Communications, 2003. IEEE International Conference on, Volume: 3 (2003), pp. 1718–1722, 2003
- [Lertsuwanakul, 2009] L.-O. Lertsuwanakul, H. Unger. An Adaptive Policy Routing with Thermal Field Approach. In: 9th International Conference on Innovative Internet Community Systems (I2CS), Lecture Notes in Informatics, pp.169-179, Jena, Germany, 2009
- [Vlasov, 1938] A. A. Власов. О вибрационных свойствах электронного газа. В: Журнал экспериментальной и теоретической физики, т. 8 (3), стр. 291, 1938 (Vlasov A.A. O Vibratsionnyh svoistvah electronnogo gaza. In: Journal of experimental and technical physics, Vol. 8(3), p 291. 1938)

Authors' Information

Olexandr Ya. Kuzemin – *Dr.ing. habil, Prof. of Information Department, Kharkov National University of Radio Electronics, Head of IMD, (Ukraine); e-mail: kuzy@kture.kharkov.ua*

Ievgen D. Kozlov – *Engineer (applicant for PhD) of Information Department, Kharkov National University of Radio Electronics; e-mail: engineer@ort.kharkov.ua*

METHODS OF ANALYSIS FOR THE INFORMATION SECURITY AUDIT

Natalia Ivanova, Olga Korobulina, Pavel Burak

Abstract: *In this article authors propose the analysis of the main information security audit methods: the active audit, the expert audit and the audit on conformity with standards applying SWOT-analysis. After this analysis authors make the suggestions for the future work in this area.*

Keywords: *information security, audit, threat, vulnerability, audit method, SWOT- analysis.*

Introduction

Information systems play an important role in our life. There is a big amount of important information flows that should be protected from unauthorized access and unauthorized modification. The information security systems are designed for the critical information protection.

The information security system is a complex of organizational, technical and legal measures that are used to protect information from unauthorized access and unauthorized modification in the process of receiving, processing, storage and transmission.

The information security systems correctness must be checked continuously to maintain the required security level. The information security audit is for this purpose.

The Information security audit is necessary for identifying the gaps in information security systems and, on the basis of the results, improving their protective functions. If the information security systems defense functions are not revealed in time, they may possibly cause the leak of confidential information which would adversely affect the information system image and would reduce the users' trust in it.

The information security threats classification

The Information security means maintaining the information confidentiality, integrity and availability and also the information authentication, the system reliability, control over the commitments' implementation [Standard, 2005]. Standard [Standard, 2005] also introduces two important definitions: the definition of the information security threats and the information security vulnerabilities. The threat is a potential cause of an undesirable incident, which may cause harm to the information system or the company. Vulnerability is a negative feature of an asset or a group of assets, through which one or more threats can be implemented. Thus, the information security may be compromised as a result of the information security threats, which are implemented through the vulnerabilities that exist in the information system.

All the information security threats can be classified, and today there are many different classifications. The authors of this research present a proper information security threats classification that is illustrated in figure 1. Table 1 presents all threats in order and each threat gets its own number. The main vulnerabilities for these threats are presented in table 2.

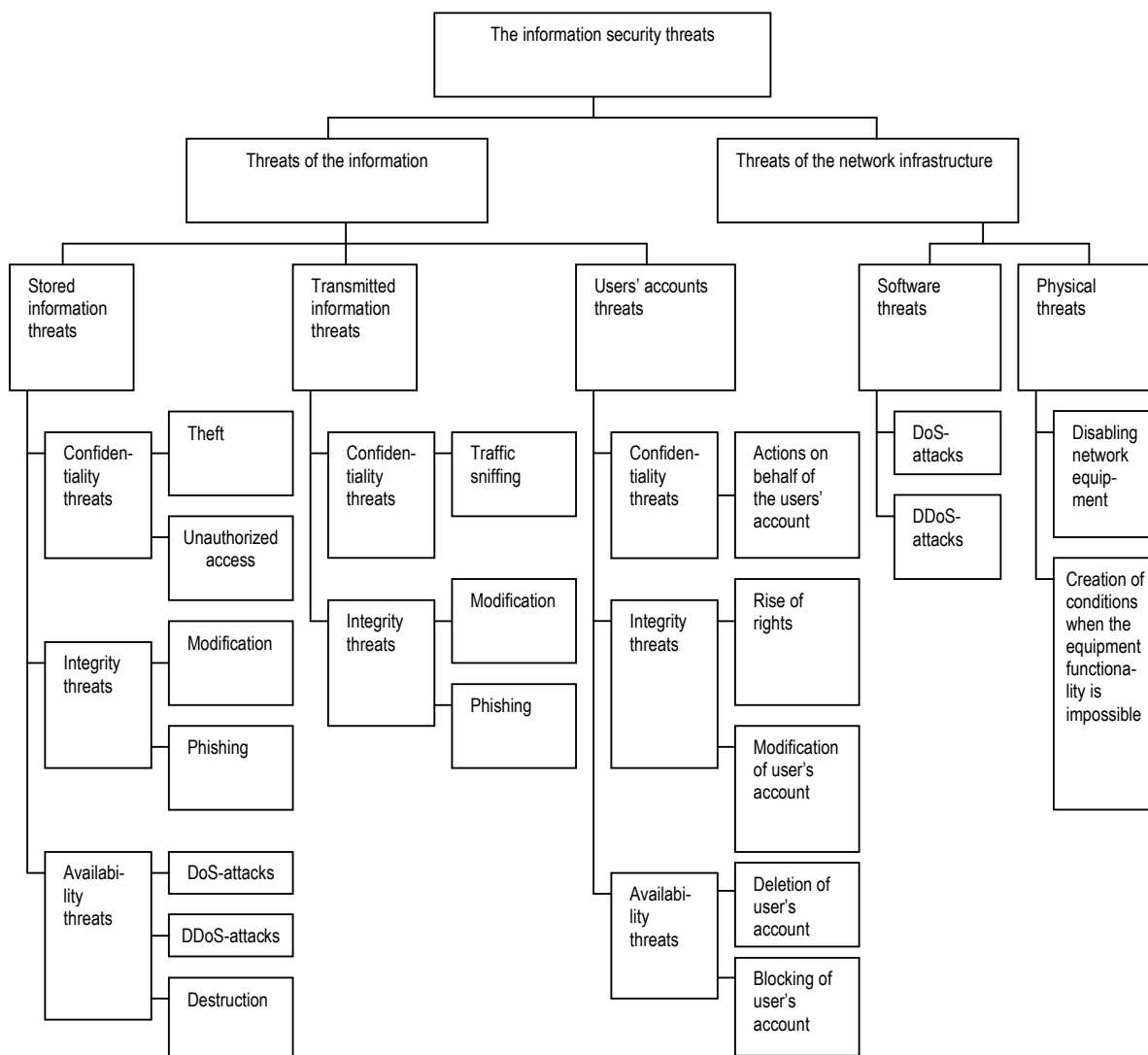


Figure 1 – Threats classification.

Table 1 – The list of threats.

No	The threat name
1	The stored information theft
2	Unauthorized access to the stored information
3	The stored information modification
4	The stored information phishing
5	DoS-attacks
6	DDoS-attacks
7	The stored information destruction

8	The traffic sniffing
9	The transmitted information modification
10	The transmitted information phishing
11	The transmitted information destruction
12	Actions in the system on behalf of the authorized user (masquerade)
13	The rise of rights
14	The users' accounts modification
15	The users' accounts deletion
16	The users' accounts blocking
17	The network equipment disabling
18	Creation of the conditions when the equipment functionality is impossible

Table 2 – The list of vulnerabilities.

Vulnerability	The ongoing threats numbers
The weak cryptographic policy	2, 3, 8, 9
The firewalls invalid configuration	1, 2, 4, 5, 6
The weak password policy	12, 14
Lack of check-point in the company	1, 17, 18
Harmful effects on the lines of force outside the company	11, 18
Disasters	11, 17, 18
Free access to communication channels outside the company	8, 10, 11
The incorrect implementation of the restricting access rules to the stored information	1, 2, 3, 4, 7
The intrusion detection systems and the intrusion prevention systems incorrect work	1, 2, 4, 5, 6, 7
The antivirus software incorrect configuration	3, 7, 9, 14, 15, 16
The use of unprotected data transfer protocols	8, 9, 10, 11
The users' rights incorrect settings	12, 13, 14, 15, 16
The integrity check mechanisms' incorrect settings	3, 9, 14
Uncontrolled technical channels of information leakage	8, 10, 11
The restrictions absence on the number of logon attempts	1, 2, 12

The information security methods

The information systems audit and control association (ISACA) provides the following definition to the term "the information security audit":

The information security audit is a process of the information gathering and the information analysis in order to establish:

- whether the organization's resources (including data) security is provided;
- whether the necessary parameters of the data integrity and the data accessibility are provided;
- are the organization's goals in the terms of the information technologies effectiveness reached.

Today there are three main the information security audit methods. They are presented in figure 2.

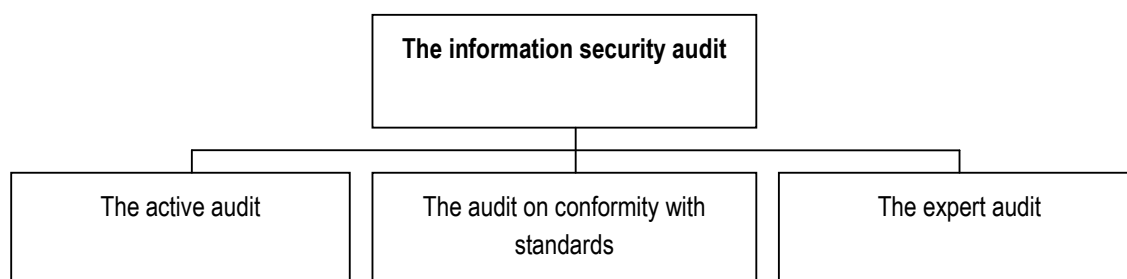


Figure 2 – The information security audit methods.

The information security active audit is a study of the information system information security from the perspective of a hacker or cracker who is highly skilled in information technologies [Prosyannikov, 2004].

During this audit method a large number of network attacks that hacker can implement is modeled. The same conditions in which the hacker works are artificially created for the auditor. The auditor is also provided only that information that can be found in open sources. The active audit result is the information about all the vulnerabilities, their severity degree and their elimination methods.

The expert audit is the information security comparison with the "ideal" description, which is based on:

- the CIO requirements;
- the "ideal" security system description based on accumulated in the audit company the worlds' and the private experience [Prosyannikov, 2004].

The performed during the expert audit actions are presented in figure 3.

The method of interviewing employees is used to collect the initial information. Technical specialists answer the questions related to the information systems operation, and the company's management explains the requirements that are applied to the information security system. The expert audit results can contain various proposals about modifying or upgrading the information security system.

During the audit on conformity with standards the information security state is compared with some abstract description found in the information security standards [Prosyannikov, 2004]. The information security standards are presented in figure 4.

The reasons for the audit on conformity with standards (and certification) may be divided into 4 categories, depending on the necessity of this service for the company:

- Compulsory certification;
- Certification due to the external objective reasons;
- Certification, which allow getting the benefits in the long term;
- Voluntary certification.

After this audit method the official report is generated. It contains the following information:

- The extent to which the information system matches selected standards;
- The extent to which the information system matches the company's' internal information security requirements;
- The number (and the categories) of disparities and received comments;
- Proposals about modifying or upgrading the information security system to bring it to conformity with standards;
- Detailed references to the company's key documents, such as security policies, descriptions of not obligatory standards and norms, applicable to the company.

Nowadays an increasing number of companies consider the certification as the confirmation of the high level information security. They use the received certificates as a "trump card" in the fight for a major client or business partner.

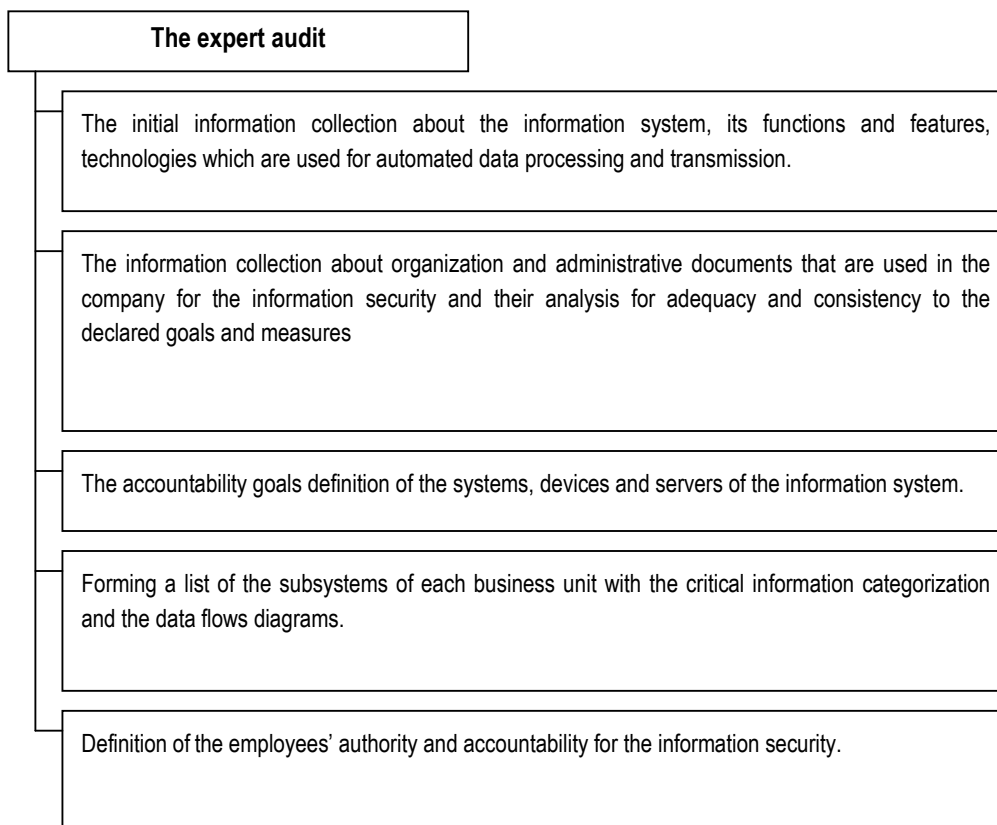


Figure 3 – The expert audit.

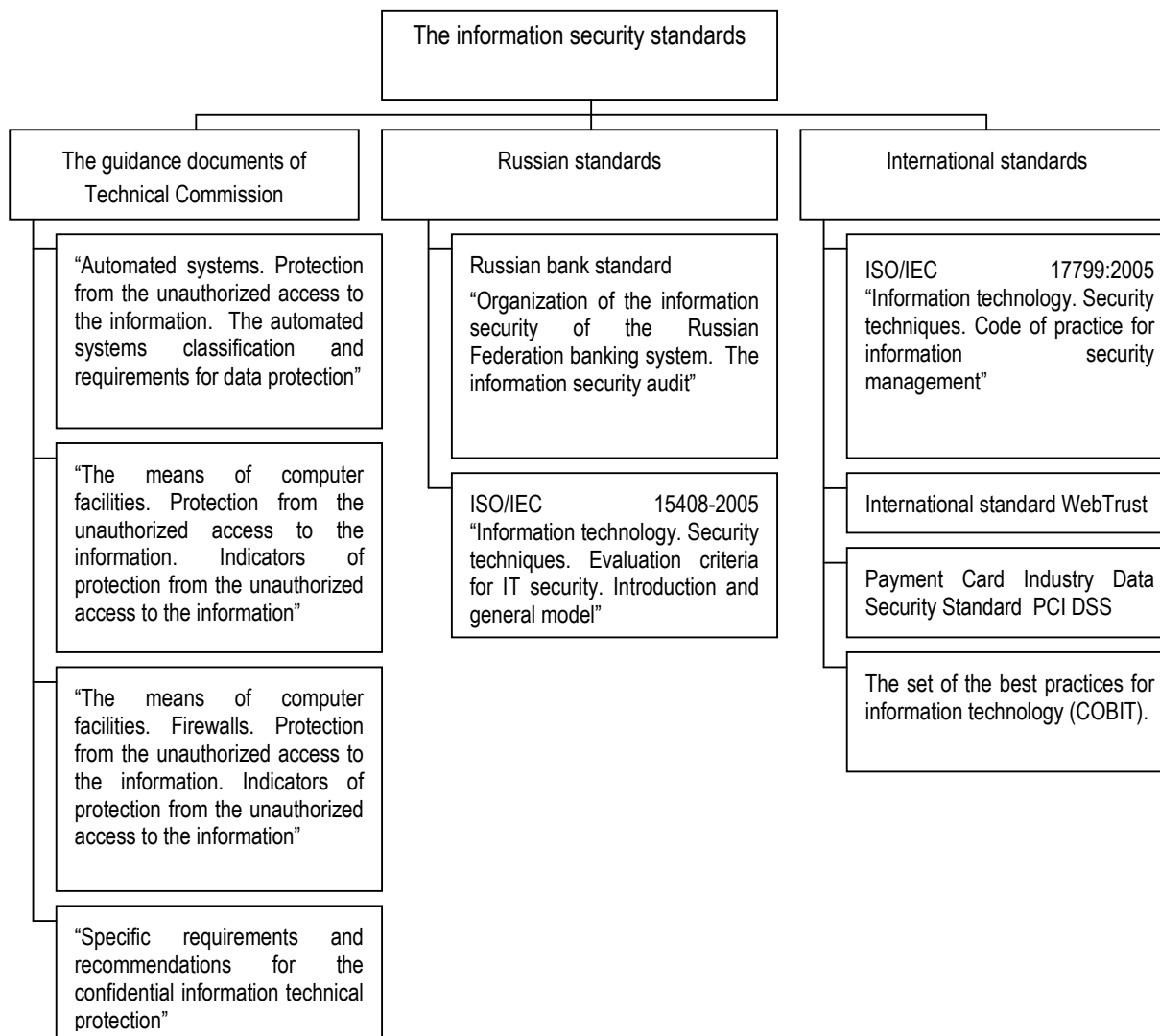


Figure 4 – The information security standards.

The audit methods analysis

The useful tool for the information security audit methods analysis is SWOT-analysis.

Initially, SWOT-analysis was created for different marketing researches but its mechanism is quite universal so the SWOT-analysis can be applied in other areas. SWOT-analysis aim is to determine the company's strengths and weaknesses (the internal environment analysis), as well as opportunities and threats of the nearest company's environment (the external environment analysis) [Gvosdenko, 2006]. The SWOT-analysis results are used to plan the company's strategy.

In this paper SWOT-analysis is used to determine the strengths and the weaknesses of the information security audit methods, as well as to identify external factors that are able to make the audit procedure easier or, on the contrary, more complicated.

Using the SWOT-analysis results, we can decide which audit method is the most perspective and how it is possible to develop its development strategy.

Tables 3, 4 and 5 represent the information security audit methods SWOT-matrixes.

Table 3 - The active audit SWOT-matrix.

Strengths	Weaknesses
<p>The audit process automation. The audit doesn't require the employees' participation. The audit frequency is not regulated. It is possible to carry out the stress test in order to determine the system productivity and stability, as well as the system resistance to DoS-attacks.</p>	<p>The additional software is required. Users should stop working with the information system before the audit beginning. Audit can identify only the known vulnerabilities.</p>
Opportunities	Threats
<p>High demand in the market. Audit can be carried out by the information security department staff. There is a large number of different software from various organizations. The biggest part of the auditors' work is automated.</p>	<p>The necessary software is expensive. Each system requires different software. The software can contain errors. There are no laws for this audit method.</p>

Table 4 - The expert audit SWOT-matrix.

Strengths	Weaknesses
<p>The additional software is not required. Users may work with the information system during the audit. The audit frequency is not regulated. The audit is based on the information security threats, thereby it is possible to cover a large number of vulnerabilities.</p>	<p>The employees should participate in the audit. The information provided by the client company should be precise. Preparative works can last long. The audit may occupy a considerable amount of time.</p>
Opportunities	Threats
<p>A big accumulated experience of the expert knowledge in the information security field. There are necessary regulatory documents. Audit can be carried out by the information security department staff. It is possible to atomize the audit process.</p>	<p>The absence of the audit process automation means. The need to trust the expert estimates. High requirements for the experts' competence. Potential conflicts among the experts' opinions.</p>

Table 5 - The audit on conformity with standards SWOT-matrix.

Strengths	Weaknesses
<p>The audit carrying out is regulated by normative documents. The reports' structure is described in the normative documents. Additional software is not required. Users may work with the information system during the audit.</p>	<p>The employees should participate in the audit. The audit should be carried out after every change in the information system. The information provided by the client company should be precise. The audit may occupy a considerable amount of time.</p>
Opportunities	Threats
<p>The security certificate, issued after the audit, raises the company's prestige. The best expert practices are reflected in the normative documents requirements. High demand in the market.</p>	<p>A large number of the normative documents. Constant changes in the normative documents. The contradictions in the normative documents. The audit can't be performed by the company itself, because the security certificate is issued only by the accredited organizations.</p>

In order to pass from the qualitative estimations to the quantitative ones, for all the factors listed in tables 3, 4 and 5 the authors identified the following values:

- The rate of the factor importance (F_impi);
- The observed value of the factor impact (F_infi);
- The uncertainty of judgments (F_probi).

The significance of each factor is calculated by the formula:

$$F_val_i = F_infi * F_probi \tag{1}$$

Then the total significance of all the factors for each parameter is as follows:

$$Val = \sum_{i=1}^n F_impi * F_val_i \tag{2}$$

The calculation results for each information security audit method are presented in table 6. Figure 5 illustrates the obtained values by the histogram.

Table 6 – The **parameters'** significance.

	The active audit	The expert audit	The audit on conformity with standards
Strengths	112,75	137,15	98,70
Weaknesses	147,20	147,00	154,35
Opportunities	174,80	184,95	163,80
Threats	165,45	181,70	184,40

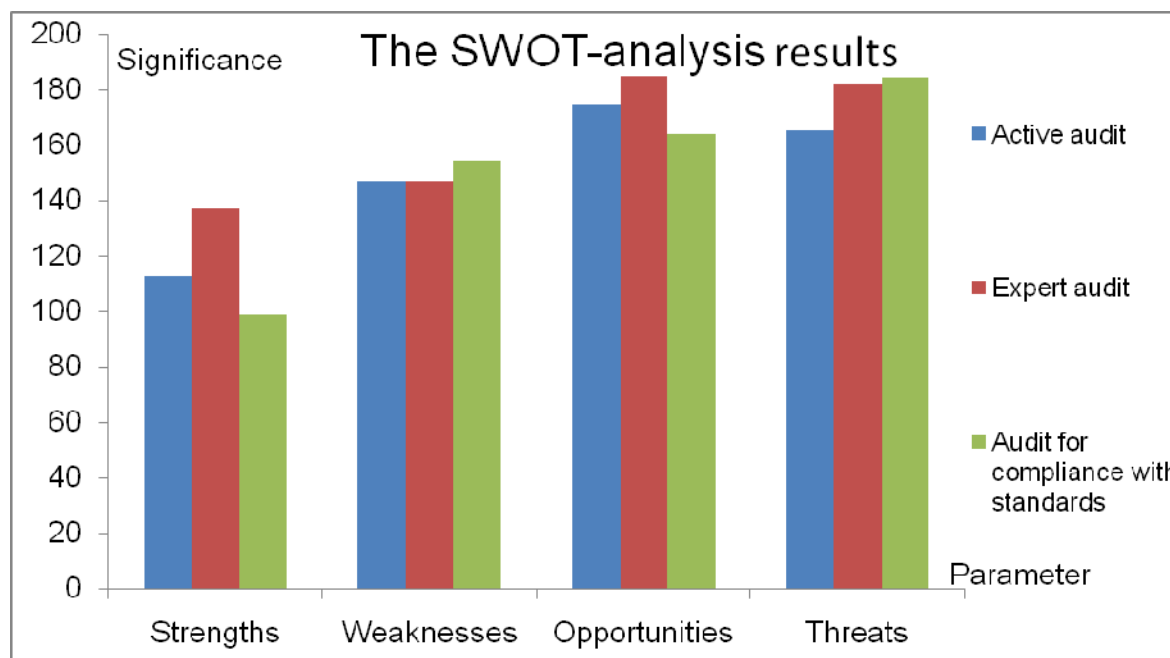


Figure 5 – The SWOT-analysis results.

From the table 6 and the histogram in figure 5 it is clear that expert audit has the best results. Consequently, this method is the most perspective information security audit method.

The expert audit requires good trained experts, but there are not many specialists of such a level. In addition, the task is poorly formalized and is based on the experts' personal experience and intuition. In this regard, we can conclude that in order to solve these problems we should use an expert system based on knowledge.

Conclusion

Information systems work with important and sometimes even critical information. This information should be protected from the unauthorized access and the unauthorized modification, to avoid harmful incidents. The information security systems are created for this purpose. Their correctness should be checked regularly, in order to maintain the demanded security level. The information security audit is the process that does this check. It is the difficult process demanding the knowledge of highly skilled experts. So in this case authors create the expert system which could help to make this difficult but very important and critical work.

Bibliography

- [Standard, 2005] ISO/IEC 17799 2005 "Information technology. Security techniques. Code of practice for information security management"
- [Prosyannikov, 2004] R.Prosyannikov. To get rid of errors: the information security audit methods. In: "Connect! The word of connection", №12/2004.
- [Gvosdenko, 2006] A. Gvosdenko. SWOT-analysis: methods of carrying out and application possibilities in the Russian enterprises. In: "Marketing and marketing researches", №2/2006.

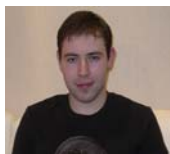
Authors' Information



Natalia Ivanova – *Ph.D.(Tech), Associated Professor of Petersburg State Transport University, department of Informatics and Information safety*
190031, Russia, Saint-Peterburg, Moskovskij prospect, 9
e-mail: natali_iv@rambler.ru



Olga Korobulina – *Post Graduate Student of Petersburg State Transport University, department of Informatics and Information safety*
190031, Russia, Saint-Peterburg, Moskovskij prospect, 9
e-mail: Olga_korobulina@list.ru



Pavel Burak – *Petersburg State Transport University, department of Informatics and Information safety*
190031, Russia, Saint-Peterburg, Moskovskij prospect, 9
e-mail: wistle3b@gmail.com

ON MEASURABLE MODELS OF PROMOTION OF NEGENTROPIC STRATEGIES BY COGNITION

Pogossian Edward

Abstract: *Could models of mind be independent from living realities but be classified as mind if the mind uses the same criteria to form the class mind? In the paper a constructive view on the models of mind, cognizers, is presented and the measurable criteria and schemes of experiments on mentality of cognizers are discussed.*

Keywords: *modeling, cognition, measures, negentropic, strategies.*

ACM Classification Keywords: *A.0 General Literature – Conference.*

1. Introduction

Due mind forms models of any realities including itself raises the question whether models of mind can be mental not being living realities (LR), assembled from LR or developed from the springs of LR?

In other words, whether are models of mind which do not depend from LR but are classified as mind possible if mind uses the same criteria when forms the class mind?

To answer the question constructive models of mind and criteria of measuring their mentality as well as the exhaustive experiments on revealing the truth are needed.

In what follows a measurable approach to the models of mind, cognizers, is presented and the criteria and experiments of testing of mentality of cognizers are questioned.

This approach to refining of cogs continues the approach started in [Pogossian,1983] and continued in [Pogossian,2005,2007] on interpretation of the recognized views on mind [Flavell,1962,Neuman,1966, Botvinnik,1984, Atkinson1993, Pylyshin,2004, Roy,2005, Winograd, 986,Mendler,2004,] by models having unanimous communalized meanings followed by experiments on validity of those models.

The paper describes the author's view on mental behavior and traditionally we should address to the readers by using words "our view" ,"we think», etc.

On the other hand, mental behavior, we assume, is identified with ourselves and we plan to discuss personalized and communalized constituents in communications.

That explains why we find possible in the paper to use the pronoun "I" for the mind along with "we" and "our" when they seem to be appropriate

2. A View on Mind

2.1. I am a *mind* and I am able to interpret, or *model* the *realities* I *perceive*, including myself, evaluate the quality, or *validity* of models and use those models to promote my *utilities*.

The models are composed from cause-effect relationships between realities, particularly between realities and utilities, and any composition of those relationships comprise the *meanings* of the realities.

The basic, or *nucleus* utilities and meanings are inborn while mind incrementally enriches them by assimilating and accommodating by Piaget [Flavel,1962, Mandler, 2004] cause-effect relationships between realities and already known utilities and meanings *solving* corresponding *tasks* and *problems* .

By Piaget “Mind neither starts with cognition of itself nor with cognition of the meanings of realities but cognizes their interactions and expanding to those two poles of interactions mind organizes itself organizing the world” [Flavell, 1962].

As much coincide ontology, or *communalized* (vs. *personalized*) meanings of realities with meanings of their models and as much those meanings are *operational*, i.e. allow to reproduce realities having equal with the models meanings, so better is the validity of the models.

In what follows a personalized model of mind, a view W , and a communalized version of W , *cognizers*, are presented with discussion of the validity of cognizers and schemas to meet the requirements.

2.2.1. Minds are algorithms for promoting by certain *effectors* the utilities of *living realities* (LR) in their games against or with other *players* of those games.

The players can be LR, assembles of LR like communities of humans or populations of animals as well as can be some realities that become players because not voluntarily but they affect LR inducing games with environments or the units like programs or devices that have to be tested and response to the actions of engineers. To compare and discuss some hypothetic mental realities like Cosmic Mind by Buddhists and Solaris by Stanislaw Lem are considered as players as well. Note, that descriptions of religious spiritual creatures resemble algorithm ones.

2.2.2. A variety of economic, military, etc. games can be processed by players. But all LR in different ways play the main negentropic games against overall increase of the entropy in the universe [Shrodinger, 1956].

In those negentropic games with the environments LR and their populations realize some versified *reproduction* and on-the-job selection *strategy elaboration algorithms* (r SEA).

The parent rSEA periodically generates springs of LR where each child of the springs realizes some particular strategy of survival of those children in on going environments. LR with successful survival strategies get the chance to give a new spring and continue the survival games realizing some versions of strategies of their parents while unsuccessful LR die.

2.2.3. The utilities of LR and their assembles initially are determined by their nucleus, basic interests in the games but can be expanded by new mental constructions promoting already known utilities. For example, the nucleus utilities of LR, in general, include the codes (genetic) of rSEA and algorithms for reconstructing rSEA using their genetic codes.

2.2.4. The periods of reproduction, the power of the springs and other characteristics of rSEA are kinds of means to enhance survival abilities of LR and vary for different LR depending, particularly, from the resources of energy available to LR and the velocity of changes of the environments of LR.

2.2.5. Minds can be interpreted as one of means to enhance the survival of LR. In fact, minds realize SEA but in contrast to on-the-job performance rSEA the strategies elaborated by minds are auxiliary relatively to rSEA and are selected by a priory modeling.

Correspondingly, the nucleus of mental LR in addition to rSEA codes include codes of mind developing algorithms like the adaptation algorithms by Piaget [Flavel, 1962, Mandler, 2004].

2.3. Thus, *modeling* SEA, or mSEA, do, particularly, the following:

- form the models of games and their constituents
- classify models to form classes and other mental constructions

- use mental constructions for a priori selection the most prospective strategies for the players
- elaborate instructions for the effectors of players using the prospective strategies.

The effectors transform the instructions into external and internal *actions* and apply to the environments of mSEA and mSEA themselves, correspondingly, for developing the environments and mSEA and enhancing the success of the players.

2.4. Whether are the models of mind which are not dependent from LR but are classified as mind possible if mind uses the same criteria when forms the class *mind*?

To answer to the question constructive models of mind and criteria of measuring their mentality as well as the exhaustive experiments on revealing the truth are needed.

2.5. Let's name *cognizers* the models of mind not depending from LR while the models of mental constructions name *mentals*.

Apparently, this ongoing view *W* on mind is a kind of cognizers, say, for certainty, *1-cognizers*, *1cogs* or *cogs* in this paper.

In what follows a constructive approach to cogs, the criteria and experiments of testing of mentality of cogs are presented.

3. Basic Approaches and Assumptions

3.1. Further refining of cogs extends the approach described above on interpretation of the recognized views on mind by models having unanimous communalized meanings followed by experiments on validity of those models to mind.

3.2.1. Later on it is assumed that cogs are object-oriented programs, say in Java.

All programs in Java are either classes or sets of classes.

Therefore, it is worth to accept that cogs and their constituents, mentals, are either Java classes or their compositions as well.

3.3. Accepting the above stated assumption the experiments on quality of cogs were run for SSRGT games.

Particularly, because chess represents the class and by variety of reasons is recognized as a regular environment to estimate models of mind [Botvinnik,1984, Pogossian,1983,2007, Atkinson,1993, Furnkranz, 2001] in what follows the constructions of mentals and experiments on mentality of cogs are accompanied, as a rule, by interpretations in chess.

3.4. Following to the view *W* cogs elaborate instructions for the effectors of players to promote their utilities. The effectors in turn transform instructions into *actions* applied to the players and their *environments*. They can be parts of the players or be constructed by cogs in their work.

It is assumed that certain *nucleus* mentals of cogs as well as the players and their effectors are predetermined and process in discrete time intervals while mentals of cogs can evolve in time.

The fundamental question on the origin of nucleus mentals and other structures needs further profound examination.

4. Refining Constituents of Cognizers

4.1.1. In general, *percepts* are the inputs of cogs and have the structure of bundles of instances of the classes of cogs composed in discrete time intervals.

The *realities* of cogs are refined as the causes of their *percepts*.

The *environments* and the *universe* of cogs are the sets and the totality of all *realities of cogs*, correspondingly.

More in details, the bundles of instances of attributes of a class *X* of cogs at time *t* are named *X percepts at t* and the causes of *X/t* percepts are named *X/t realities*.

It is worth to consider *t percepts* and *percepts* as the elements of the unions of *X/t* percepts and *t* percepts, correspondingly, and assume that there may be multiple causes for the same percept.

Analogically, *t realities* and *X/t realities* are defined.

In case percepts are bundles of instances of attributes of certain classes of cogs the realities causing them are the classes represented by those attributes.

Otherwise, cogs learn about the realities by means of the percepts corresponded to realities and by means of the responses of those percepts when cogs arrange actions by effectors.

Due cogs are continuously developed they start with percepts formed by nucleus classes followed by percepts formed by the union of new constructed and nucleus classes.

4.1.2. Cogs promote utilities by using links between utilities and percepts. They continuously memorize percepts, by certain criteria unite them in classes as *concepts* and distinguish realities to operate with them using *matching* methods associated with the concepts.

In addition some concepts are nominated by *communicators* to communicate about the realities of the domains of the concepts with other cogs or minds and enhance the effectiveness of operations of cogs in the environments.

4.2.1. The base criteria to unite percepts in concepts are *cause-effect relationships* (*cers*) between percepts, particularly, between percepts and utilities.

For revealing *cers* cogs **form and solve** tasks and *problems*.

Tasks are requirements to link given percepts (or realities) by certain *cers* and represent those *cers* in frame of certain classes.

4.2.2. The basic tasks are the *utility* tasks requiring for given percepts to find utilities that by some *cers* can be achieved from the percepts. In chess utility tasks require to search strategies for enhancing the chances to win from given positions.

The *generalization*, or *classification* tasks unite percepts (as well as some classes) with similar values into more advanced by some criteria classes and associate corresponding matching procedures with those classes to distinguish the percepts of the classes and causing them realities.

The *acquisition* tasks create new classes of cogs by transferring ready to use classes from other cogs or minds while the *inference* tasks infer by some general rules new classes as consequences of already known to cogs classes.

The *question* tasks can be considered as a kind of formation tasks inference tasks which induce new tasks applying syntax rules of question tags to the solutions of already solved tasks.

The *modeling* tasks require revealing or constructing realities having certain similarities in *meanings* with the given ones.

Before refining meanings of realities let's note that to help to solve the original tasks some approximating them model tasks can be corresponded.

4.2.3. *Problems* are compositions of homogeneous tasks and *solutions of problems* are procedures composing the solutions of constituent tasks.

The problems can be with *given spaces* of possible *solutions* (GSS) or without GSS, or the *discovery* ones.

Tasks formation and *tasks solving procedures* form and solve tasks types.

4.3.1. To refine the meanings of realities and mentals it is convenient to interpret the percepts, uniting them concepts, nucleus classes and the constituents of those mentals as the nodes of the *graph of mentals* (GM) while the edges of GM are determined by utility, cers, attributive, part of and other relationships between those nodes.

Then the *meaning of a percept C* can be defined as the union of the totality of realities causing C and the connectivity sub graph of GM with root in C.

The *meaning of a concept X* is defined as the union of the meanings of the nodes of the connectivity sub graph of GM with the root in X.

The *meaning of realities R* causing the percept C is the union of the meanings of the nodes of the connectivity sub graph of GM with the root in the percept C.

4.3.2. Later on it is assumed that the *knowledge* of cogs unites, particularly, the cogs, GM and their constituents.

4.4.1. Processing of percepts and concepts is going either *consciously* or *unconsciously*. While unconsciousness, usually, addresses to the *intuition* and needs the long way of research efforts for its explanation, the consciousness is associated with the named concepts and percepts in languages and their usage for communications. Particularly, the vocabularies of languages provide names of variety of concepts and realities causing those percepts.

Mind operates with percepts, concepts and other mentals while names realities causing those mentals when it should communicate.

Particularly, this ongoing description of cogs follows to the rules for named realities while internally refers to corresponding mentals.

4.4.2. When mind operates *internally* with the representations of realities it is always able to address to their meanings or to *ground* those representations [8].

For *external* communications mind uses representations of realities, *communicators*, which can be separated from the original carriers of the meanings of those realities, i.e. from the percepts of those realities, and become *ungrounded*.

The role of communicators is to trigger [12] the desired meanings in the partners of communications. Therefore, if partners are deprived of appropriate grounding of the communicators special arrangements are needed like the ones provided by ontologies. If the communicators are not sufficiently grounded well known difficulties like the ones in human-computer communications can rise.

Note, that if the model R' is a grounded reality the meaning of R' can induce new unknown aspects of the meaning of the original ones.

4.5. Realities R' represent realities R, or R' is a *model* of R, if meanings of R' and R intersect.

Model R' is *equal* to R if R' and R have the same meanings. The more is the intersection of the meanings of R and R' relative to the meaning of R the greater is the *validity* of R' . For measuring the validity of models a variety of aspects of the meanings of original realities can be emphasized. Particularly, descriptive or behavioral aspects of the meanings can be considered, or be questioned whether the meanings are views only of the common use or they are specifications.

5. Questioning Validity of Mind

5.1. Modeling problems require constructing realities having certain similarities in meanings with the original ones. When those realities are problems as well cogs correspond model problems to the original ones, run them to find model solutions and interpret them back to solve the original ones.

Apparently, solutions of problems are the most valid models of those problems but, unfortunately, not always can be found in frame of available search resources.

Valid models trade off between the approximations of the meanings of solutions of problems and between available resources to choose the best available approximations.

Due of that inevitable trade off the models are forced to focus on only the particular aspects of those solutions.

If communication aspects are emphasized the *descriptive* models and criteria of validity can be in use require the realities-models be equal only by communicative means of the communities.

On-the-job or *behavioral* criteria evaluate validity of models by comparing the performances of corresponding procedures.

The records of computer programs provide examples of descriptive models while when processed programs become the subject of behavioral validity. Sorts of behavioral validity provide functional testing and question-answer ones like Turing test.

Productive behavioral validity criteria compare the results of affection of the outputs of realities and their models on the environment. Fun Newman requirement on self-reproducibility of automata [Neuman, 1966] provides an example of productive validity. In its interpretation as *reflexive reproducibility* (RR) validity that criterion requires to construct 1-models of realities able to produce 2-models equal to the 1-models and able to chain the process.

5.2. To formulate criteria of validity of cogs it is worth to summarize the refined to this end views on mind as the following:

mind is an algorithm to solve problems on promotion of utilities of LR in their negentropic games

mind is composed from certain constituent algorithms for forming and solving tasks of certain classes including the utility, classification, modeling, questioning classes

mind uses solutions of problems to elaborate instructions for certain effectors to make the strategies of LR more effective and the environments of LR more favorable to enhance the success of LR in negentropic games.

5.3. Criteria of validity of cogs to mind have to answer whether cogs have meanings that minds have about themselves.

On the long way in approaching to valid cogs a chain of inductive inferences is expected aimed to converge eventually to target validity.

Inductive inferences unite science with arts and, unfortunately, the term of their stabilization can not be determined algorithmically. Nevertheless, what can be done is to arrange those inferences with the trend to converge to the target stabilization in limit [19].

To approach to valid cogs it is worth to order the requirements to the validity of cogs and try to achieve them incrementally, step by step.

The requirements v1- v4 to validity of cogs condition them to meet the following:

- v1. be well positioned relatively to known psychological models of mind
- v2. be able to form and solve the utility, classification, modeling and question tasks with acceptable quality of the solutions
- v3. be able to use the solutions of tasks and enhance the success of the players
- v4. be able to form acceptable models of themselves, or be able to *self modeling*

The requirements v2 - v4 follow the basic views on mind while v1 requires positioning cogs relatively, at least, to the recognized psychological models of mind to compare and discuss their strengths and weaknesses.

Note, that parent minds of LR reproduce themselves in the children minds in indirect ways using certain forms of cloning, heritage and learning procedures.

Some constituents of reproduction of LR can already be processed artificially, i.e. by regular for the human community procedures.

The requirement v4 is questioning, in fact, whether completely artificial minds, cogs, can reproduce new cogs equal themselves and to the biological ones.

5.4. What are the validity criteria to make cogs equal by meaning to mind and whether cogs valid by those criteria can be constructed?

It is a long way journey to answer to these questions and elaborate some approaches to implement.

6. Conclusion

Valid cogs, if constructed, confirm the assertion that mind is a modeling based problem formation and solving procedure able to use knowledge gained from the solutions to promote the utilities of LR in their negentropic games.

Synchronously, mental cogs provide a constructive model of mind as the ultimate instrument for cognition. Knowledge on the nature of instruments for revealing new knowledge gives a new look on the knowledge already gained or expected and raise new consequent questions.

Therefore, revealing by cogs the new knowledge on the instruments of cognition it is worth to question the new aspects of relationships between mind and the overall knowledge mind creates and uses.

Ongoing experiments on study of cogs are based on the technique of evaluating adaptive programs and their parts by local tournaments and use the game solving package with its kernel Personalized Planning and Integrated Testing (PPIT) and Strategy Evaluation units [Pogossian,1983,2005,2007].

Bibliography

- [Atkinson,1993] G. Atkinson Chess and Machine Intuition. Ablex Publishing Corporation, Norwood, New Jersey, 1993.
- [Botvinnik,1984] M.Botvinnik Computers in Chess: Solving Inexact Search Problems. Springer Series in Symbolic Computation, with Appendixes, Springer-Verlag: NY, 1984.
- [Flavell,1962] J. Flavell The Developmental Psychology of Jean Piaget, D.VanNostrand Company Inc., Princeton, New Jersey, 1962.

- [Furnkranz, 2001] J.Furnkranz Machine Learning in Games: A Survey in "Machines that Learn to Play Games", Nova Scientific, 2001.
- [Mandler,2004] Mandler J. The Foundations of Mind: Origins of Conceptual Thought. Oxford Univ. Press, 2004.
- [Neuman, 1966] John von Neuman.Theory of Self-reproducing Automata. University of Illinois Press, 1966.
- [Pogossian,2007] E.Pogossian. On Measures of Performance of Functions of Human Mind. 6th International Conference in Computer Science and Information Technologies, CSIT2007, Yerevan, 2007, 149-154
- [Pogossian ,2006] E.Pogossian. Specifying Personalized Expertise. International Association for Development of the Information Society (IADIS): International Conference Cognition and Exploratory Learning in Digital Age (CELDA 2006), 8-10 Dec., Barcelona, Spain (2006) 151-159
- [Pogossian,2005] E. Pogossian. Combinatorial Game Models For Security Systems. NATO ARW on "Security and Embedded Systems", Porto Rio, Patras, Greece, Aug. (2005) 8-18
- [Pogossian,2007] E.Pogossian, V. Vahradyan, A. Grigoryan. On Competing Agents Consistent with Expert Knowledge", Lecture Notes in Computer Science, AIS-ADM-07: The Intern. Workshop on Autonomous Intelligent Systems - Agents and Data Mining, June 5 -7, 2007, St. Petersburg, 11pp.
- [Pogossian,1983] E.Pogossian. Adaptation of Combinatorial Algorithms (a monograph in Russian), 293 pp. 1983. Yerevan.,
- [Pylyshyn,2004] Z. Pylyshyn Seeing and Visualizing: It's Not What You Think, An Essay On Vision And Visual Imagination, <http://ruccs.rutgers.edu/faculty/pylyshyn.htm>,2004.
- [Roy,2005] D.Roy Grounding Language in the World: Signs, Schemas, and Meaning Cognitive Machines Group ,The Media Laboratory, MIT (<http://www.media.mit.edu/cogmac/projects.html>) 2005.
- [Searle,1990] Searle J. Is the brain's mind a computer program? Scientific American 262, pp26-31, 1990.
- [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.
- [Shrodinger,1956] E.Shrodinger . Mind and Matter. Cambridge, 1956.
- [Winograd,1986] T.Winograd, F.Flores. Understanding Computers and Cognition (A new foundation for design). Publishers, Chapter 2, pp. 11–59, Huntington, NY, 1986.

Authors' Information



*Head of the Cognitive Algorithms and Models Laboratory at the Academy of Science of Armenia (IPIA) , professor at the State Engineering University of Armenia , Marshall Bagramyan Av.24, Academy of Sciences, Yerevan, 0019, epogossi@aua.am,
Major Fields of Scientific Research: models of cognition and knowledge processing, strategy elaboration, algorithms.*

SYSTEMOLOGICAL CLASSIFICATION ANALYSIS IN CONCEPTUAL KNOWLEDGE MODELING

Mikhail Bondarenko, Nikolay Slipchenko, Kateryna Solovyova,
Viktoriia Bobrovska, Andrey Danilov

Abstract: *It is difficult to exaggerate the importance, the urgency and complexity of “good” classifications creation, especially in knowledge management, artificial intelligence, decision making. To what extent it is possible within a short paper, the peculiarities and advantages of the new system method of the systemological classification analysis for the classifications of concepts creation were discussed. It is noted that the systemological classification analysis on the basis of the natural classification improves considerably the quality and the power of the classification knowledge models and ontologies, allows taking into account the deep knowledge of any, including ill-structured, domains. In the process of the research conduction the system models of the domain fragment of the ontologies on the basis of the parametric classification were created. Some results of the actual domain “Social Networks in Internet” analysis and modelling and the ontology fragments, realized in the ontologies engineering tool Protégé 3.2, are also considered. The systemological classification analysis application has allowed proving the obtained classifications of social networks functions, taking into account the objects essential properties. It has also successfully recommended itself for deep knowledge acquisition; the basic hierarchy of classes, “good” classifications and ontologies creation; possesses predictive power, simple logically relevant structure, ensures the possibility of the correct inference on knowledge.*

Keywords: *conceptual knowledge, knowledge systematization, natural classification, ontology, systemological classification analysis, social network, hierarchy, systemology, artificial intelligence.*

ACM Classification Keywords: *1.2 Artificial Intelligence – 1.2.6 Learning: Knowledge Acquisition*

Introduction

The development of knowledge management, artificial intelligence, decision making and many other actual scientific and practical directions is determined by knowledge and its quality. As we know, knowledge, intellectual capital is the main competitive advantage, the foundation of modern organizations, enterprises, society, human and nations' welfare and important component of decision making support systems.

In different spheres of knowledge acquisition and application conceptual models of subject domains (and of problem domains – in the terminology of E. V. Popov) play a leading role. "Historically," the species of domain models are: dictionaries, thesauri (in linguistics), conceptual models (infological, semantic models - in databases), UML diagrams (of classes, of use cases, ... - in object-oriented analysis and modeling), models of knowledge (semantic nets, frames, ... - in artificial intelligence), ontologies (from the viewpoint of the realization and application one of the most modern kind of a domain model, aimed primarily at the knowledge application in Internet).

The basis of such models is the relationships of the hierarchy between concepts (concepts classification), in the first place, the relations *genus-species* and *part-whole*, about two millennia known in formal logic. These relations in the theory of classification are called the relations of *taxonomy* and *meronomy*, in artificial intelligence – *genus-species*: *Isa* (class - class), *Instance-of* (class - element) and *part-whole*: *Part-of*; in object-oriented analysis and modeling – *generalization / specialization* and aggregation (in some cases, *composition*), respectively, etc. In

systemology to these relations corresponds one relation of the support of the functional ability of the whole, respectively, for system-classes and concrete systems (which are reflected in general and single concepts).

How effective are the methods of the concepts classification creation - the basis of modern models of knowledge of domains? The analysis shows that in most domains the classifications are subjective; many of them do not meet even the requirements of formal logic. That is why it is proposed to apply a new unique method of the systemological classification analysis based on the natural classification [E. A. Solovyova, 1999; E. A. Solovyova, 1991; E. A. Solovyova, 2000], which has successfully recommended itself for deep knowledge acquisition, the basic hierarchy of classes, "good" classifications and ontologies creation in all, including ill-structured domains.

Introduction to the Systemological Classification Analysis on the Basis of the Natural Classification

As noted, this work is not about data classification into existing classes. We work with knowledge classifications and besides with the conceptual deep knowledge, on the conceptual level, determine classes (entities), properties and relations, and besides in accordance with their position in the domain, in the reality, in accordance with the systemic of the reality. Naturalists and other scientists interested for many centuries in the problem of "good" classification creation, *the position of objects in which reflects the reality (the domain), is determined by essential properties and relations of objects* and therefore possessing predictive power. This "good" classification was called systematics, or the natural classification, the first meaningful criteria of which were introduced by the Englishman Wavell more than 150 years ago; then by A. A. Liubishchev, Y. A. Schrayder and other scientists, for example, the natural classification - *is a form of the laws of nature presentation..., expresses the law of the systems of reality relationship, allows to reach the maximum number of goals, because it takes into account the essential properties, etc.* Such criteria are useful for fundamental science, but are not constructive for computer modeling, application in knowledge models and ontologies. That is why in Knowledge Acquisition Laboratory and at the Social Informatics Department for more than 20 years the systemic research of conceptual knowledge and natural classification has been conducted. For the first time the constructive criteria of the natural classification and a new method of systemological classification analysis which allow to take into account deep knowledge, objects essential properties and relations in domain models in the most objective way, have been obtained [E. A. Solovyova, 1999; E. A. Solovyova, 1991; E. A. Solovyova et al; 2000, E. A. Solovyova, 2000, etc.]. This method for the first time synthesizes system and classification analysis. The natural classification criteria correspond completely to the formal-logical criteria and also deepen and generalize them.

These fundamental results have not only theoretical but also an important practical value. They allow creating knowledge models and ontologies which take into account essential properties and causal-investigative relations, possess predictive power, simple logically relevant structure, allow generalization and unlimited knowledge refinement without redesigning classification, ensure the possibility of the correct inference on knowledge, recommendations and decisions making support, interface with the concepts of natural language application.

It is proved mathematically and systemologically and (with the use of the category theory and the categorical-functorial model of the natural classification obtaining) that the natural classification is the parametric one (including properties of all its elements), in which the properties classification determines (isomorphic) the objects classification, the properties properties classification – deep layer properties – the properties classification, etc.). In practice, the consideration of one level of properties (their genus - species classification) allows making the classification model founded and really effective for solving on its basis the various tasks that require knowledge application.

Functional systemology - the systemic approach of the noospheric stage of science development – was created for and is aimed at complex, qualitative, ill-structured problems solving, it differs profitably from the traditional

systemic approaches and for the first time really takes into account the systemic effect. Systemology, taking into account the principles of systemic, integrity and diversity, considers all objects, processes and phenomena as systems functioning to support the supersystem functional abilities. Systemology as modern system methodology does not regard system as a set but as a functional object which function is assigned by supersystem. Systemology in particular allows overcoming problems of traditional methods of system analysis at the expense of using conceptual knowledge as well as formalizing procedures of analysis and synthesis of complex systems and creating knowledge-oriented software tools for their simulation. The development of the concrete (internal) systems systemology of G. P. Melnikov for the system of classes allows deep knowledge getting and modeling for all, including ill-structured, domains [Bondarenko et al, 1998; E. A. Solovyova, 1999].

Ontologies Analysis Fragment

Ontologies is the subject of interest in intelligence technologies, diverse research areas as there are applications of ontologies with commercial, industrial, academical and, many other focuses [N. Guarino, 1996]. They are used within a great number of domains, in this paper - in the domain "Social Networks" and accomplish many various tasks. As ontologies is the sphere of interest for numerous researchers and practitioners there are many definitions of the term "ontology" itself. The different aspects appearing in different ontologies definitions are mainly caused by the concrete ontologies applications within concrete problem domains, that is by the tasks for accomplishing of which the ontologies are created. The following ontology definition proposed by Gruber is considered as the most cited one: "ontology is a formal, explicit specification of conceptualization" [M. Auxilio, M. Nieto, 2003].

The diversity of ontologies tasks and applications causes, for its turn, the absence of a single ontology classification. This fact determines the need of the domain "Ontologies" research by means of the systemological classification analysis. The considered method usage allows, in particular, to evaluate each classification from the viewpoint of its validity, reflection of the objects essential properties in it, the possibility of the objects properties detection and prediction according to their place in the classification, from the viewpoint of the possibility of the classification application as a tool for theoretical analysis in the correspondent domain [E. A. Solovyova, 1999].

The existing ontologies classifications are numerous and various. This is explained by the fact that different researchers choose different division bases when they create their classifications. The analysis of the ontologies classifications by means of the systemological classification analysis involves the essential properties of the ontologies revelation in order to determine their place in the Natural Classification (NC) and to connect these properties with the supersystem functional query. To find the supersystem functional query it is necessary to answer the question "What ontologies are needed for?". The ontologies are used for knowledge representation in many different domains. The analysis of the mentioned above division bases used for ontologies classification shows that the division bases must take into account the ontologies functional aspect. The use of another division bases (for example, connected with the structural aspect) will not allow creating the ontologies classification corresponding to the NC criteria. The use of the systemological classification analysis allows taking into account the essential properties of the considered ontologies during the knowledge systematization about them. The ontologies classifications analysis shows that there are two main types of the division bases proposed to classify the ontologies. They are: division bases reflecting the functional aspect of the ontologies, division bases reflecting the structural aspect of the ontologies. The analysis of the ontologies classifications by the division bases reflecting the functional aspect has given the following results: 1) the division bases declared by the authors often do not coincide with the real division bases; 2) the concepts definition rules of the formal logic are often violated; 3) the ontologies essential properties often are not taken into account (it concerns some ontologies classifications by the functional aspect and all the ontologies classifications by the structural aspect).

In the process of the research conduction the system models of the ontologies structure such as ontologies parametric classification fragment, semantic nets were created. The obtained version of the ontologies ontology was realized in Protégé 3.2 in the form of a frame net.

The full results of this research have been reported on the conferences and their fragment has been included into the research work for the examination in the course Knowledge Management – Knowledge Technologies (Social Informatics Department, KhNURE and Stockholm University) in 2010.

Social Networks Functions Classification

The conceptual knowledge modelling will be accomplished on the example of the actual domain of **social networks, including the ontology creation**. Nowadays the need to solve complex problems requiring the knowledge of the domain specialists appears increasingly. To train highly qualified professionals progressive companies propose to use the conception of learning organizations. A learning organization as a tool for solving problems related to the company professional level improving. To create and acquire knowledge the company needs to be constantly in the process of self-improvement. One of the advanced methods of the organization development is the social networks use. The social networks in Internet functions research will allow understanding better the expediency of their use, to use the social networks more effectively in decision making, for further knowledge systematization in the social networks domain.

Resulting from the research the developed social networks classifications were not found. There are several articles where the social networks in Internet functions but not their classifications are mentioned. For example, the following main functions of social networks in Internet are allocated:

- profiles, communities, blogs dogear, activities [Byelenkiy, 2008];
- functions of personal profiles creation, of users interactions, of common goal achieving by means of the cooperation, of resources interaction, needs satisfaction due to the resources accumulation [Kuzmenko, 2009].

The analysis shows that in the first division base, for example, the communication (messages interchange) functions class is absent. In the second division base the search functions are absent and it is also not clear what is meant by the functions of common goal achieving by means of the cooperation. The authors of the given divisions do not exemplify the functions which refer to the classes of these divisions.

Thus, the knowledge systematization in the domain of social networks is needed. Subsequently it will allow not only to obtain the social networks ontology but also to improve the considered nets from the functional viewpoint, to expand the set of their functions, to improve the meaningful placement of the menu functions in concrete social networks. The results of the social networks systematization may be applied for a new social network creation taking into account the advantages and disadvantages of the existing social networks.

In this case we consider the classification creation by the functionality as the knowledge systematization in the given domain [Solovyova, 1999]. The advantage of the proposed classification of the social networks in Internet functions is that it includes the functions considered in popular social networks «В Контакте.ру» (<http://vkontakte.ru>), «Википедия» (<http://ru.wikipedia.org/wiki>), «Мой Мир» (<http://my.mail.ru/mail>), «Connect.ua», «МойКруг» (<http://moikrug.ru>), «Science-community.org».

For these networks the functions classifications by the relation "part - whole" were created that has given the possibility to develop the recommendations or the meaningful placement of the menu functions of the social networks according to the requirements of systemology and formal logic. As an example, in Figure 1 our recommended classification of functions of the first level of hierarchy by the relation "part-whole" for the social network of scientists «Science-community.org», implemented in a software tool Protégé 3.2, is shown.

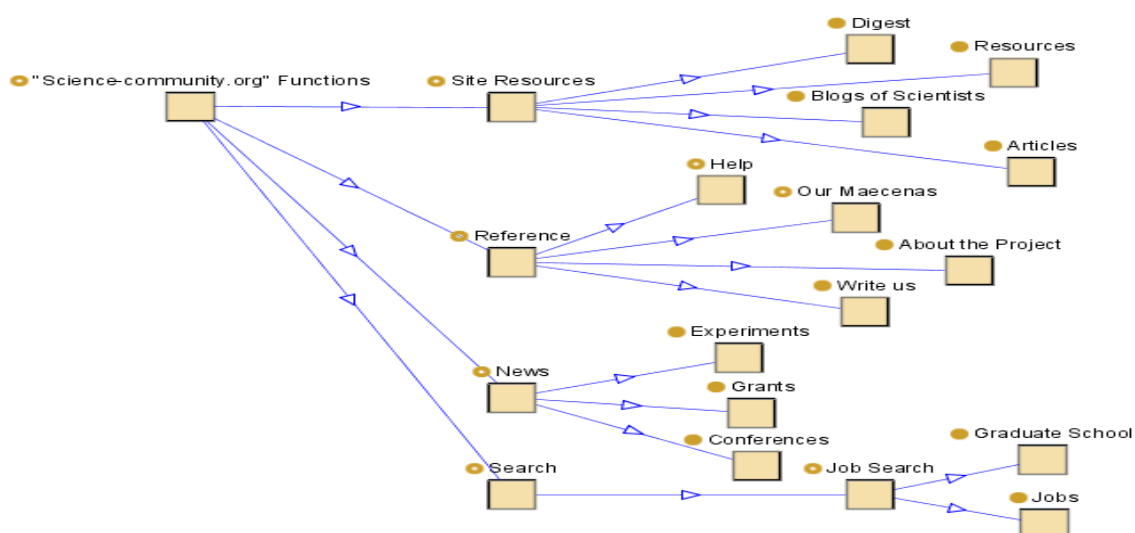


Figure 1. Recommended classification of the functions of the first level of the hierarchy for the social network «Science-community.org» by the relation «part-whole»

The systemological classification analysis application has allowed justifying the obtained classifications of social networks in Internet, to take into account the objects essential properties in them. This classification gives the possibility to detect and predict the objects properties by their position in the classification, i.e. from the viewpoint of the possibility to apply the classification not only as an effective practical tool but also as a tool of the theoretical analysis in the correspondent domain.

The use of the systemological classification analysis allows formulating recommendations for the hierarchical structure of functions implementation in the social network, for their meaningful placement in the menu in accordance with the created classification. Such natural placement will allow to reduce significantly the load on the user, will improve his work, networks and the principles of their functioning mastering.

The obtained classifications of the social networks in Internet functions allow to determine easy which class this or that concrete function of social networks refers to with which the user may meet while working with social networks in Internet. The greatest number of functions refers to the functions of "search" and "work with network resources," the functions of "communication" are also important. This classification of the social networks in Internet functions can be viewed as a parametric (including the classification of properties) one, because the classes functionality is seen from their names. Resulting from the functions of various social networks research the functions classification fragment, shown in Figure 2, was built. The created classification fragment allows determining to which class refer the functions of the first level of the hierarchy of the social networks: «В Контакте.ру» (<http://vkontakte.ru>), «Википедия» (<http://ru.wikipedia.org/wiki>), «Мой Мир» (<http://my.mail.ru/mail>), «Connect.ua», «МойКруг» (<http://moikrug.ru>), «Science-community.org». The functions search was done by means of the practical use of a concrete function to verify its functionality. First the functionality for each concrete function was determined, and then the function appurtenance to the concrete class was determined. The obtained fragment of the classification «social networks functions» was realized in the software tool Protégé 3.2. is shown in Figure 2. This software tool was chosen due to a number of advantages [Shcherbak, 2008, etc.].

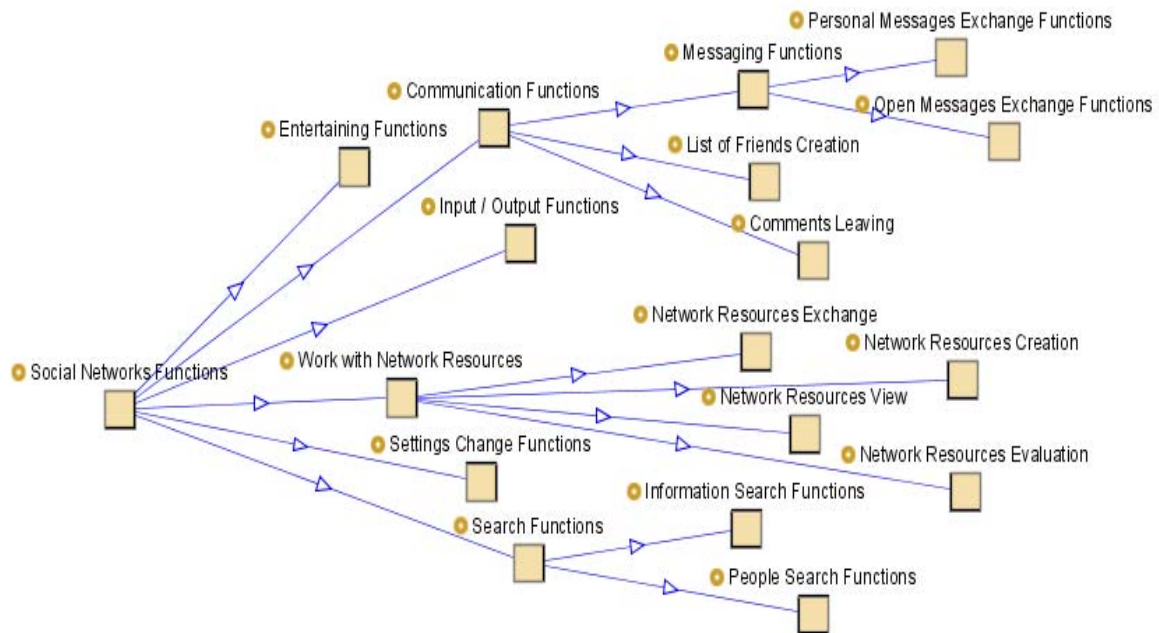


Figure 2. The fragment of the social networks in Internet functions classification by the functionality, realized in the software tool Protégé 3.2.

The obtained fragment of the social networks in Internet functions classification will allow becoming faster familiar with the functions of social networks in Internet, to choose more effectively the social network for registration, taking into account the functionality of the social network. The obtained results should be used for further knowledge systematization in the field of social networks in Internet.

Use of the Systemological Classification Analysis in the Social Networks Construction

Increase of the social networks in Internet influence of on the society has convinced many people to use social networks in business. Large corporations can afford to order a strong social network from firms of developers, but creation of such a network will require a lot of money. The enterprises (low-budget organizations) with a small income have not such a possibility, they may or attempt to use already functioning network or to attempt to create a social network by themselves. The latter variant is more advantageous, as the company itself regulates who will be the participant of the network, what tasks the social network must solve within the organization, etc. To create a social network in Internet it is necessary to use software for social networks creating.

Nowadays Internet is filled with a variety of software for the own social network creation. Many of them paid and (or) require deep knowledge in programming. There is also a number of software proposing to create a social network for free. This software proposes some free set of functions for a simple social network creating, there is also the possibility to use the supplement paid services.

The analysis of the software «Socialtext», «IBM Lotus Connections», «Jive SBS», «СвояСеть», «Connectbeam», «Ning», «Taba.ru» allows to make the conclusion that «Ning» (<http://www.ning.com/>), «Taba.ru» (<http://taba.ru/>), «СвояСеть» (<http://svoyaset.ru/getform.html#>) are the most acceptable for writing the recommendations to the social networks creation. They are conditionally free and do not require deep knowledge in programming. The disadvantage of the program service «Ning» is the absence of the interface in Russian. This

disadvantage is significant for the recommendations to the social networks creation. In connection with it the software «Табa.ru», and «СвояСеть» were chosen. While creating the social network in «Табa.ru» it is recommended to use the social networks in Internet classification fragment shown in Figure 2.

In the process of writing recommendations the alternative menu creation of the social network has been tested using the systematological classification analysis. The social networks functions alternative menu created taking into account the results mentioned above was maximally approximated to the menu corresponding to the formal logic and systemological classification analysis. Unfortunately, the considered designers have the limited functionality and do not allow applying fully the results of the conducted research. In the process of work guidelines and recommendations to social networks creation in Internet in the software «Табa.ru», «СвояСеть» have been developed, the shortcomings and benefits of a social network creation in the selected designers have been revealed, as examples the demoversions of social networks in each of the designers have been created.

The proposed results of the social networks may be used in the process of a learning organization creation, for decision making, intelligence technologies and artificial intelligence development.

Conclusion

The classifications of concepts are the basis of each science and are applied for solving various scientific-practical tasks. Now the classifications has got “the second birth” and are an integral element of ontologies, computer models of knowledge, object-oriented analysis and modeling, intelligence technologies, knowledge management, decision making and artificial intelligence, etc. That is why the role and the necessity of “good” classifications of concepts have increased even more. Systemology application has allowed synthesizing system and classification analysis, discovering new criteria of systematics (natural classification) and their applying for knowledge systematization in any domain. The results of the systemological research partially included in the paper may be used for the further knowledge systematization, creation of more effective alternative menus, etc.

Acknowledgements

The paper is partially financed by the project **ITHEA XXI** of the Institute of Information Theories and Applications FOI ITHEA and the Consortium FOI Bulgaria (www.ithea.org, www.foibq.com).

Bibliography

- [E. A. Solovyova, 1991] E. A. Solovyova. Mathematical Modeling of Conceptual System: a Method and Criteria of a Natural Classification (Article). New York: Allerton Press, Inc., V. 25, No. 2., 1991.
- [E. A. Solovyova, 1999]. E. A. Solovyova. Natural Classification: Systemological Bases [In Russian], Kharkov: KhNURE, 1999.
- [E. A. Solovyova, 2000]. Mathematical and Systemological Foundations of Natural Classification (Article). New York: Allerton Press, Inc., V. 33, No. 4., 2000.
- [E. A. Solovyova, et al, 2000] D.B. Elchaninov, S.I. Matorin]. Application Of Categories Theory To Research and To Modeling Of Natural Classification (Article). New York: Allerton Press, Inc., V. 33, No. 2., 2000.
- [Bondarenko et al, 1998] M. F. Bondarenko, E. A. Solovyova, S. I. Matorin. Foundations of Systemolgy, [In Russian], Kharkov : KhTURE, 1998.
- [N. Guarino, 1996]. N. Guarino. Understanding, Building, and Using Ontologies, Padova: Corso Stati Uniti, 1996. <http://ksi.cpsc.ucalgary.ca/KAW/KAW96/guarino/guarino.html>

[M. Auxilio, M. Nieto, 2003]. M. Auxilio. An Overview of Ontologies: Technical Report, Mexico: Center for Research in Information and Automation Technologies, 2003.

[Byelyenkiy, 2008]. A. Byelenkiy Business Perspectives of Social Networks // <http://www.compress.ru/article.aspx?id=18650&iid=865> [In Russian].

[Kuzmenko, 2009] Kuzmenko. Social Network // http://www.itpedia.ru/index.php/Социальная_сеть [In Russian].

[Shcherbak, 2008] S. S. Shcherbak. A Few Words about the Protocol Open Knowledge Base Connectivity (OCBC) and about the ontologies redactor Protégé // <http://semanticfuture.net/index.php?title> [In Russian].

Authors' Information

Mikhail Bondarenko – Rector of Kharkov National University of Radio Electronics, Corresponding Member of the National Academy of Sciences of Ukraine, Lenin Ave., 14, Kharkov, 61166, Ukraine; e-mail: rector@kture.kharkov.ua

Major Fields of Scientific Research: Intelligence Technologies, Information and Knowledge Management, System Analysis, Artificial Intelligence, Decision Making, Knowledge Research and Application, Natural Language Processing, Knowledge and Natural Language Modeling, Business Intelligence, Competitive Intelligence, Modern (e-) Learning, Knowledge-Based Systems and Technologies, Systemological Analysis.

Nikolay Slipchenko - Chief of Scientific Department, Kharkov National University of Radio Electronics, Professor, Doctor of Technical Sciences, Lenin Ave., 14, Kharkov, 61166, Ukraine; e-mail: slipchenko@kture.kharkov.ua

Major Fields of Scientific Research: System Analysis, Intelligence Technologies, Information and Knowledge Management, Modern (e-) Learning, Competitive Intelligence, Knowledge-Based Systems and Technologies, Artificial Intelligence, Decision Making, . Systemological Analysis.

Kateryna Solovyova - Chief of Social Informatics Department and Knowledge Management Center, Professor, Doctor of Technical Sciences, Kharkov National University of Radio Electronics, Lenin Ave., 14, Kharkov, 61166, Ukraine; e-mail: si@kture.kharkov.ua

Major Fields of Scientific Research: Knowledge Classification, Systematization, Elicitation, Acquisition and Modeling, Knowledge Management, Ontological Engineering, Systemological Analysis, Knowledge Research and Application, Decision Making, Knowledge-Based Systems and Technologies, Artificial Intelligence, Business Intelligence, Modern (e-) Learning, Competitive Intelligence, Cognitive Modeling.

Viktoriia Bobrovskia – Social Informatics Department, Kharkov National University of Radio Electronics, MA Student, Lenin Ave., 14, Kharkov, 61166, Ukraine; e-mail: si@kture.kharkov.ua

Major Fields of Scientific Research: Knowledge Modeling, Ontological Engineering, Competitive Intelligence, Decision Making Support, Artificial Intelligence, Knowledge Research and Application, Knowledge Management, Modern (e-) Learning, Intelligence Technologies, Systemological Analysis

Andrey Danilov – Social Informatics Department, Kharkov National University of Radio Electronics, MA Student, Lenin Ave., 14, Kharkov, 61166, Ukraine; e-mail: si@kture.kharkov.ua

Major Fields of Scientific Research: Social Networks, Ontological Engineering, Competitive Intelligence, Decision Making, Intelligence Technologies, Knowledge Research and Application, Knowledge Management, (e-) Learning, Artificial Intelligence, Systemological Analysis.

SEARCH AND ADMINISTRATIVE SERVICES IN ICONOGRAPHICAL DIGITAL LIBRARY

Desislava Paneva-Marinova, Radoslav Pavlov, Maxim Goynov,
Lilia Pavlova-Draganova, Lubomil Draganov

Abstract: Today there are a large number of digital archives, libraries and museums with rich digital collections representing the European cultural and historical heritage. The new challenge shifts from having online access to resources to making an effective use of them and avoiding information overload. A possible answer to this challenge is the approach, used for the development of the “Virtual encyclopedia of the Bulgarian Iconography” digital library (BIDL). It is a complete web-based environment for registration, documentation, access and exploration of a practically unlimited number of Bulgarian iconographical artefacts and knowledge. The key for its efficiency is the provision of strictly designed functionalities, powered by a long-term observation of the users’ preferences, cognitive goals, and needs, aiming to find an optimal functionality solution for the end users. A special attention was pay to search and administrative services, trying to cover a wide range of possible solutions such as keyword search, extended keyword search, semantic-based search, complex search, search with result grouping, tracking services, DL data exportation, etc. This paper presents these services in detail, their functional specifications and used algorithms. The ontology of the East-Christian Iconographical Art is discussed, because of its important role for the semantic description and search of iconographical artefacts and knowledge in the library.

Keywords: multimedia digital libraries, systems issues, user issues, online information services

ACM Classification Keywords: H.3.5 Online Information Services – Web-based services, H.3.7 Digital Libraries – Collection, Dissemination, System issues.

Introduction

In an attempt to answer the need for presentation and preservation of the Bulgarian iconography, a team from the Institute of Mathematics and Informatics has developed a multimedia digital library called Virtual Encyclopedia of Bulgarian Iconography (<http://mdl.cc.bas.bg>). It was designed so as to provide wide accessibility and popularization of the works of the Bulgarian iconographers, and moreover to enable future precise restoration of the icons at risk.

The “Virtual encyclopedia of the Bulgarian Iconography” digital library (also called Bulgarian Iconography Digital Library, BIDL)¹ is a complete web-based environment for registration, documentation, access and exploration of a practically unlimited number of Bulgarian iconographical artefacts and knowledge. It provides a rich knowledge base for the iconographical art domain, enabling its usage for content annotation, preview, complex search, selection, group and management. The ontology of the East-Christian iconographical art was developed and used for semantic annotation of the library content [Pavlov et al., 2010] [Pavlova-Draganova et al., 2007b] [Paneva et al., 2007].

¹ The first release of the BIDL was developed five years ago during the project “Digital Libraries with Multimedia Content and its Application in Bulgarian Cultural Heritage” (contract 8/21.07.2005 between the Institute of Mathematics and Informatics, BAS, and the State Agency for Information Technologies and Communications), aiming to lay the foundations of the registration, documentation, and the exploration of a practically unlimited number of Bulgarian icons [Pavlova-Draganova et al., 2007a] [Pavlov et al., 2006].

A very important task during the BIDL development was the provision of the strictly designed functionalities. A special attention was pay to search and administrative services, trying to cover a wide range of possible solutions such as keyword search, extended keyword search, semantic-based search, complex search, search with result grouping, tracking services, DL data exportation, *etc.*

This paper extends the BIDL functionality presentation of [Pavlov et al., 2010], where the content creation and preview services was mainly discussed. The paper structure is the following: Section 2 makes a short overview of the BIDL architecture, covering its main service panels, repositories and their relationships. The search and administrative services are presented in details in Section 4 and Section 5, where their functionality and base algorithms are described. The ontology of the East-Christian Iconographical Art knowledge is discussed in Section 3, because of its key role for the semantic search of iconographical artefacts and knowledge in the library. Section 5 summarizes the achieved results and traces the directions for future development of BIDL.

BIDL Architecture

The BIDL environment, depicted in figure 1, integrates:

- Appropriate repositories and services for management of two types of objects:
 - MDL objects – multimedia digital objects, described by technical and semantic metadata and saved in a *Media Repository*;
 - User profiles, presenting user's data and behavior, saved in a User Profile Repository.
- Two main service panels, named *Object data management* and *Administrative services*, provide the base BIDL functionality.

The *Object data management* panel refers to the activities related to: content creation: add (annotate and semantic indexing), store, edit, preview, delete, group, and manage multimedia digital objects; manage metadata (see [Pavlov et al., 2010]); search, select (filter), access and browse digital objects, collections and their descriptions. The *Administrative services* panel mainly provides user data management, data export, tracing and analysis services, presented bellow.

For every MDL object all semantic and technical metadata are saved in the Media repository. These metadata are represented in catalogue records that point to the original media file/s associated to every MDL object. The User profile repository manages all user data and their changes.

There are several internal relations between the separate components in the service panels. For example, in the Object data management panel:

- the *Add object services* are related to the *Preview and Edit services*;
- after the *Preview (services)*, the *Edit or Delete services* can be executed;
- the Search object services point to *Preview, Edit, Delete and Group objects services*;
- the *Group objects services* are related to *Preview services*;
- after the *Edit (services)*, the *Preview services* can be executed.

There are several relations between the components of the two main service panels, for example, the Tracing of MDL objects from the Administrative services panel is connected to Add object, Preview, Delete, Search, Edit and Group services from the Object data management panel.

All existing internal and external relations for the service panels provide the internal interoperability and the flexibility of the library.

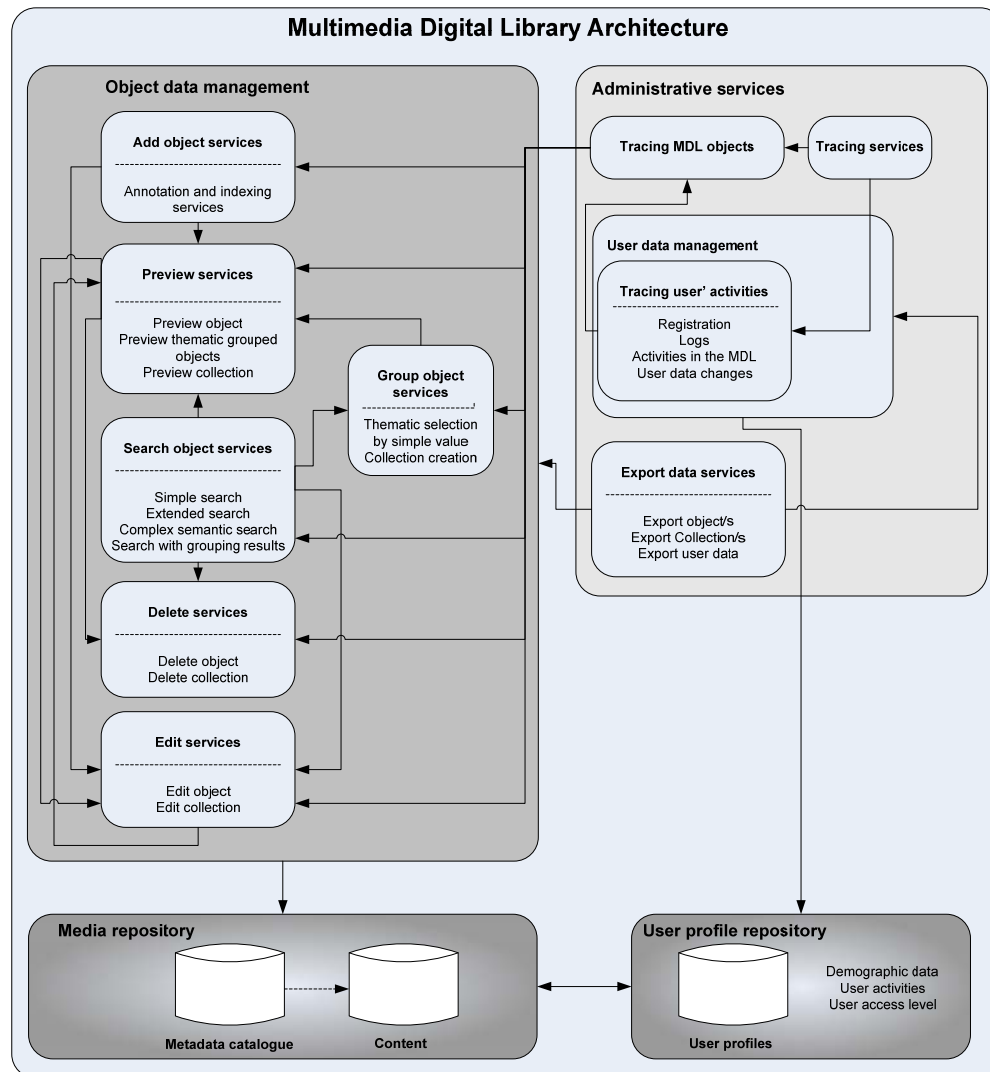


Figure 1: BIDL Architecture

Semantic Description of Iconography Art World

The semantic metadata description of Bulgarian icon art is determined by the domain ontology of the East-Christian iconographical art (also called iconography ontology)¹. It presented the iconographical art world by three "thematic entities"² (also called levels of knowledge). Every one of these entities is enriched with a set of sub-levels, covering wide range of characteristics. The first one is the "Identification" entity (see Figure 2), which consists of general data identifying aspects such as IO title, type, author and biographical data for the object's author, its clan, iconographic school, period, dimensions, current location and source, and object identification notes, iconographic school description.

¹ The ontology of the East-Christian iconographical art was developed for resource semantic annotation for the project SINUS "Semantic Technologies for Web Services and Technology Enhanced Learning" (№ D-002-189).

² A development methodology used during the creation of the *Ontology of the iconographical objects (artefacts)* [Paneva et al., 2007] [Pavlova-Draganova et al., 2007b].

Identification entity of the Ontology of the East-Christian iconographical art

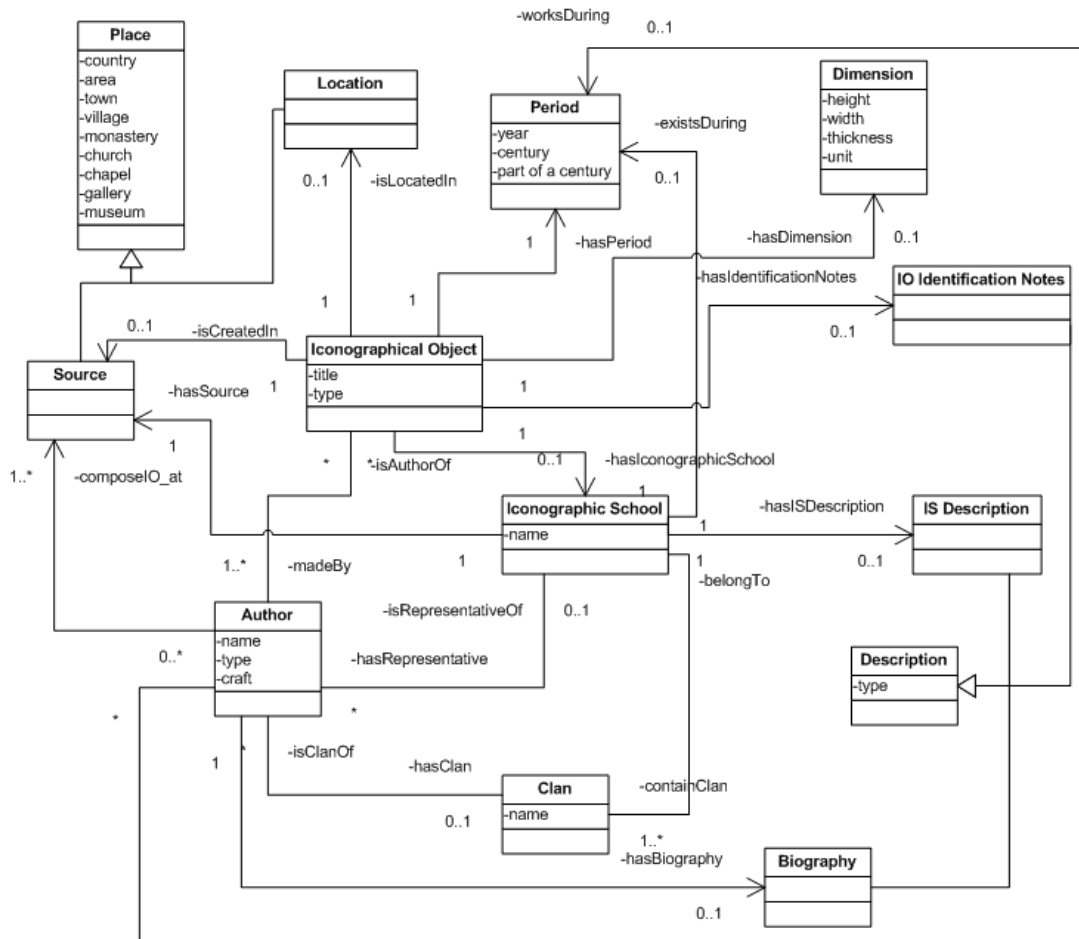


Figure 2: Identification entity of the Ontology of the East-Christian iconographical art

The second entity (see Figure 3) covers information concerning the descriptive details of the theme and forms of representation, providing a better understanding of the content.

The main concepts included are: depicted character/s, iconographical scenes, character/s in the scene/s, symbol/s in the scene/s, character' gestures, character' vestment, detailed description of the depicted content, *etc.* The third entity covers technical information revealing iconographic techniques, base materials, varnishes, gilding, *etc.*, used in the creation of the iconographical object/collection, and also concerning examinations of the condition, such as diagnosis or history of the conservation treatment.

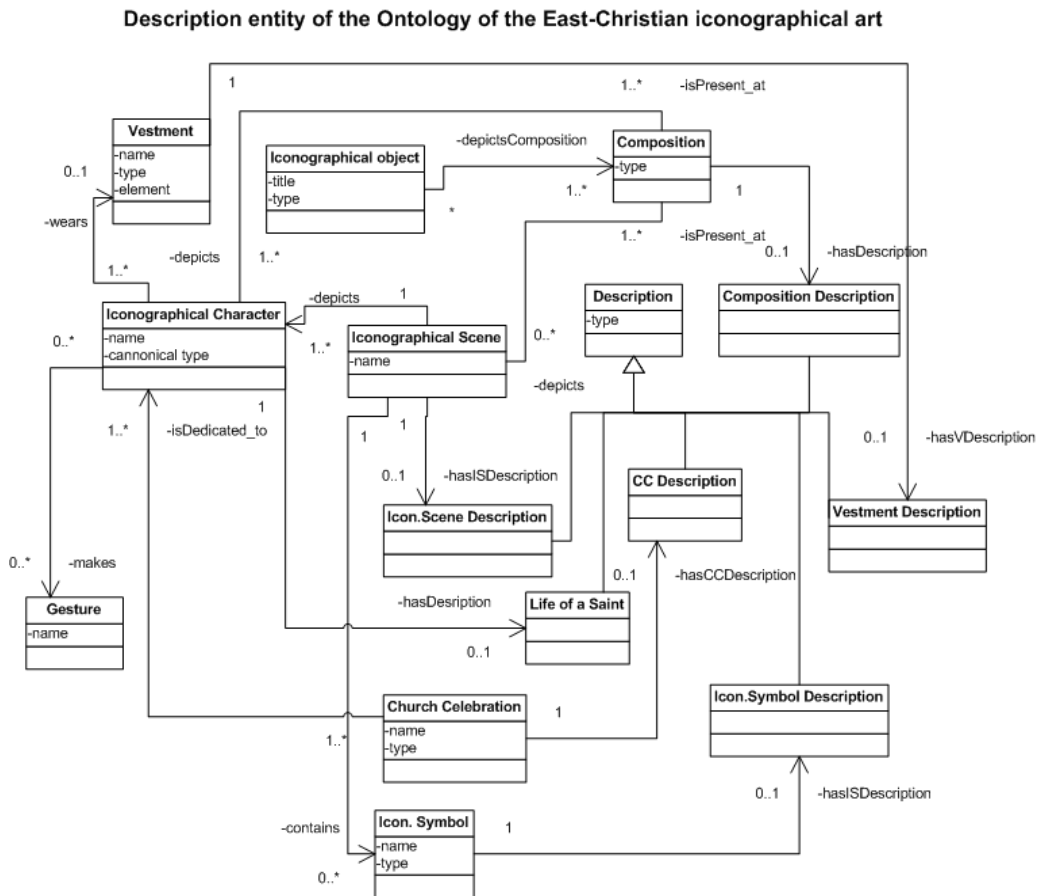


Figure 3: Description entity of the Ontology of the East-Christian iconographical art

Search Services

BIDL provides a wide range of search services, such as keyword search, extended keyword search, semantic-based search, complex search, and search with grouping results. This section presents the complex search algorithm that is base of all other search possibilities.

Let $U = O \times C$, O is the set of objects and C is the set of characteristics and U is the set of all objects and their characteristics. Let $v(o,c)$ is a function : $v : O \times C \rightarrow V$, where V is the set of values of the characteristics.

$p(c,v)$ is a condition for the characteristic c and the value v . In the first version of our search service, there was only one type of condition: $p(c,v) \Leftrightarrow$ "objects having value v for characteristic c ". Let P be the set of all possible conditions for $c \in C$ and $v \in V$.

Let define the search function $s(p,u)$, where $p \in P$ and $u \in U, s : P \times U \rightarrow U$. The result is a set $S \subseteq U$.

Let assume that we search on n characteristics.

So, in the first version of our searching service, we used the following algorithm:

$$S_1 = s(p_1, U) \rightarrow \text{time for execution} = t_1 = t$$

$$S_2 = s(p_2, U) \rightarrow \text{time for execution} = t_2 \approx t$$

$$S_3 = s(p_3, U) \rightarrow \text{time for execution} = t_3 \approx t$$

...

$$S_n = s(p_n, U) \rightarrow \text{time for execution} = t_n \approx t,$$

where p_n are the conditions for all n characteristic.

The result of our search will be:

$$R = S_1 \cap S_2 \cap S_3 \dots \cap S_n \text{ (See figure 4)}$$

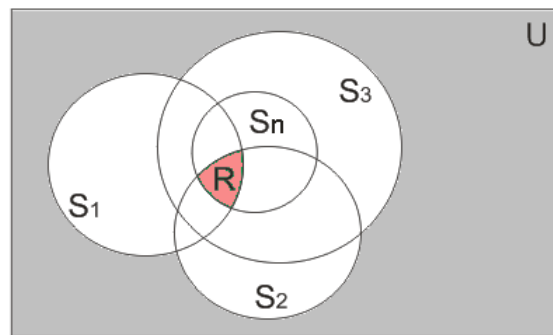


Figure 4: Result set of the search in the first release of the BIDL

If we assume that the time for making one search iteration over the all set of objects and their characteristics U is t , therefore the execution of the whole algorithm will spend $t_{v1} = t_n$ time + the time needed for the intersection of the results in the first release of searching services.

The current version of the searching service had the following changes:

The types of conditions raised to 5:

1. "objects having value = v for characteristic c " – the same as in version 1
2. "objects having value $\neq v$ for characteristic c "
3. "objects having numeric value $\geq, \leq, <, >, or = v$ for characteristic c "
4. "objects having characteristic c "
5. "objects NOT having characteristic c "

The algorithm for the search function changed to (see figure 5):

$$S_1 = s(p_1, U) \rightarrow \text{time for execution} = t_1 = t$$

$$S_2 = s(p_2, S_1) \rightarrow \text{time for execution} = t_2 \leq t_1$$

$$S_3 = s(p_3, S_2) \rightarrow \text{time for execution} = t_3 \leq t_2$$

...

$$S_n = s(p_n, S_{n-1}) \rightarrow \text{time for execution} = t_n \leq t_{n-1}$$

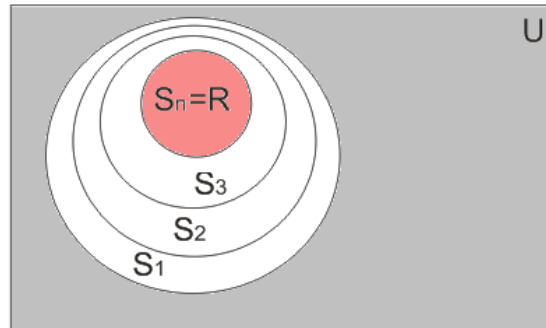


Figure 5: Set of results of the search in the current release of the BIDL

The result R will be equal to S_n , so no intersection will be needed. The time for execution $t_i \leq t_{i-1}$ is because at each iteration the search set $S_i \subseteq S_{i-1}$, therefore the time for processing a search decreases.

In this way, the overall time for execution will be $t_{v2} = \sum_{i=1}^n t_i \Rightarrow t_{v2} \leq t_n$ and $t_{v2} < t_{v1}$, t_{v2} is the time needed for results generation in the current release of searching services.

Administrative Services

The *Administrative services* panel mainly provides user data management, data export, tracking services, and analysis services. The user data management covers the activities related to registration, data changes, level set, and tracking activities of the user. The tracking services have two main branches: tracking of MDL objects, tracking of MDL user' activities (example, figure 6). The tracking of MDL objects spies on the activities of add, edit, preview, search, delete, selection, export to XML, and group of MDL objects/collections in order to provide a wide range of statistic data (for frequency of service usage, failed requests, etc.) for internal usage and generation of inferences about the stable work (stability), the flexibility, and the reliability of the environment. The tracking of MDL user' activities spies user logs, personal data changes, access level changes and user behavior in the BIDL.

The QlickTech® QlinView® Business Intelligence¹ software is the analysis services provider. It is connected to the BIDL tracking services and objects data base by preliminary created data warehouse¹.

¹ Business Intelligence is an architecture and a collection of integrated operational as well as decision-support applications and databases that provide easy access to great amount of (business) data.

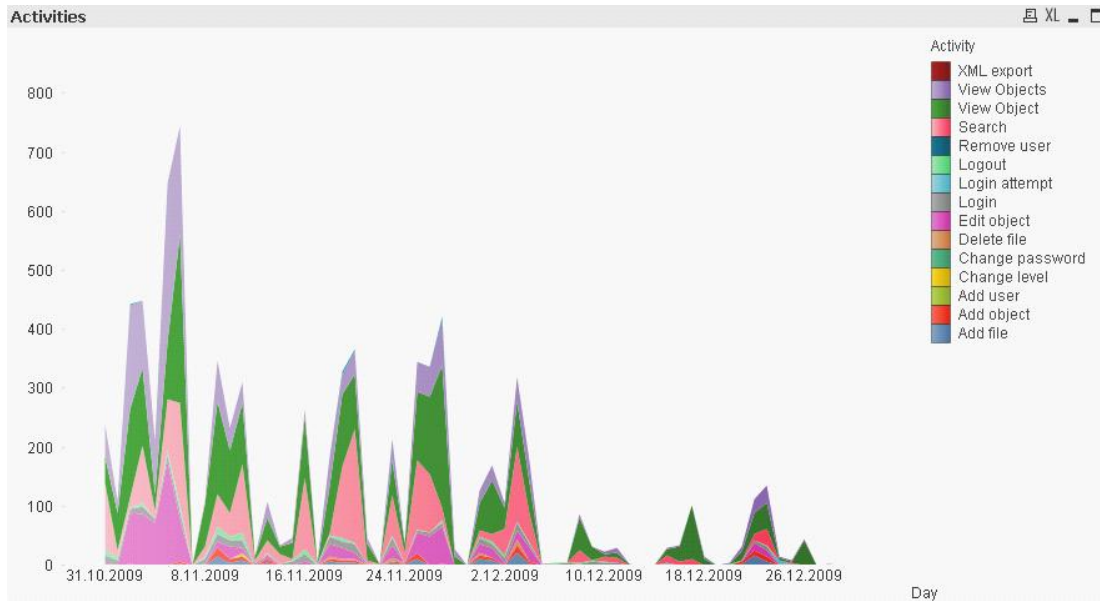


Figure 6: Users' activities during the period 10 – 12. 2009

The ETL (Extract, Transform, Load)² is completely automatic process and is performed by administrator request.

The QlickTech® QlinView® Business Intelligence Software is deployed in order to provide fast, powerful and visual in-memory analysis of the data in the warehouse. It is a data access solution that enables you to analyze and use information from different data sources. It is based on online analytical processing (OLAP), which provides an approach to quickly answer multi-dimensional analytical queries [Codd et al., 1993].

Figure 7 depicts PIE diagram making canonical sub-types analysis.

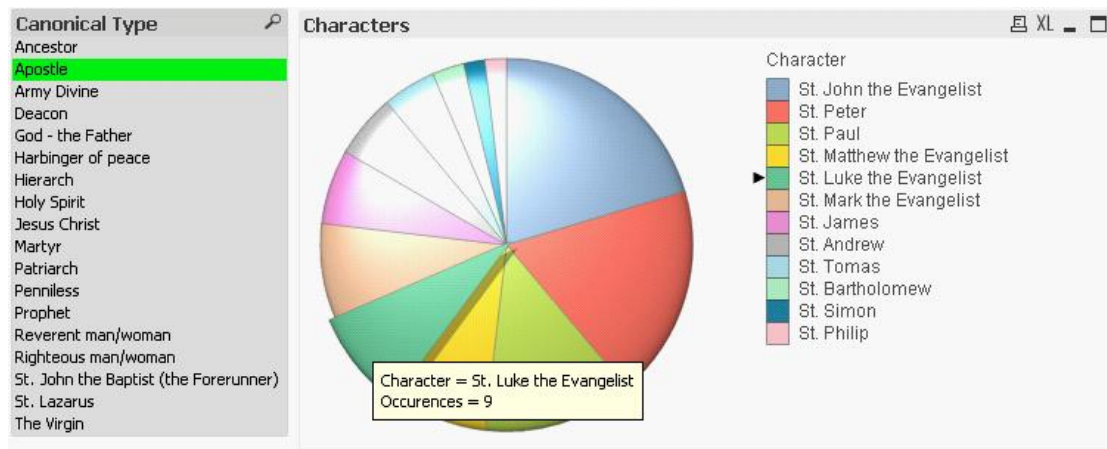


Figure 7: PIE diagram of canonical sub-types for Apostle canonical type

¹ A data warehouse is a repository of an organization's electronically stored data. Data warehouses are designed to facilitate reporting and analysis [Inmon, 1995].

² Extract, transform, and load (ETL) is a process in database usage and especially in data warehousing that involves: extracting data from outside sources, transforming it to fit operational needs (which can include quality levels), and loading it into the end target (database or data warehouse).

The variety of generated statistic information about BIDL data using QlickTech® QlinView® provides a rich extension of the tracking services and the base for profound analysis of extracted data.

The export data from the administrative services panel provides the transfer of information packages (for example, packages with BIDL objects/collections, user profiles, etc.) compatible with other systems managing data bases. For example, with these services a package with BIDL objects could be transported in a XML-based structure for a new external usage in e-learning [Paneva-Marinova et al., 2008] [Pavlov and Paneva, 2007] [Paneva-Marinova et al., 2009] or e-commerce applications.

The following code shows an instance of object data, exported in XML.

```
<object id="1">
  <characteristics>
    <chr name="Идентификация" id="1_0_0" value="">
      <chr name="Заглавие" id="2_0_0" value="">
        <chr name="lang:bg" id="3_0_0" value="Св. Богородица Катафиги (убежище) и св. Йоан Богослов"/>
        <chr name="lang:en" id="4_0_0" value="The Virgin Cataphuge (Refuge) and St. John the Evangelist"/>
      </chr>
      <chr name="Тип на иконографския обект" id="5_0_0" value="">
        <chr name="Икона" id="6_0_0" value=""/>
      </chr>
      <chr name="Автор" id="20_0_0" value="">
        <chr name="lang:bg" id="21_0_0" value="неизвестен"/>
        <chr name="lang:en" id="22_0_0" value="Unknown"/>
      </chr>
      <chr name="Иконописна школа" id="29_0_0" value="">
        <chr name="lang:bg" id="30_0_0" value="неизвестна"/>
        <chr name="lang:en" id="31_0_0" value="Unknown"/>
      </chr>
      <chr name="Период" id="32_0_0" value="">
        <chr name="От" id="33_0_0" value="">
          <chr name="Година" id="34_0_0" value="1395"/>
        </chr>
      </chr>
      <chr name="Размери (см)" id="53_0_0" value="">
        <chr name="височина" id="54_0_0" value="93"/>
        <chr name="ширина" id="55_0_0" value="61.5"/>
      </chr>
      <chr name="Местонахождение" id="57_0_0" value="">
        <chr name="Държава" id="58_0_0" value="">
          <chr name="lang:bg" id="59_0_0" value="България"/>
          <chr name="lang:en" id="60_0_0" value="Bulgaria"/>
        </chr>
        <chr name="Област" id="61_0_0" value="">
          <chr name="lang:bg" id="62_0_0" value="София"/>
          <chr name="lang:en" id="63_0_0" value="Sofia"/>
        </chr>
        <chr name="Галерия" id="85_0_0" value="">
          <chr name="lang:bg" id="86_0_0" value="Национална художествена галерия"/>
          <chr name="lang:en" id="87_0_0" value="National Art Gallery"/>
        </chr>
      </chr>
    </chr>
    <chr name="Описание" id="128_0_0" value="">
      <chr name="Персонажи" id="129_0_0" value="">
        <chr name="Име на персонаж" id="130_0_0" value="">
          <chr name="lang:bg" id="131_0_0" value="Св. Богородица Катафиги (Убежище)"/>
          <chr name="lang:en" id="132_0_0" value="The Virgin Cataphuge (Refuge)"/>
        </chr>
        <chr name="Каноничен тип на персонаж" id="133_0_0" value="">
          <chr name="lang:bg" id="134_0_0" value="Св. Богородица"/>
          <chr name="lang:en" id="135_0_0" value="The Virgin"/>
        </chr>
      </chr>
    </chr>
    <chr name="Технология" id="146_0_0" value="">
      <chr name="Иконографска техника" id="147_0_0" value="">
        <chr name="lang:bg" id="148_0_0" value="Темперка"/>
        <chr name="lang:en" id="149_0_0" value="Tempera"/>
      </chr>
      <chr name="Основа" id="153_0_0" value="">
        <chr name="lang:bg" id="154_0_0" value="Дърво"/>
        <chr name="lang:en" id="155_0_0" value="Wood"/>
      </chr>
    </chr>
  </characteristics>
  <files>
    <file id="1" original_name="1.jpg" savedas="1.jpg" />
  </files>
</object>
```

Conclusions and Future Work

A tendency from the last few years points towards the use of digital libraries as a source of digital knowledge and environment for its delivery. This tendency determines the development of new methods and techniques for functionality provision, aiming to satisfy user's needs and preferences. The new MDL systems aim to find optimal functionality solutions, mainly by improving the content annotation, search and presentation, metadata management, environment administration, *etc.* In this paper we presented the core and the motor of the BIDL architecture: the search and administrative services, covering the used techniques and algorithms. The next step will be the implementation of personalized work space and dynamic content adaptation service, assisting individual user's content observation. A profound research is done for the provision of innovative techniques and tools for digital preservation and restoration of valuable artefacts of the Iconographical art world. The investigations are also directed towards the development of tools for aggregating iconographical content and ensuring its semantic compatibility with the European digital library EUROPEANA, thus providing possibilities for pan-European access to rich digitalised collections of Bulgarian Iconographical heritage.

Acknowledgements

This work is partly funded by Bulgarian NSF under the project D-002-189 SINUS "Semantic Technologies for Web Services and Technology Enhanced Learning".

This work is partially supported by Project BG051PO001/07/3.3-02/7 as a part of the grant scheme "Support for the Development of PhD Students, Post-doctoral Students, Post-graduate Students and Young Scientists" under the Operational programme "Human Resource Development" of the European Social Fund and the Bulgarian Ministry of Education and Science.

Bibliography

- [Pavlov et al., 2010] Pavlov, P., D. Paneva-Marinova, M. Goynov, L. Pavlova-Draganova. Services for Content Creation and Presentation in an Iconographical Digital Library, International Journal "Serdica Journal of Computing", 2010 (in print)
- [Paneva et al., 2007] Paneva, D., L. Pavlova-Draganova, L. Draganov. Towards Content-sensitive Access to the Artefacts of the Bulgarian Iconography, In the Proceedings of the Fifth International Conference "Information Research and Applications" – i.Tech 2007 (ITA 2007 - Xth Joint International Scientific Events on Informatics), 26 June – 01 July, 2007, Varna, Bulgaria, vol. 1, pp. 33-38.
- [Paneva-Marinova et al., 2008] Paneva-Marinova, D., L. Pavlova-Draganova, R. Pavlov, M. Sendova. Cross-media and Ubiquitous Learning Applications on Top of Iconographic Digital Library. In the Proceedings of the 14th International Conference on Virtual Systems and Multimedia, Limassol, Cyprus, 20-25 October 2008, pp. 367-371.
- [Paneva-Marinova et al., 2009] Paneva-Marinova D., L. Pavlova-Draganova, L. Draganov, R. Pavlov, M. Sendova. Development of a Courseware on Bulgarian Iconography for Ubiquitous On-demand Study. In: Szucs A. (Ed.) Proceedings of Open Conference "New Technology Platforms for Learning – Revisited". Budapest, Hungary, January 2009, pp. 37-46.
- [Pavlov and Paneva, 2007] Pavlov R., D. Paneva. Toward Ubiquitous Learning Application of Digital Libraries with Multimedia Content, International Journal "Cybernetics and Information Technologies", vol. 6 (2007), № 3, pp. 51-62.
- [Pavlova-Draganova et al., 2007a] Pavlova-Draganova L., V. Georgiev, L. Draganov. Virtual Encyclopaedia of Bulgarian Iconography, Information Technologies and Knowledge, vol.1 (2007), №3, pp. 267-271.
- [Pavlova-Draganova et al., 2007b] Pavlova-Draganova, L., D. Paneva, L. Draganov. Knowledge Technologies for Description of the Semantics of the Bulgarian Iconographical Artefacts, In the Proceedings of the Open Workshop "Knowledge Technologies and Applications", Kosice, Slovakia, 31 May - 1 June, 2007, pp. 41-46.

- [Pavlov et al., 2006] Pavlov R., L. Pavlova-Draganova, L. Draganov, D. Paneva, e-Presentation of East-Christian Icon Art, In the Proceedings of the Open Workshop "Semantic Web and Knowledge Technologies Applications", Varna, Bulgaria, 12 September, 2006, pp. 42-48.
- [Pavlova-Draganova et al., 2009] Pavlova-Draganova, L. D. Paneva-Marinova. A Use Case Scenario for Technology-Enhanced Learning through Semantic Web Services, International Journal „Information Technologies & Knowledge”, vol. 3, 2009 (in print)
- [Codd et al., 1993] Codd E.F., S.B. Codd, C.T. Salley. Providing OLAP (On-line Analytical Processing) to User-Analysts: An IT Mandate. Codd & Date, Inc., 1993, pp. 1-31.
- [Inmon, 1995] Inmon, W.H. Tech Topic: What is a Data Warehouse? Prism Solutions. Volume 1. 1995.

Authors' Information



Desislava Paneva-Marinova – PhD in Informatics, Assistant Professor, Institute of Mathematics and Informatics, BAS, Acad. G. Bonchev Str., bl. 8, Sofia 1113, Bulgaria; e-mail: dessi@cc.bas.bg

Major Fields of Scientific Research: Multimedia Digital Libraries, Personalization and Content Adaptivity, eLearning Systems and Standards, Knowledge Technologies and Applications.



Radoslav Pavlov – PhD in Mathematics, Associated Professor, Institute of Mathematics and Informatics, BAS, Acad. G. Bonchev Str., bl. 8, Sofia 1113, Bulgaria; e-mail: radko@cc.bas.bg

Major Fields of Scientific Research: Multimedia and Language Technologies, Digital Libraries, Information Society Technologies, e-Learning, Theoretical Computer Science, Computational Linguistics, Algorithmic, Artificial Intelligence and Knowledge Technologies.



Maxim Goynov – Programmer, Institute of Mathematics and Informatics, BAS, Acad. G. Bonchev Str., bl. 8, Sofia 1113, Bulgaria; e-mail: maxfm@abv.bg

Major Fields of Scientific Research: Multimedia Digital Libraries and Applications.



Lilia Pavlova-Draganova – Assistant Professor, PhD Student, Laboratory of Telematics, BAS, Acad. G. Bonchev Str., bl. 8, Sofia-1113, Bulgaria; e-mail: pavlova.lilia@gmail.com

Major Fields of Scientific Research: Multimedia Digital Libraries, East Christian Iconography, 2D and 3D Computers Graphics and Animation, Design and Modelling Cultural Heritage Objects, Preservation and Restoration of Digitalized Artifacts, eLearning.



Lubomil Draganov – Assistant Professor, PhD Student, Institute of Mathematics and Informatics, BAS, Acad. G. Bonchev Str., bl. 8, Sofia 1113, Bulgaria; e-mail: lubomil@gmail.com

Major Fields of Scientific Research: Multimedia Digital Libraries, East Christian Iconography, 2D and 3D Computers Graphics and Animation, Design and Modelling Cultural Heritage Objects, Preservation and Restoration of Digitalized Artifacts.

THE INFLUENCE OF THE COMPUTER GAME'S VIRTUAL REALITY UPON THE HUMAN PSYCHOLOGY

Helen Shynkarenko, Viktoriya Tretyachenko

Abstract: *The influence of computer technologies upon the psyche of a person is depicted; 3D computer games as implementors of subconscious (including aggressive) desires of a person in particular and possible variant of psycho-correctional work with the usage of computer virtual reality.*

Keywords: *computer virtual reality, game addiction, subconscious, person's needs, self-estimation, aggression, mechanisms of psychological defense.*

ACM Classification Keywords: *K.4.2 Computers and Society - Social Issues*

Introduction

In translation from Latin "virtual" ("virtualis") means being in latency, but having an opportunity to become apparent, to happen. Projecting the subjective emotional states into the world of virtual realities through the interaction with the computer, a person often appears in captivity of the opportunity to change own mood artificially.

The dependence on the computer and computer games in particular has become one of the serious social and medical problems during recent years. The pathological addiction to the game is registered in all age groups of population and turns more and more into a kind of gambling epidemics. Increasing occurrence of pathological addiction to the game led a lot of people to the poverty and some of them to the suicide. Urgency of the problem of pathological addiction to the game is viewed in connection with three main reasons: 1) social and financial problems which pathological gamers have - 23% of gamers have financial problems, 35% of them are divorced, 80% have bad personal relationships in marriage; 2) existing of unlawful acts – nearly 60% of addicted gamers commit crimes; 3) high risk of suicides – from 13 till 40% among pathological gamers try to commit suicide, 32-70% have suicide thoughts.

This problem exists and demands researching. The present article does not claim to complete scientific explanation of the questions defined by this theme or to substantiation of theory. Its aim is to depict the trends of research and propose some corrections into existing today scientific concepts about interaction between a human being and a computer in the course of contact. We shall try to analyze some aspects of the influence of the computer virtual reality upon the human psychology and the reasons of deep, unconscious development of person's psychological addiction to the computer games.

The Psychology of Game Mania: Modern Status of Problem

The first computer game is considered to be "Tennis for two" invented in 1958 by American scientist William Higinbotham. Today we have several directions into which computer games have developed: game automatic machines, television-game devices and computer games themselves. In 80-90s a lot of researches were conducted in Western psychology devoted to the possible influence of the videogames upon a child. The first psychological research of the addiction to the computer was held by M. Shotton. At the end of 1994 Kimberly Yang worked out and placed on a web-site a special questionnaire. She obtained nearly 500 answers among which 400 were sent by addicted people according to selected by her criteria. In 1995 Goldberg proposed a set of diagnostic criteria for defining the addiction to the Internet, based upon the features of pathological addiction to

gambling. In 1997-1998 the research and consulting psychotherapeutic web-services devoted to this problem were developed. In 1998-1999 K. Yang, D. Grinfeld and K. Surrat published the first monographs. To the end of 1998 according to Surrat Internet addiction appeared to be legalized – not as a clinical trend in narrow sense, but as the sphere of researches and sphere of giving people practical psychological help. In the CIS countries the Internet addiction was viewed by psychologists and psychiatrists: A.E. Vojskunsii, M. Ivanov, Yu.V. Fomicheva, A.G. Shmelyov, I.V. Burmistrov, T.N. Dudko.

Modern researchers divide all psychological reasons of development of addiction into two categories:

The first category is the influence of environment (i.e. outer factors): the development of technical progress (including computer technologies); the manifestation of social anomia in modern society, i.e. traditional values do not exist any more and new system of norms and values does not formed yet (Merton, Darendorf, Curek and others); the formation by youth groups own subcultures which sometimes have disintegrated character; destroying of social institutions including families; hypo- and hyperguardianship of the child by parents.

The second category is connected with the peculiarities of the structure of the personality of the addicted person. A great impact into the studying of this problem was made by psychoanalytical researches. The main attention they pay to the subconscious and its influence upon the personality, to its role in the formation of addicted behaviour. Psychoanalytics do not make great differences between chemical (alcoholism, drag addiction and others) and emotional (work addiction, sex addiction, anorexia, computer game addiction and others) addictions. They suppose their inner mechanisms to be the same.

From the point of view of modern psychoanalysis addiction is the delitescence of suicide, i.e. the attempt to commit suicide which is prolong in time. Psychoactive object is used as the mean of self-destruction. The suicide itself is the attempt of going away from the disease, psychosis or the intention of overcoming the inner antagonism. Another aspect being paid attention to by psychoanalitics is identity. The difficulties in own identity's formation, its disturbance can become the additional risk factors in formation of addiction.

There is no common view concerning the mechanisms and reasons of addiction's formation among psychoanalytical researches but they are quite alike.

In domestic psychology the attitude to the entertaining games was quite scornful till the last time: "... in a whole gambling games are harmful as they lead to thoughtless squandering of time" (V.V. Rubtsov. A Learner before the computer: what is allowed or not // The Basics of Social-genetic psychology. – M., Voronezh, 1996). But in the process of development of innovational technologies in modern society a lot of specific problems have been accumulated: starting from the tendencies of current moment and the appearance of blending of game and movie till the opportunities of using computer programs in the sphere of education and problems of negative influence of computer technologies upon the consciousness of person. In connection with this the researchers pay attention to this sphere again. They analyze in details the opportunities and demands to the creation of developing teaching programs, the ways of organizing the work of a child and an adult in situation of using computers, the functions of the computer in learning activity. At the same time entertaining computer games, including plot games installed on personal computers and television-game devices are viewed superficially in the majority of cases. Though they gain increasing popularity and obtain one of the first places in the frequency of usage as personal entertainment. More than that, they become the first form of interaction with the computer available to a child.

It is difficult to overestimate the importance of deeper investigations of this problem because in spite of widely studied influence of computer games upon social adaptation in western researches the contradictory data about the connection between self estimation, social skills, successfulness of people and the time they spend playing computer games were obtained. On the one hand, the game is a supporting environment in which one can gain results and strengthen oneself but, on the other hand, the addiction to the computer intensifies the problems in the sphere of social contacts isolating the person, giving it the opportunity to go away from the problem. At the

same time the games themselves become the reason for communication, the topic of discussions and rivalry in obtaining the game's results, i.e. they can serve as a mean of socialization.

In domestic psychology the level of development of this problem is not high enough. Especially little attention is paid, to our mind, to the influence of the interaction of a person with the computer upon deep, subconscious sides of human psychology.

The Mechanism of Formation of Addiction to the Role Computer Games

Succeeding to psychoanalytics, we consider as well that under the influence of negative social conditions the person seeks the ways to release the inner tension formed as a result of inappropriate parental upbringing (emotional rejection; hypoguardianship, hypoprotection, hyperguardianship, contradictory upbringing in the family). Additionally the passion to the computer games forms within a player the needs which do not contribute to the person's adaptation to the environment, to the society in which he/she lives. That is, a vicious circle arises and only the person motivated to self changes can ruin it under the influence of awareness of subconscious (hidden deep in the mentality) motives of own behaviour.

In connection with this we held the investigation of three groups of people having the easy access to the computer games. The respondents were proposed a set of questions giving an opportunity to find out the level of their involvement in playing activity and the test for studying the level of self estimation with the aim of learning the connection between the affection to games and self estimation. The inquiry was held among first-year students (230 participants) of Kirovograd state pedagogical university (Group I). It was held in two steps: in April and in November of 2009. Additionally 14 people who are engaged into the development of new programs (programmers) were questioned. Besides, the author of the article as a psychological practitioner consulted 12 persons at the age from 15 till 20 years, who can be classified (under the criteria of Yu.V. Popov and V.D. Vid) as people with developing computer addiction (Group III).

It was found out that in case of having inadequate self estimation the student (116 persons) preferred not aggressive logical games playing sometimes aggressive games as well (we may assume that the students consciously kept back the opportunity of their affection to aggressive computer games, but judging from the communication with them, the given answers in a whole were true). 18 persons with low estimation and 10 persons with high estimation preferred to play aggressive games more. Practically all students with adequate self estimation preferred to play aggressive games. They mentioned, besides, that they play mostly for the sake of entertaining but after the game the spirits became higher. We may assume that in this situation the game helps either to form adequate self estimation or to keep it in stable state if self estimation was formed before the acquaintance with the games. But on the other hand, we cannot say more precisely what the reason is and what the result of this situation is. For more objective evaluation of obtained data one must hold additional investigations taking into account the factor of time.

Also nearly all students pointed out that other people who played aggressive games could be irritated either before the game or during it or after it, but nobody acknowledged the presence of aggression in themselves in such situation (in reality often while answering this question the person's mechanism of psychological defense snap in action, which was found out during personal conversation with respondents: they felt aggression). While holding inquiry no game addicted person was found.

The fact of non-existence of game addicted persons among this category of respondents we consider to be natural. One of the questions showed that all students who are playing computer games now started playing them being young or middle teenagers. It means that the personality had enough time to have the formed enough motivation for entering the university before finishing school. The game addiction presupposes the many hours abstractedness from other types of action. The students' motivation presupposes spending a lot of time for

studying. They play games to fill in the time, for the sake of pleasure from the game itself, for the result, for the satisfaction of the motive of achievement, possible rivalry with other players and so on. In this case the game is combined with other types of activity. A person communicates normally with other people and uses computer game during the leisure time.

To pay attention to the second group of respondents and having summarized the personal impressions and the answers of 14 people we came to such general conclusions: the constant desire to use computer is presented, but we cannot call it an addiction yet. This group of people under consideration as well has other interests not connected with virtual reality, though during their playing computer games the partial change in reality's perceiving already happens. There is narrowing of consciousness with great concentration on the giving task, inner personal insularity on the system: everything concerning this loses the interest. The feeling of heat during the finding way out in decision difficult computer task is characteristic for such people. While creating own perfectly functioning program the person feels satisfaction, the feeling of self-importance appears. At first narcissism appears and it brings more satisfaction than the presence of witnesses during the triumph. And only then, in some time, the desire to show own work to others appears. And even if the tiredness comes, the eyes hurt; the inner dependence of the desire to work with the computer, to seek for something new still remains. These people are characterized also by the passion to the computer games, but only for pleasure, for having rest. While analyzing answers of this group of respondents we found out that it was possible to divide them into two subgroups. For the people of the first subgroup computer games help to decrease the level of aggressiveness hidden in deeps of unconsciousness; with the people of the second subgroup a definite social barrier which blocks the asocial behaviour is destroyed. It is difficult to say what factors influence the appearance of such reactions. To do this we need additional aimed researches.

Besides the respondents remarked that after the game finished the emotions appeared during the game (including aggression) remained with them in real world from five minutes till two hours.

People playing computer games regularly every day for several hours were referred to the third group. The psychological corrective work was conducted with them. It was found out that in most cases these are people with pronounced non-adequate self estimation, with distinctive restraint which hid inner spite and the desire to isolate oneself from outer world. It is the result of non-forming and inefficiency of ways of person's psychological defense which could give him an opportunity to reduce the emotional tense even for some time; of the person's inability to overcome productively the situation of complicated satisfaction of actual, vitally important needs; of the deformity of the system of values; of the inclination to react inadequately to the frustrating circumstances; of the existence of psycho-traumatic situation the way out of which cannot be find; inability to perceive the situations connected with the necessity to overcome the life difficulties adequately, with the building the relations with the people around and regulation of own behaviour. Thus the personality appears to be helpless before the negative states overwhelming it and it resorts to changing of its state with the help of virtual reality. Very often it helps hesitated and timid person to release from the fear and uncertainty. But on the other hand in the process of developing game addiction the psychological state of many people starts changing: the inner conflicts intensify; the week psychological adaptation becomes more and more vivid; the release of adequate perception of emotions through mimicry, gestures and poses of another person is noticed including expression of own once; the increased level of emotional expression is registered; the leveling of sex differentiations in emotional sphere between boys and girls appears.

Besides, in Group III of respondents in 10 cases out of 14 the unconscious suicide thoughts were registered. It is necessary to notice that two persons refused from psychological corrective work in spite of parents' demands to continue the communication with the psychologist (young men of 23 and 24 years old).

The changes of the needs' character under the condition of game addiction

We as many other researchers witnessed the changes of needs of people having game addiction.

Taking into account the classical pyramid (hierarchy) of needs by A. Maslow, we should notice that the needs under the situation of game addiction have a bit different character according to the point of view of specialists.

Below we present a table of needs on non-playing people (non gamers) and gamers

Non gamers	Gamers
Need in self actualization	
Usage of own potential opportunities for personal growth, self development, career and professional growth, implementation of own wishes and fantasies in reality.	The satisfaction of this need adds up to achieving new records in computer games, obtaining sharper feeling with the help of game, realization of own fantasies and desires with the help of computer games.
Aesthetic needs	
Acceptance of rules and norms of the society, acting according to aesthetic norms.	There is no prohibitions and limitations accepted in society in computer games, that is why gamer feels freely and comfortably in virtual reality (code of ethic norms is defined by the game itself) having an opportunity to perform any deeds with impunity (pogroms, explosions, killing of "aliens") and to the contrary feels disgust for real world probably understanding own defectiveness.
Cognitive needs	
Everything mysterious, unknown, unexplained attracts, as the result the desire to receive knowledge and skills for becoming acquainted with it appears.	The desire to know something new refers only towards new cool computer games. Anything outside the virtual reality interests not much or does not interest at all.
Need to be respected	
The intention to achieve respect of people, to rise own prestige in society, to receive professional and personal recognition, not to spoil own reputation by any deed.	The intention to rise own prestige in the circle of gamers, the desire to become the best player, to become a leader among them, a "legend of gamers' circle".
Need in affiliation and love	
A person needs the emotional relations with people, in taking its own place in its group, comes to this goal intensively.	A person tries to avoid extra social contacts, preserving them only in the circle of gamers. Love is understood through the prism of virtual reality in which he/she is a "superhero", all-powerful wizard, unconquerable warrior etc.
Need in security	
A person seeks for the order and stability in real vital situations. Facing the threat he/she tries to avoid it or to obviate it.	A person is not confident in himself concerning real world which is perceived as strange, full of traps and unexpectedness, that is why he/she tries to stay longer in virtual reality in which he/she feels self-confidently and quietly.
Physiological (organic) needs	
The satisfaction of these needs in natural, normal way.	Often non adequate replace of such need as sexual satisfaction by virtual sex, games sex – modulators. Being involved into the game a person often does not feel hungry or thirsty.

Main mechanisms of developing game addiction are based upon the needs of a person in taking the role and going away from reality. A person who adapts normally in society does not try to avoid from reality. Only nonadapted people who estimate themselves non-adequately go away from reality taking other role on them. A feeling of being not protected is a sign of increased anxiety the level of which increases together with the increasing of contradictions between reality and virtual reality in the consciousness of game addicted person. On

the one hand the adaptation to the virtual world rises, but on the other hand the nonadaptation to real world increases. The intention to go into virtual world becomes a mean of satisfaction of a need in security and serves as a kind of protection from reality. In connection with this the changes in physiological and social needs of a person take place. Very often the game addicted person tries to "hide" with the help of the mechanism of psychological defense as far as possible the real, hidden and located in subconsciousness needs which are in conflict with ones adopted by the society. In such cases the following mechanisms function:

- exclusion – not the facts of game addiction are replaced from the consciousness but the psychologically traumatic circumstances coming with the rupture from the society. Such circumstances can be the following: impracticality of the known gamers towards the circumstances of living, accusation of them in theft of money from family members with the aim to continue the game, the refusal to communicate with surrounding people and so on.
- rationalization – the defensive mechanism with the help of which a person tries to find acceptable explanation to own extra affection to games, i.e. he/she rationalizes pathological, not realized enough addicted behaviour. The following rational explanations are the most popular: for calming oneself; my friends play; I do not drink alcohol, do not take drugs; it is easier to communicate their and so on.
- projection – a person becomes free from the feeling of fear and guilt arrogating own negative features to other people. The demonstration of rational projection is registered within the young people. The game addicted person knows that by means of game he/she shields oneself from those around one, knows that it creates the problems in communication with other people, but he/she arrogates the misunderstanding of his inner state, ungrounded petty objections, hard-heartedness, aggressiveness and other negative emotions concerning him to other people. All the more "everybody plays different games now and nothing awful happens".

Concerning the fact of lowering the level of aggressiveness during the game we may suppose, that it happens due to the mechanism of projection and transfer which helps to satisfy subconscious desire to punish offenders. To understand their influence properly we need study the process of development of game addiction within the person

The Mechanism of Forming the Psychological Addiction to Role Computer Games

To come back to the early childhood one may notice that at this age the child lacks words for expressing his inner state. And he learns to use his emotions and at the same time he studies the ways of manifestation of emotions by other people. This ability of studying emotional sphere of a person is given to a child by nature as an integral part of psychological development in early age. But not always this studying happens as positive moment. Being practically helpless a child is made an endless row of unconditional and uncompromising demands. On the one hand there is an absolute necessity (fixed genetically) to evacuate bowels when a child wants, to destroy, to express feelings and pleasant emotions connected with movements and discoveries. On the other hand there are firm demands from the people around and parents first of all for a child to refuse from these primary pleasures in order to receive the parents' approval as a reward. This reward which can disappear as quickly as it appeared is the inconceivable mystery for a child who has not learnt yet the connection between the reason and the result.

Displeased glance addressed to him can course the feelings increasing the negative data about him. The more a child receives negative information from outer world the more the uneasiness in relations with world around grows. He himself cultivates the feeling of own inferiority and as a result of it the displeased by himself appears. But revealing own displeasure through emotions a child risks to be punished by people around in case of this displeasure is not accepted by them. Coming out of the fact that every person subconsciously aims at self-defense, at adjusting to survive in different conditions (because of this in early age a child seeks first of all after receiving

parents' acknowledgment as communication with them often is the condition which guarantees such survival). Depending upon the type of relationships which are established between a child and its parents every child finds own way to attract the parents' attention or in other words creates own system of communication with surrounding. It happens quite spontaneously: a child does not know yet what is good and what is bad. Depending on his actions, results and situation definite reaction of those surrounding appears. Parents and other adults often give their evaluation to what is happening in which connection in most cases a child perceives this evaluation as the evaluation of him not of his action. That is why if some actions do not produce expected parents' reaction and do not attract their attention, but to others parents react, so a child has fixed intention to do those actions which produce the parents' reaction. Thus the formed stereotype of own relationships under the influence of communication with parents and own manner of communication become fixed.

Besides, children are very inclined to the investigation of adults' behaviour. Boryshevskiy M.I. explains it by the fact that the ability to analyze critically the adults' behaviour is not developed well enough in small children yet. Noncriticality of a child is especially vividly displayed towards the parents who are for him an irrefutable authority. He loves them and relies on them. That is why negative emotions and feelings which appear in him raise fear as they can result into the rejection of a child by his parents or by other adults who seem to be all-powerful to him at this stage of development. He/she starts pretending that there are no negative emotions in him. But as they do not disappear a child starts to project them outside, i.e. the surrounding world seems to be more cruel and awful to him than it is in reality. A child starts to feel better on the one hand but on the other it becomes worse as now he feels himself be in alien surrounding. At these stages the world is still equal for a child to his mother and father. And now instead of trying to tear those to pieces in his mind because of hunger and fury being punished by them a child starts to project these feelings on his parents; adults are imagined by a child as cruel and awful. In compliance with this, surrounding world seems to be alien and cruel as well. And all other people are included into this world. As the initial perception usually assigns to subconsciousness then he will unconsciously perceive all other people as persons who relate to him aggressively, i.e. the aggressive attitude towards them is formed inside of him (attitude to outer world which later conceals by later development referring to up-bringing). Coming out of this some outer threat always exist subconsciously for a child who is ready for defense every minute. Naturally it would be better to destroy the "alien" world but as it is impossible in reality or there is no object on which such situation can be played or such an object resists to such transferal so this is transferred into the virtual world.

It means that if the scheme of the computer game coincides with the scheme of the situation which is important and uncomfortable for the gamer's "I", it gives him pleasure to play this game. At the same time quick victory brings the feeling of dissatisfaction and pressing emptiness. That is why the gamer sometimes creates by himself barriers on the way to victory. After several difficult losses the victory is very significant. The images of those with whom the gamer constantly or accidentally comes into definite relations are also included into the game's scheme and these relationships develop according to the scheme proposed by the player. And as it known usually it is the scenario, the scheme of important for player relationships. That is, a kind of psychotherapy takes place which leads to the decrease of inner tension. But if the frustration is high enough then the aggressiveness shown during the game continues to be present for time from 45 minutes till two (and more) hours and may develop in the relationships with those people around.

At the same time the most difficult thing is to take the first step. That is why if the intention to destructive actions presents within the person subconsciously (for example, to beat certain person) but it cannot be realized under existing circumstances (it can be very close person or a person who will not allow to do it) so designing similar situation a person takes the first step to some extent: destructive behaviour in virtual world. With the help of such actions the control over subconscious desires within such person is reduced. When the first step is taken and a

person obtains definite moral satisfaction out of done, a kind of psychological relaxation takes place. But this relaxation is temporary and unstable as the source of tension exists in reality. If a person is inclined to satisfaction of desires and the defensive mechanisms restraining it within generally accepted behaviour are weak enough, the dream may come true. More than that during the game subconscious tension of muscles and accumulation of potential energy take place which demands the release. So to avoid the uncontrolled reaction one must create the conditions under which this release would take place under the control of conscious, otherwise a great possibility of development of conflict situations exists.

Why is such relaxation possible in this situation? As it is known in the person's brain there are many different centers responsible for this or that functions leading to the satisfaction of our demands and desires. When one of them is stimulated by non-realized definite demand or desire a person feels discomfort. When combined unsatisfied desires influence upon a person, he/she feels suffering, i.e. corresponding brain centers will be in the state of stimulation for a long time providing orientation of person's behaviour to the achieving of goal. The stronger such stimulation is the greater effort the person takes to realize his orientation.

The intensive searches of means of releasing this tension start. The means socially excepted together with asocial ones can be included here. During last time computer has started to be included here more often. It helps to release tension within definite type of a person. During this the "brain center" of interaction with the computer is forming. If you work on it, the center of stimulation inhibits and the satisfaction comes. Thus, a person fulfills the "will" of this "center". Every time the computer's interaction dependence of a person increases. The addiction to the work with the computer, a kind of computer mania starts developing step by step.

But in contrast to alcoholism and drug addiction there are more chances to release person from such an addiction. To do this one needs first of all analyze scrupulously the set of games, the game's process and psychological state of a person before and after the game.

We took an attempt to use computer games with psychotherapeutic aim and to trace their influence upon the psyche of a person who is in the state of anxiety, in posttraumatic situation. In the capacity of subsidiary material we used the game "Quake II" as a quite dynamic and easy learning game. At the initial stage we registered the feeling of sickness, slight dizziness and at the same time the feeling of competition. If a person conceals quite strongly his aggressiveness (he feels sorry to kill people even during the game, i.e. the mechanisms of psychological defense are quite strong), then the "revelation" of aggressiveness inside him may frustrate greatly.

One of the reasons of it is the resistance of the mechanisms of psychological defense, which are aimed to suppress subconsciously or not to allow into the consciousness the information which contradicts the demands of conscience, some definite moral qualities of a person and can traumatize him or provoke awareness of something the person does not want to know, which can be a kind of trauma as well.

When a person sits down the game which coincides in its scheme with the person's demands the process of "involving into the murder" starts. At the beginning "murder" of a rival happens under the influence of a factor "if you do not kill they will kill you". Then the awareness of the fact that you are the same as others takes place, i.e. you do not have a sense of pity towards "killed" but you have only heat and desire to "survive". In this case there is a great opportunity of development of headaches as subconscious aggressiveness coming out destroys to some extent the "I-Image". But on the other hand, displaced earlier emotion "attracted" to itself energetic potential of a person for its blocking, i.e. a person under the circumstances of up-bringing subconsciously took a lot of efforts for rejection of such attractions, which led automatically to the constriction of consciousness. It led to inadequate estimation of events happening around a person and in its turn led the inadequate reaction to those events. Actually a person reacted more to its own emotions towards the surrounding world than to the events which happened in it.

After several séances of successful game using the method of introspection and test for finding out the level of self-estimation, the lowering of anxiety was fixed, self-estimation increased, it became easier for the person to say "no" in the situations which do not satisfy him and so on. We consider that the developing of new programs and their influence upon the person depends to much extent upon moral qualities of programmers.

Taking into account everything mentioned we pose before the psychologists and developers of computer programs especially computer games the tasks connected with the increasing of psychological culture of interaction of computer users and working out psychological and psycho-preventive strategies and forms of work.

Under the condition of having strong computer addiction, it is possible to overcome it only when the stronger motivation of receiving pleasure from something else than the pleasure from the game is found, or when the strengthening of "I-Image" of a person or the influence of both factors simultaneously take place

Conclusion

We cannot speak about complete studying of mechanisms of addiction's formation. The problem of the influence of the computer upon the person is very extensive and many-sided. We analyzed three main psychological mechanisms of role computer game addiction's formation: the necessity of going away from the reality, the exception of the role of other and implementation of the desire to take revenge on oneself for previous pain (in our case for parents) in virtual reality. As practice shows, the awareness and realization of such desire very often (but not always unfortunately) brings to weakening (in some cases to disappearance) of subconscious addiction to desires that is to computer games. These mechanisms always work simultaneously, but one of them can prevail over the other in its strength of influence upon addiction's formation. They are based on the process of compensation of negative vital experience, so we can presume that they will not work if a person is completely satisfied with his/her life, does not have psychological problems and consider his life to be happy and productive. But there are not so many of such persons, that is why the majority of people can be considered to be potentially predisposed to the formation of psychological addiction to role computer games. There are some assumptions about the reasons of strengthening of aggressive manifestations in some cases, about the decreasing of them in other cases and staying without changes in third cases. But factors influenced upon these processes are still not studied enough and demand additional serious investigations

Bibliography

- [Вільямс, Маклін, 1988] Вільямс Р., Маклін К. Комп'ютери в школі. К., 1988.
 [Воронов, 1990] Воронов Ю.П. Компьютеризация: шаг в будущее. Новосибирск, 1990.
 [Макото, 1988] Макото Арисава. Что такое компьютер? К., 1988.
 [Нестеренко, 1990] Нестеренко А.В. ЭВМ и профессия программиста. М., 1990.
 [Ракитов, 1991] Ракитов А.И. Философия компьютерной революции. М., 1991.
 [Рошак, 1999] Рошак Н.М. Англійська мова з комп'ютером // Комп'ютер у школі та сім'ї. №2, 1999.
http://www.makarova.net/index.php?option=com_content&task=view&id=110
<http://www.narcom.ru/publ/info/249>

Authors' Information



Helen Shynkarenko – Kirovograd State Pedagogical University named by V. Vynnychenko, candidate of philological science, associate professor, Kirovograd, Ukraine, e-mail: oshink@kspu.kr.ua



Viktoriya Tretiyachenko – East Ukrainian National University by Volodymyr Dahl Doctor of Psychological Sciences, Professor, Lugansk, Ukraine e-mail: psihology-snu@mail.ru

EDUKIT: INFO-EDUCATIONAL PLATFORM ENABLING TO CREATE WEBSITES FOR SECONDARY SCHOOLS

O.Y. Stepanovsky, S.O. Ryzhikova, D.P. Nechiporenko, B.S. Elkin, O.B. Elkin, I.V. Garyachevska, O.M. Dyachenko, D.V. Fastova

Abstract: *This article focuses on computerisation and informatisation of Ukrainian secondary schools. Some of the problems can be solved with the help of modern technologies, for instance, by presenting information on websites. Website development for government-financed educational institutions presents a number of difficulties which often cause schools to give up the idea of creating their own official websites. This article presents EDUKIT, an open social info-educational platform, which enables national schools to create websites independently and at no cost and further develop them into a single school network as well as build an internal local school system.*

Problems and prospects of school website development in Ukraine. Problem statement

The paper is published with financial support by the project ITHEA XXI of the Institute of Information Theories and Applications FOI ITHEA (www.ithea.org) and the Association of Developers and Users of Intelligent Systems ADUIS Ukraine (www.aduis.com.ua).

Computers and information technologies have penetrated all areas of people's life, not to mention such socially significant sector as education.

To ensure a new type of interaction between the society and the system of education, we need to prepare, publish and spread information on the performance of the whole education system in general and each educational institution in particular. Recent years have shown that people are becoming aware of the growing need for more information. This demand is satisfied by means of periodicals and one-off publications, TV news stories, reports and public speeches describing the state and development of educational establishments. Apart from that, you can also draw information from modern technology sources, in particular websites.

The domestic and foreign experience of building websites for educational institutions shows that school website development presents a number of technical, technological, financial, organisational and managerial problems [Kravchina,2008], [Eelmaa, 2005] which, in most cases, cause developers to give up this idea. However, the new laws and the plan of school informatization set deadlines for creating electronic school representations in the Internet (e.g. as early as the end of 2009 in Kharkiv region [Pinchuk, 2008]). The study of the Ukrainian IT market (School Site – www.edusite.ru, uCoz Web-Services – www.ucoz.ru, Net School Ukraine – net.elnik.kiev.ua and others) has revealed that there is no simple, barrier-free, user-friendly, goal-oriented, yet gratuitous, solution of the above-mentioned task.

The EDUKIT platform as an info-educational system

EDUKIT is a non-commercial social initiative presented by the Stella Systems Company and the Eastern-Ukrainian Branch of the International Solomon University and sponsored by the Kharkiv Board of Education. The main aim of the project is to promote the introduction of information technologies into the Ukrainian educational system by creating official websites for secondary schools in Kharkov and Kharkiv region.

The project is based on the materials developed by Dialog WebDesign GmbH (Frankfort on the Main, Germany), whose official representative in Ukraine is Kharkiv IT-company Stella Systems.

The EDUKIT Project is an info-educational platform whose goal is to help schools create websites, further develop them into a single school network and, if necessary, set up an internal local school system. This business solution facilitates the information and experience exchange both at the level of a separate educational institution and at the level of the secondary education system as a whole.

School websites created on the EDUKIT platform are not just a set of autonomous information units, but an integral info-educational system in which all school websites are elements of a hierarchic structure equipped with process control and unification. It has to be pointed out that, apart from developing the platform, we are also using different ways of product distribution to deliver it to the end user as well as collecting client feedback. In other words, there is a close interaction between the manufacturer and the consumer.

Technical aspects of the platform and its performance capabilities

From the technical point of view, EDUKIT is a constantly developing Internet platform which enables building an unlimited number of websites united by a single structure and functions (see Pic. 1). All platform websites are organized into an integral system which centralises the storage of important information and provides easy access to it with the help of various resources. For example, after signing in at a school website, a teacher can look through the teaching aids of a colleague who has placed them there through a different website of the system.

This problem solution is based on the multifunctional content management system (CMS) called Pulsar developed by a group of authors with the help of free up-to-date technologies. Pulsar doesn't require any special technical knowledge or programming skills. This system enables you to add, edit or delete any website content, without resorting to anyone's assistance. All the changes you make to a school website created with the help of the system are carried out on a real-time basis. CMS Pulsar has an interface designed for numerous users which allows simultaneous editing of different website sections by several members of school staff whose access rights are controlled by the system.

The website editing system is distributed as an Open Source, which frees its users from any further licence obligations or material liabilities.

The platform is developed with the help of top-notch foreign technologies such as Zend-Framework (www.zendframework.com), WYMeditor (www.wymeditor.org), jQuery (jquery.com).

The main modules of the platform are [Stepanovski, 2009]:

- News and events. You can publish news and/or events on a website and always keep them up-to-date. The home page displays only the news and events selected by the administrator. This kind of information always attracts the attention of website visitors and is indicative of the website's popularity.
- Service Order Form. Allows visitors to view the range of services on offer (for example, school clubs), select services they are interested in and order them on-line. After the user has applied, the information stated in the application form is sent to the administrator's address (or to the address of any other person in charge).
- Calendar of Events. The calendar will help organise and schedule events and activities for any dates. You can also divide the events into categories and highlight them with the help of editing tools.
- Opinion Polls. The purpose of this module is to conduct surveys, votes and polls of website visitors. This tool allows you to ask questions with the option of one or several answer choices, set time frames for polls and calculate statistical data.

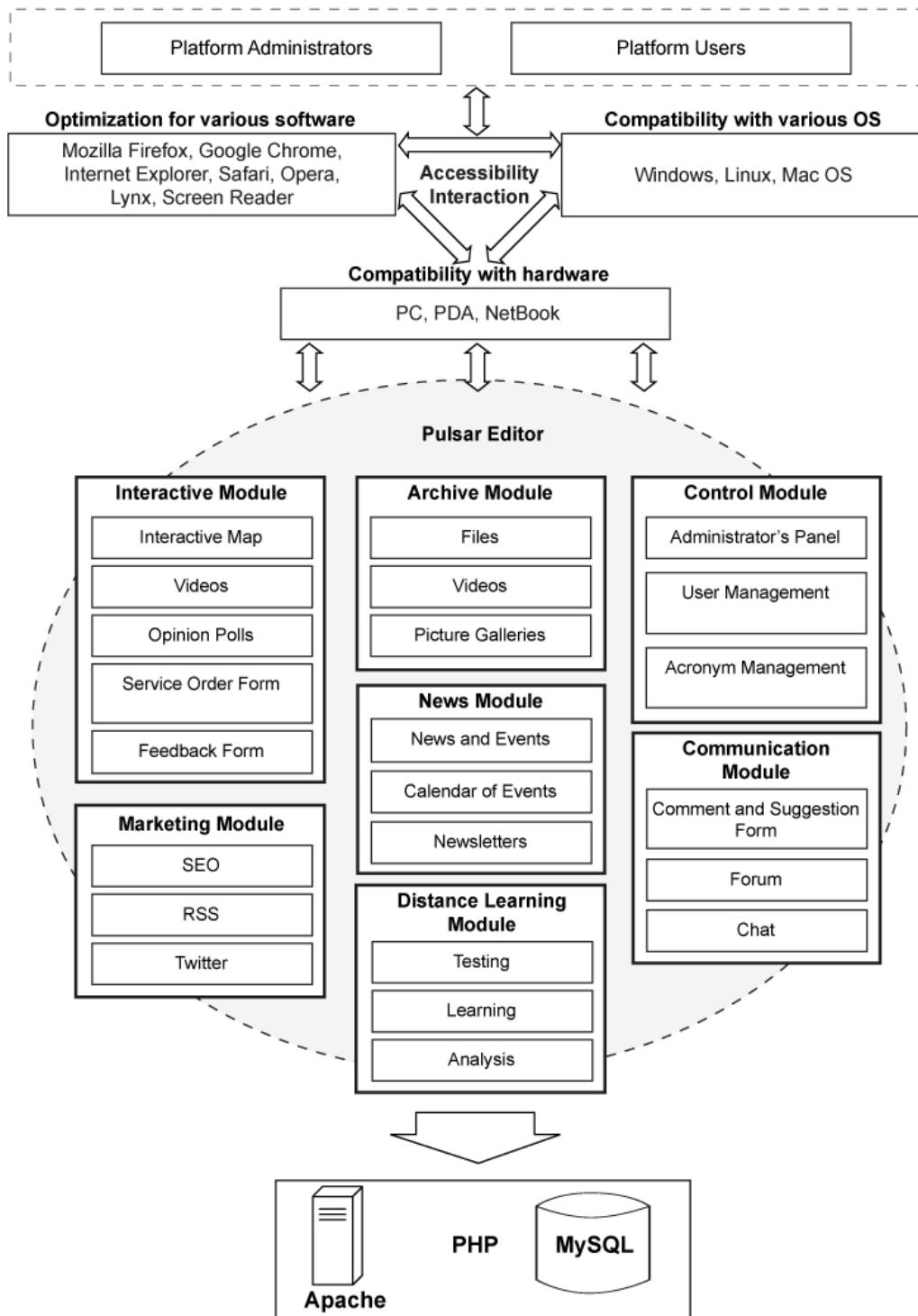


Figure. 1. EDUKIT Info-Educational System

- **Feedback Form.** Allows users to get feedback from the representatives of a school by filling in the form with their contact information. The completed form is sent to the indicated school address by e-mail.

- **Comment and Suggestion Form.** Allows users to leave comments and suggestions concerning the school performance on the website, which helps further optimise business procedures and satisfy the needs of the target audience. With the help of the editing system, you can add this form to any website page. To avoid obscene comments, we have integrated a function of moderation and a swear word filter of unprintable words and expressions. The comments are published on the page only after being confirmed by the administrator.
- **Interactive Map.** The Interactive Map with an itinerary marked on it will help users quickly find the way to the school they want to get to and work out how long it will take them to get there. The integration of Google Maps enables users to work their way to their point of destination.
- **Newsletter.** This module enables website visitors to subscribe for regular newsletter delivery and manage a great number of subscribers. The newsletter delivery template is adapted to the corporate design. Users subscribe for newsletter delivery (or unsubscribe) on their own, without school staff's interference.
- **Search Function.** The function of search within the website is necessary for Internet resources which contain great amounts of information as it helps visitors find their bearings on the website. The function of simple search is usually located on top of the website page. If you are not satisfied with the search results, you can use advanced search, which allows you to specify your request.
- **Video Gallery.** Enables you to upload on the website an unlimited number of videos in the *.flv format. To play a video, you can use the player integrated into the website which has a function of full-screen view.
- **Picture Galleries.** The system allows you to create and edit Picture Galleries at any section of the site without anybody's assistance. You can easily move pictures from one collection to another, change their captions or delete them from the website. The size of the pictures is automatically adapted to fit the screen. Full-screen view is also available.
- **File Archive.** The file archive can store documents in *.doc and *.pdf formats as well as photo-, audio- and video-materials which users can download and store on their computer. The module enables a school to keep all their files in one place as well as systematise and group important materials.

Websites created with the help of EDUKIT meet the international website quality standards of World Wide Web Consortium (www.w3.org) and are barrier-free. In other words, this means that school websites are user-friendly, and their main functions work on monitors with any screen resolution, including mobile devices. The content of such websites is accessible to screen readers, special programs which reproduce the screen content either in the form of speech or with the help of Braille display. Thus, new school websites embody one of the main concepts of education – accessibility to the maximum number of people, regardless of their technical and physical abilities.

Platform approbation and distribution

The experiment of software product distribution began in February 2009, when we gathered the first group of 18 schools and 3 computer technology laboratories of the Kharkiv Board of Education.

To facilitate testing and ensure its continuity, we created a technical support website www.edu.kh.ua (see Figure.2) which contains not only information about the project, news and contact information, but also a **full user manual** [Stepanovski, 2009] **and video lessons on how to use the platform**. We have created a forum for EDUKIT users, where anyone can anonymously get specialists' consultations on any questions they are

interested in and share experience with other project participants. In addition, there is a blog of the project, social network groups and a telephone hotline available five days a week.

Figure. 2. EDUKIT technical support website (www.edu.kh.ua)

Thanks to the efficient feedback from the target group representatives, we have been able to carefully study their needs and wants. As a result, the platform is developing in the right direction – taking into account the particularities of secondary schools.

The testing period among the first project participants and the exchange of opinions gave EDUKIT impetus to further development. At the request of the public, we added new features to the platform, including small graphic changes to the website design, language selection according to the school's language of instruction, integration of any number of additional language versions, etc.

To help new project participants use the platform with the highest efficiency and at the same time to receive first-hand recommendations and comments, EDUKIT managers conducted dozens of seminars-presentations in Kharkiv and Kharkiv region. They gave practical training sessions and explained the facilities of the platform. Moreover, school representatives had an opportunity to get individual consultations and communicate with the developers.

As of April 2010, as many as 219 Kharkiv schools and 37 Kharkiv region schools have joined the project, which constitutes 97% and 4% of the city's and region's schools accordingly. Most schools decided to use the product because of:

- its simplicity and user-friendliness;
- the common principle of editing for all websites;
- the efficient technical support;
- the introduction of a multi-module structure – a true enhancement to the platform;
- the product accessibility, regardless of users' material and technical abilities.

At the end of the first stage of EDUKIT introduction, in October–November 2009, we held the Contest for the Best School Website created using as many EDUKIT facilities as possible (see the sample of a website winner – Figure 3).

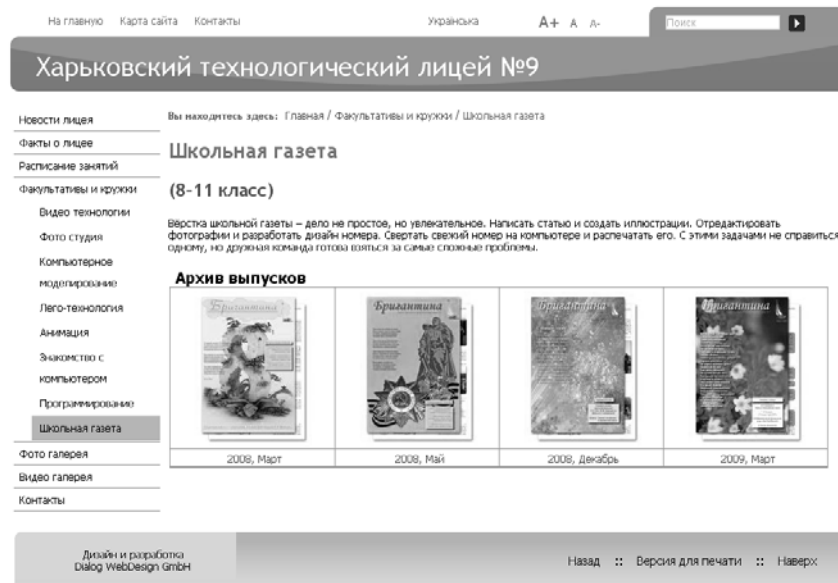


Figure 3. The website of Technological Lyceum # 9 created on the EDUKIT platform (www.lyceum9.edu.kh.ua)

Conclusion

At present, the EDUKIT Project is steadily developing: we are continuing to collect client feedback as well as working on the technical enhancement of the product, producing new versions and making presentations at topical conferences [NevaCamp, 2009]. We are also planning to introduce the platform in other regions of Ukraine and other countries of the CIS.

The experience of November 2009, when school websites succeeded in distributing homework assignments among students at the time of the lengthy quarantine due to the flu epidemic, shows that the platform could be further used in distance learning. We are already making the first steps in this direction – trying out the developed module of computer-aided students' knowledge testing, which is expected to be in demand owing to the introduction of compulsory external independent testing in secondary schools.

Bibliography:

- [Kravchina,2008] Кравчина О. Є. Інформатизація організаційно-управлінської діяльності в загальноосвітній школі / О. Є. Кравчина // Інформаційні технології і засоби навчання [Електронний ресурс]. – К.: Інститут інформаційних технологій і засобів навчання АПН України, 2008. – № 3 (7). – Режим доступу до журналу <http://www.nbu.gov.ua/e-journals/ITZN/em7/emg.html>.
- [Eelmaa, 2005] Ээльмаа Ю. В. Школьный сайт – это просто? / Ю. В. Ээльмаа // Проблемы автоматизации управления образованием. – Москва, 2005. – № 1.
- [Pinchuk, 2008] Пинчук А. В. Школьные сайты: проблемы и перспективы / А. В. Пинчук // Информационные и коммуникационные технологии в образовании и научной деятельности, 21-23 мая 2008 г. – Хабаровск, 2008. – С. 206–208.

- [Matkin, 2008] Маткин А. А. Сайт школы – не роскошь, а средство продвижения / А. А. Маткин [Электронный ресурс] // Справочник классного руководителя. – 2008. – № 5. – Режим доступа к журналу <http://klass.resobr.ru/archive/year/articles/565/>.
- [Stepanovski, 2009] Степановский А. Ю. Социальный образовательный проект «EDUKIT»: разработка сайтов для средних учебных заведений / А. Ю. Степановский, С. О. Рыжикова // Одиннадцатая научная конференция ВУФ МСУ «История. Компьютерные науки. Экономика»: Тез. докл. – Х.: Тарбут Лаам, 2009. – С. 36–41.
- [NevaCamp, 2009] Материалы международной конференции в области веб-технологий, стартапов, инвестиций «NevaCamp», 26-28 июня 2009 г. [Электронный ресурс] – Санкт-Петербург, 2009 – Режим доступа к материалам конференции <http://nevacamp.com/ru/nevacamp09>.
- [Stepanovski, 2009] Руководство пользователя по работе с редакционной системой платформы «EDUKIT» (www.edu.kh.ua) / [А. Ю. Степановский, С. О. Рыжикова, Д. П. Нечипоренко и др.]. – Х.: ВУФ МСУ, 2009. – 66 с.

Authors' Information

- Oleksiy Y. Stepanovskiy** – LLC «Stella Systems», director. e-mail: elkin3son@mail.ru
- Svitlana O. Ryzhikova** – LLC «Stella Systems», social programs coordinator. e-mail: elkin3son@mail.ru
- Dmitriy P. Nechiporenko** – LLC «Stella Systems», head of innovations department. e-mail: elkin3son@mail.ru
- Borys S. Elkin** – Eastern-Ukrainian Branch of the International Solomon university, director, professor. e-mail: elkin3son@mail.ru
- Oleksandr B. Elkin** – Eastern-Ukrainian Branch of the International Solomon university, deputy director, associate professor. e-mail: elkin3son@mail.ru
- Iryna V. Garyachevska** – Eastern-Ukrainian Branch of the International Solomon university, chair of software engineering, associate professor. e-mail: elkin3son@mail.ru
- Olesya M. Dyachenko** – Eastern-Ukrainian Branch of the International Solomon university, chair of software engineering, associate professor. e-mail: dolesya@mail.ru
- Darya V. Fastova** – Eastern-Ukrainian Branch of the International Solomon university, chair of software engineering, associate professor. e-mail: elkin3son@mail.ru