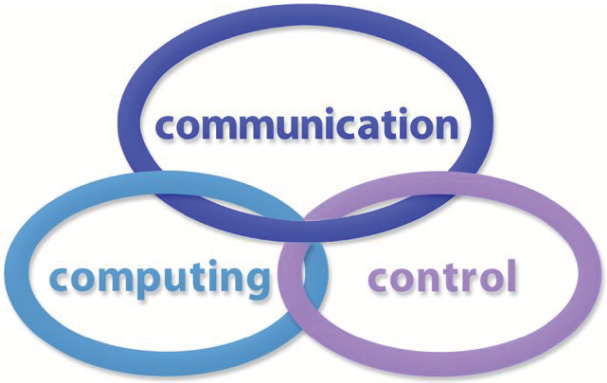
INTERNATIONAL JOURNAL

of

COMPUTERS, COMMUNICATIONS & CONTROL

With Emphasis on the Integration of Three Technologies

IJCCC



Year: 2010 Volume: 5 Number: 5 (December)

Agora University Editing House

CCC Publications

www.journal.univagora.ro

International Journal of Computers, Communications & Control

EDITORIAL ADDRESS:
R&D Agora Ltd. / S.C. Cercetare Dezvoltare Agora S.R.L.
Piaţa Tineretului 8, Oradea, jud. Bihor, Romania, Zip Code 410526
Tel./ Fax: +40 359101032
E-mail: ijccc@univagora.ro, rd.agora@univagora.ro, ccc.journal@gmail.com
Journal website: www.journal.univagora.ro

DATA FOR SUBSCRIBERS
Supplier: Cercetare Dezvoltare Agora Srl (Research & Development Agora Ltd.)
Fiscal code: RO24747462
Headquarter: Oradea, Piata Tineretului Nr.8, Bihor, Romania, Zip code 410526
Bank: MILLENNIUM BANK, Bank address: Piata Unirii, str. Primariei, 2, Oradea, Romania
IBAN Account for EURO: RO73MILB0000000000932235
SWIFT CODE (eq.BIC): MILBROBU

International Journal of Computers, Communications & Control

EDITORIAL BOARD

Mario de J. Pérez Jiménez
Dept. of CS and Artificial Intelligence
University of Seville
Sevilla, Avda. Reina Mercedes s/n, 41012, Spain
marper@us.es

Dana Petcu
Computer Science Department
Western University of Timisoara
V.Parvan 4, 300223 Timisoara, Romania
petcu@info.uvt.ro

Radu Popescu-Zeletin
Fraunhofer Institute for Open
Communication Systems
Technical University Berlin, Germany
rpz@cs.tu-berlin.de

Imre J. Rudas
Institute of Intelligent Engineering Systems
Budapest Tech
Budapest, Bécsi út 96/B, H-1034, Hungary
rudas@bmf.hu

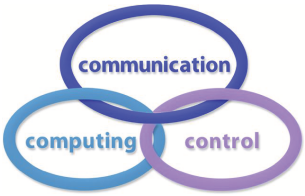
Athanasios D. Styliadis
Alexander Institute of Technology
Agiou Panteleimona 24, 551 33
Thessaloniki, Greece
styl@it.teithe.gr

Gheorghe Tecuci
Learning Agents Center
George Mason University, USA
University Drive 4440, Fairfax VA 22030-4444
tecuci@gmu.edu

Horia-Nicolai Teodorescu
Faculty of Electronics and Telecommunications
Technical University "Gh. Asachi" Iasi
Iasi, Bd. Carol I 11, 700506, Romania
hteodor@etc.tuiasi.ro

Dan Tufiş
Research Institute for Artificial Intelligence
of the Romanian Academy
Bucharest, "13 Septembrie" 13, 050711, Romania
tufis@racai.ro

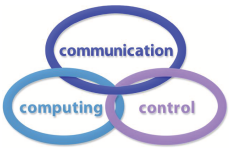
Lotfi A. Zadeh
Department of Computer Science and Engineering
University of California
Berkeley, CA 94720-1776, U.S.A.
zadeh@cs.berkeley.edu



International Journal of Computers, Communications & Control

Short Description of IJCCC

Title of journal: International Journal of Computers, Communications & Control
Acronym: IJCCC
International Standard Serial Number: ISSN 1841-9836, E-ISSN 1841-9844
Publisher: CCC Publications - Agora University
Starting year of IJCCC: 2006
Founders of IJCCC: Ioan Dzitac, Florin Gheorghe Filip and Mişu-Jan Manolescu
Logo:



Number of issues/year: IJCCC has 4 issues/odd year (March, June, September, December) and 5 issues/even year (March, September, June, November, December). Every even year IJCCC will publish a supplementary issue with selected papers from the International Conference on Computers, Communications and Control.
Coverage:

- Beginning with Vol. 1 (2006), Supplementary issue: S, IJCCC is covered by Thomson Reuters - SCI Expanded and is indexed in ISI Web of Science.
- Journal Citation Reports/Science Edition 2009:
  - Impact factor = 0.373
  - Immediacy index = 0.205
- Beginning with Vol. 2 (2007), No.1, IJCCC is covered in EBSCO.
- Beginning with Vol. 3 (2008), No.1, IJCCC, is covered in SCOPUS.

Scope: IJCCC is directed to the international communities of scientific researchers in universities, research units and industry. IJCCC publishes original and recent scientific contributions in the following fields: Computing & Computational Mathematics; Information Technology & Communications; Computer-based Control.
Unique features distinguishing IJCCC: To differentiate from other similar journals, the editorial policy of IJCCC encourages especially the publishing of scientific papers that focus on the convergence of the 3 "C" (Computing, Communication, Control).
Policy: The articles submitted to IJCCC must be original and previously unpublished in other journals. The submissions will be revised independently by at least two reviewers and will be published only after completion of the editorial workflow.

Contents

Int. J. of Computers, Communications & Control, ISSN 1841-9836, E-ISSN 1841-9844
Vol. V (2010), No. 5, pp. 613-615

Writing as a Form of Freedom and Happiness
Celebrating the 60th birthday of Gheorghe Păun

G. Ciobanu

Gabriel Ciobanu
Romanian Academy, Institute of Computer Science
and A.I.Cuza University of Iasi, Romania
E-mail: gabriel@info.uaic.ro

Gheorghe PĂUN (born on December 6, 1950) graduated the Faculty of Mathematics of the Bucharest University in 1974 and got his PhD at the same faculty in 1977. He has won many scholarships, in Germany, Finland, The Netherlands, Spain, etc. Presently he is a senior researcher at the Institute of Mathematics of the Romanian Academy, Bucharest, and a Ramon y Cajal research professor at Sevilla University, Spain. Since 1997 he is a Corresponding Member of the Romanian Academy, and since 2006 a member of Academia Europaea. His main research fields are formal language theory (regulated rewriting, contextual grammars, grammar systems), automata theory, combinatorics on words, computational linguistics, DNA computing, membrane computing (this last area was initiated by him in 1998). He has (co)authored and (co)edited more than fifty books in these areas, and he has (co)authored more than 400 research papers. In the last two decades he has visited many universities from Europe, USA, Canada, Japan, also participating to many international conferences, several times as an invited speaker. He is a member of the editorial board of numerous computer science journals and professional associations.

Figure 1: A copy cropped from [1]

Essentially writing is form of thinking on paper, and a way of learning. According to Winston Churchill, writing a book is an adventure. "To begin with, it is a toy and an amusement; then it becomes a mistress, and then it becomes a master, and then a tyrant. The last phase is that just as you are about to be reconciled to your servitude, you kill the monster, and fling him out to the public." On the other hand, writing could be a form of freedom by escaping the madness of a period, and reducing the anxiety. In many situations the authors write to save themselves, to survive as individuals.

Gheorghe Păun is an example of a person affirming his own existence by writing. He is a prolific writer with a huge number of papers: tens of scientific books, hundreds of articles, several novels, poems, and books on games. A list of his scientific publications is posted at `http://www.imar.ro/~gpaun/papers.php` [2], while his books are listed at `http://www.imar.ro/~gpaun/books.php` [1] His way of distributing information is not by speaking, but by writing. Gheorghe Păun did not like very much to teach in universities. He preferred a form of "teaching by researching", combining ideas with nice metaphors and distributing his knowledge in articles and books. In this way he wrote several papers having a high impact in the scientific community. His seminal paper "Computing with membranes" published in Journal of Computers and System Sciences in 2000 and his fundamental book on computation theory "Membrane Computing" (Springer, 2003) has over 1,000 citations [6] (and his author was recognized as an "ISI highly cited researcher" [5]). He has defined new branches, new theories. The field of membrane computing was initiated by Gheorghe Păun as a branch of natural computing [3]; P systems are

inspired by the hierarchical membrane structure of eukaryotic cells [4]. An impressive handbook of membrane computing was published recently (2010) by Oxford University Press.

After 1990 he becomes a traveling scientist, visiting several countries and receiving many research fellowships and awards. Fruitful scientific collaboration at Magdeburg University (Germany), and at University of Turku (Finland). The trio Gheorghe Păun, Grzegorz Rozenberg and Arto Salomaa is well-known for several successful books. The last years were spent in Spain, first in Tarragona and now in Sevilla. Several collaborations were possible during his trips, and there are over 100 co-authors from many countries. His scientific reputation is related to the large number of invited talks provided at many international conferences and universities. He is a member of the editorial boards for several international journals, corresponding member of the Romanian Academy (from 1997), and member of Academia Europaea (from 2006).

It is not possible to understand the personality of Gheorghe Păun without mentioning his activity as writer of novels and poems; he is a member of the Romanian Writers Association for a long time. Another aspect of his life is related to the intellectual seduction of games; he was the promoter of GO in Romania, writing many books about GO and other "mathematical" games.

Personally, I am impressed by the speed of his mind (it is enough to say few words about some new results, and he is able to complete quickly the whole approach), his wide-ranging curiosity and intelligence, rich imagination and humor, talent and passion. He is highly motivated by challenging projects, and work hard to conclude them successfully. There are very few scientists having such an interesting profile, and I am very happy to learn a lot from him.

Celebrating his 60th birthday, we wish him a good health, long life, and new interesting achievements!

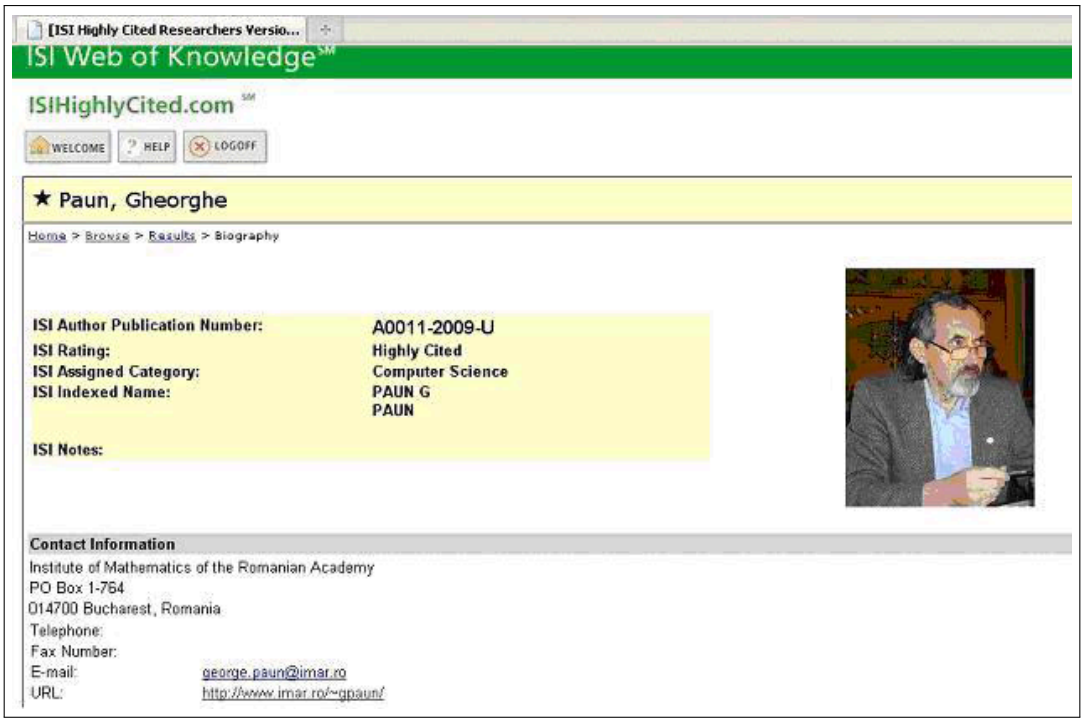


Figure 2: G. Păun-An ISI Highly Cited Researcher (A copy cropped from [5])

Bibliography

[1] Gheorghe Păun, *One More Universality Result for P Systems with Objects on Membranes*, International Journal of Computers, Communications & Control, 1(1): 25-32, 2006 (Free access at http://www.journal.univagora.ro/download/pdf/21.pdf)

[2] http://www.imar.ro/ gpaun/

[3] http://esi-topics.com/

[4] http://ppage.psystems.eu/

[5] http://hcr3.isiknowledge.com/

[6] http://interaction.lille.inria.fr/ roussel/projects/scholarindex/index.cgi

Int. J. of Computers, Communications & Control, ISSN 1841-9836, E-ISSN 1841-9844
Vol. V (2010), No. 5, pp. 616-624

Energy-Efficient Algorithms for k-Barrier Coverage In Mobile Sensor Networks

D. Ban, W. Yang, J. Jiang, J. Wen, W. Dou

Dongsong Ban, Wei Yang, Jie Jiang, Jun Wen, Wenhua Dou
National University of Defense Technology
School of Computer
Changsha, Hunan, P.R.China
E-mail: {dsban,weiyang,jiejiang,junwen,whdou}@nudt.edu.cn

Abstract: Barrier coverage is an appropriate coverage model for intrusion detection by constructing sensor barriers in wireless sensor networks. In this paper, we focus on the problem how to relocate mobile sensors to construct k sensor barriers with minimum energy consumption. We first analyze this problem, give its Integer Linear Programming(ILP) model and prove it to be NP-hard. Then we devise an approximation algorithm AHGB to construct one sensor barrier energy-efficiently, simulations show that the solution of AHGB is close to the optimal solution. Based on AHGB, a Divide-and-Conquer algorithm is proposed to achieve k-barrier coverage for large sensor networks. Simulations demonstrate the effectiveness of the Divide-and-Conquer algorithm.
Keywords: k-barrier coverage, energy-efficient relocation, mobile sensor network

1 Introduction

Monitoring and surveillance are very important applications of wireless sensor networks, such as detecting the intruders when they cross international borders, or detecting the spread of pollutant when sensors are deployed around critical regions(chemical plants,etc.) [1]. *Barrier coverage* which is achieved by barriers of sensors, is known to be an appropriate model of coverage for these applications. Different from the *full coverage* [2] where the goal is to achieve coverage for all points in the surveillance field, *barrier coverage* is unnecessary to cover every point in the field, thus *barrier coverage* requires much fewer sensors than *full coverage*.

Every sensor node is assumed to be stationary in most of *barrier coverage* literatures[1-5]. However, there are two problems for constructing sensor barriers in a randomly deployed stationary sensor network:1) The stationary network may contain gaps, as a result, constructing sensor barriers becomes impossible. 2) To avoid gaps, it will waste many sensor nodes by increasing the deployment density. To tackle these problems, we can utilize mobile sensors. After initial random deployment, a mobile sensor communicates with others to obtain the information of networks, and compute its new position for relocation, then relocate itself to new position. Therefore, it will avoid gaps in the sensor network with mobile sensors and guarantee to construct sensor barriers with much fewer sensors than stationary networks.

For mobile sensors, it is a prominent problem that how to construct k sensor barriers with the minimum energy consumption. In *barrier coverage* model, any individual sensor cannot locally determine whether the given network provides k-barrier coverage or not [2]. However, if using global information, much communication overhead and computation cost will be brought, and it is unrealistic for large scale sensor networks. Thus, it is challenging to devise the efficient algorithm for constructing k-barriers in large scale networks with low communication and computation cost.

This paper is focused on how to achieve k-barrier coverage by the relocation of mobile sensors with the minimum energy consumption. We present an energy-efficient algorithm to construct

1-barrier coverage, and a Divide-and-Conquer algorithm to construct k-barrier coverage in large scale sensor networks.

2 Related Work

Kumar et al. [2] first introduce the notion of *weak barrier coverage* and *strong barrier coverage*, and in [2] [3] the critical conditions of weak and strong barrier coverage in a randomly deployed stationary sensor network are derived. Chen et al. [1] introduce *L-local barrier coverage*, and devise a local algorithm for providing barrier coverage. For line-based deployed sensor network, Saipulla et al. [4] establish a tight lower-bound for the existence of barrier coverage. By exploiting collaborations and information fusion between neighboring sensors, Yang et.al [5] propose a centralized algorithm to find the smallest sensor set which can information-cover the barrier. All of the above works are focused on how to achieve barrier coverage in stationary network.

Recently, some research works are investigated for barrier coverage in mobile sensor network. Bhattacharya et al. [6] consider the problem how to optimally move sensors from the interior to the perimeter of a simple polygon region for detecting intruders. Yang [7] employs a game theoretic approach to study the sensor movement strategy to defend against barrier intrusions. Shen et al. [8] propose a centralized algorithm *CBarrier* and a distributed algorithm *DBarrier* to achieve 1-barrier coverage with mobile sensors. However neither of the proposed algorithms in [8] is energy-efficient because all sensors in the network are required to move. All aforementioned works only consider 1-barrier coverage problem with mobile sensors, and do not provide the effective solution for k-barrier coverage, which is important to enhance the successful probability of movement detection. Our work can achieve k-barrier coverage energy-efficiently with mobile sensors.

3 Problem Statement

M sensors are randomly deployed in a two-dimensional rectangular strip A with the width w and the length l. $S = \{s_1, s_2 \ldots, s_M\}$ denotes the set of all sensors. The initial position of s_i is (x_i, y_i) and its relocated position is $(x_i^{'}, y_i^{'})$. The sensing model is disk, each sensor has the identical sensing radius R_s and can get its own geographic location information.

The movement of mobile sensors consumes much energy, thus how to relocate them with as less energy as possible becomes very important.And after relocating, the mobile sensors should achieve barrier coverage. We first formulate the problem how to achieve 1-barrier coverage with the minimum energy consumption as 1-BCMS (1-Barrier Coverage of Min-Sum of moving distance) problem, which is referred as follows.

Definition 1 (*1-BCMS*)**.** Given a surveillance field A and a set of mobile sensors S, find a subset S_c in S and the destination position $(x_i^{'}, y_i^{'})$ of each sensor s_i in S_c, such that the sum of moving distance of all relocating sensors is minimized, meanwhile, the relocating sensors construct one sensor barrier in A.

4 Problem Analysis

For simplicity, A is divided into N equal-sized grids, where $N = n_l \times n_w$, $n_l = \lceil l/2R_s \rceil$, $n_w = \lceil w/2R_s \rceil$. The set of central points of grids is $G = \{g_1, g_2, \ldots, g_N\}$, where g_i denotes the central point of the i^{th} grid. It is obvious that a grid can obtain 1-barrier coverage when just only one sensor is located at its central point. Thus, this paper assumes that the destination

position of relocation of each sensor is selected from G. Under the above grid model, we call a sensor barrier as a *grid barrier*, which is shown in Figure 1. In a grid barrier, every sensor is located at the central point of a grid, and the distance between neighboring sensors is no more than $2R_s$.

As shown in Figure 1, the surveillance field is 1-barrier covered after constructing a grid barrier with mobile sensors. The problem how to construct a grid barrier with the minimum moving distance could be formulated as 1-GBMS (1-Grid Barrier Min-Sum of moving distance) problem. Obviously, based on grid model, 1-GBMS problem is equivalent to 1-BCMS problem. Therefore, we study the 1-GBMS problem in the following.

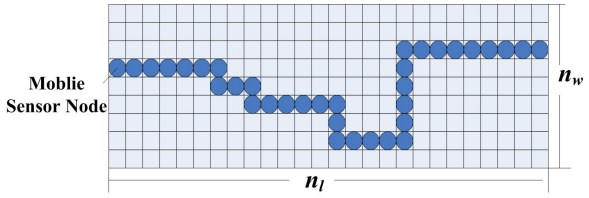


Figure 1: A grid barrier

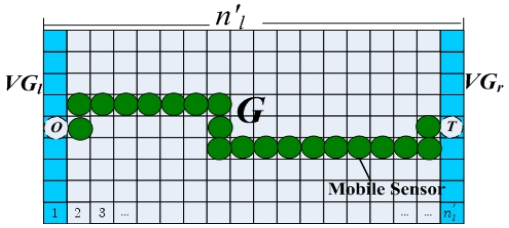


Figure 2: Region A after adding virtual vertices and virtual columns

Definition 2 (*1-GBMS*)**.** Given a set of mobile sensors S and the set G of central points of grids in a surveillance field A , find a subset S_g in S and the destination position p_i from G for each sensor s_i in S_g, such that the sum of moving distance is minimized, meanwhile, the relocating sensors construct a grid barrier in A.

In this section, the Integer Linear Programming Model of 1-GBMS is derived. As shown in Figure 2, We first add two *virtual columns* of grids beside the leftmost and rightmost column, and two distinguished vertices O and T that represent the *virtual origin* and the *virtual destination* of any barrier, respectively. The set of grids in the left virtual column is denoted by VG_l, and the set of grids in the right one is denoted by VG_r. $AG = G \bigcup VG_r \bigcup VG_l$ denotes the set of all grids, $AS = S \bigcup \{O, T\}$ denotes the set including all sensor nodes and two virtual vertices. We give sequence number to vertices in AS and AG, respectively. Specifically, The sequence number of O is 1, and that of T is M', all sequence numbers of the actual sensors are given from 2 to $M' - 1$. All grids in AG are given from 1 to N'. n'_l denotes the number of grids in one row. Let $x_{ij} = 1$, when the sensor s_i moves to the grid g_j, otherwise $x_{ij} = 0$. The distance between s_i and g_j is d_{ij}. The ILP model is described in Figure 3.

Constraint (2) and (3) enforce that a sensor is only allowed to move into at most one grid, and a grid is only allowed to be covered by at most one sensor.A barrier could be considered as a flow from O to T. Constraint (4) constrains that the flow is 1, which entering a grid in leftmost column of G from virtual vertex O , and constraint(5) guarantees that the flow leaving a grid in rightmost column of G to T is 1. By flow conservation law, if one sensor moves into a grid g_j, there must exist two sensors in the four adjacent grids(left, right, top, bottom) of g_j to make sure that the flow of entering g_j and leaving g_j are 1, respectively. This constraint is enforced by constraint (6).

We show by Theorem 1 that the 1-GBMS problem is NP-hard by restrictions to the famous Knapsack Problem.

Theorem 3. *1-GBMS Problem is NP-hard.*

Proof: Firstly, we make some restrictions of 1-GBMS problem as follows.

Minimize: $\sum_{i\in AS}\sum_{j\in AG} d_{ij}x_{ij}$ (1)

Subject to:

$\sum_{j\in AG} x_{ij} \le 1 \quad \forall i \in AS$ (2)

$\sum_{i\in AS} x_{ij} \le 1 \quad \forall j \in AG$ (3)

$x_{1k} \le \sum_{i\in S} x_{i(k+1)} \quad \forall k \in VG_l$ (4)

$x_{Mk} \le \sum_{i\in S} x_{i(k-1)} \quad \forall k \in VG_r$ (5)

$\sum_{i\in AS} x_{i(j-n'_l)} + \sum_{i\in AS} x_{i(j+n'_l)} + \sum_{i\in AS} x_{i(j-1)} + \sum_{i\in AS} x_{i(j+1)} = 2, if \sum_{i\in AS} x_{ij} = 1, \quad \forall j \in G$ (6)

$$d_{ik} = \begin{cases} 0 & i = 1, \forall k \in VG_l; \text{ or } i = M, \forall k \in VG_r \\ d_{ij} & \forall i \in S, \forall k \in G \\ +\infty & \text{otherwise} \end{cases} \quad (7)$$

$$x_{ij} = \begin{cases} 0 \text{ or } 1 & \forall i \in AS, \forall j \in AG \\ 0 & \text{otherwise} \end{cases} \quad (8)$$

Figure 3: ILP Model

Restriction1: $\forall j, k \in G, \forall i \in S$, let $d_{ij} = d_{ik}$, i.e., for any sensor s_i, the distance between s_i and any grid is equal.

Restriction2: The number of sensors that construct a grid barrier is no more than K. In original 1-GBMS problem, there is no restriction to this number.

According to Restriction1, the moving distance of a sensor is just related to its own position, i.e., if we select one sensor, the moving distance of this sensor is fixed. Thus, the sum of moving distance is just related to the set of selected sensors. We assume c_i is the cost when s_i moves to g_j, let $c_i = 1/d_i$, as a result, c_i increases as d_i decreases. Then 1-GBMS problem is reduced as a new problem how to select no more than K sensors from S, such that the sum of cost is maximized. The new problem formulation is shown as follows.

Maximize: $\sum_{i=1}^{M} c_i x_i$ (9)

Subject to:

$\sum_{i=1}^{M} 1 \times x_i \le K$ (10)

$x_i = \{0, 1\}, i = 1, 2, 3, ...M$ (11)

(11) shows that the number of selected sensors is no more than K. This formulation is the same as Linear Programming Model of Knapsack Problem [9]. Since Knapsack Problem is NP-hard,the original problem 1-GBMS is NP-hard.

□

The solution of 1-GBMS problem by ILP model is optimal, but it is polynomial unsolvable as the network size increases since it is NP-hard. Thus a polynomial approximation algorithm is required.

5 AHGB Algorithm

Under grid model, there are n_w rows of grids in A. If we deploy one sensor in each grid in a selected row, it is said to construct a *horizontal grid barrier* in A. As a result , A is 1-barrier covered. Kumar [2] proved that by constructing *horizontal grid barrier* the fewest sensors are required to achieve 1-barrier coverage. Thus the mobile sensors needed to relocate for 1-barrier coverage by constructing *horizontal grid barrier* is fewest, and the sum of moving distance in this relocation method could be less than other method with great probability. Based on above observations, we propose an approximation algorithm AHGB(Approximate to Horizontal Grid

Barrier), which mainly has two steps :1) Horizontal Grid Barrier Selection, which finds a row from A as a *horizontal grid barrier*. 2) Optimal Movement, which finds the optimal movement strategy for relocating mobile sensors to the selected *horizontal grid barrier*, subject to minimize the sum of moving distance.

5.1 Horizontal Grid Barrier Selection

If each of the mobile sensors which construct the barrier moves to its nearest grid, the sum of moving distance is minimized. Based on the above idea, we propose HGBS approach. For any grid g_k, we first assign a weight to it. The weight is the distance between g_k and the nearest sensor s_i to it. Then we compute the sum of weight of each row, and find the row which has the smallest weight as barrier_position. The HGBS approach is described as shown in Figure.4.

5.2 Optimal Movement

If we select the i^{th} row as the position of *horizontal grid barrier* by HGBS, the destination positions of mobile sensors needed to move are known. The destination positions are the n_l grids in i^{th} row. Then we select n_l nodes from M sensors, and find the optimal movement strategy to relocate the selected sensors to the n_l grids in horizontal grid barrier. We call this problem as Optimal Movement Problem.

When we find the optimal movement strategy subject to minimize the sum of moving distance, we get a one-to-one matching between the selected sensors and grid positions with the minimum sum of weight. Thus Optimal Movement Problem is equivalent to Bipartite Weighted Matching Problem [9]. The Hungarian Method is known as an optimal solution to Bipartite Weighted Matching Problem, which can be used as the solution method for Optimal Movement Problem. The computation complexity of Hungarian is $O(m^2n)$ [9], m is the size of the set which has fewer nodes, and the size of the other set is n.

HGBS Approach
begin
1. Initialize all GridSensorDistance to $+\infty$
2. for each $s_i \in S$
3. g_k=the nearest grid to s_i
4. d=distance(s_i, g_k)
5. if($d <$ GridSensorDistance(g_k))
   GridSensorDistance(g_k)=d
   end if
6. end for
8. for each $r_i \in R$
   SumRow(i)=sum of all GridSensorDistance in r_i
   end for
9. r_{min}=min(SumRow)
10. $Barrier\_Position$=Location(r_{min})

end

Figure 4: HGBS Approach

Algorithm AHGB
begin
1. Get All sensors's position information SP
2. Partition the surveillance field A to N grids,Get position information of all grids GP,$N = n_l \times n_w$, $n_w = \lceil w/2R_s \rceil$,$n_l = \lceil l/2R_s \rceil$.
3. $Barrier\_Position$=HGBS(SP, GP)
4. E=MakeEdge(S, GB)
5. C=MakeCost(E, SP, GP)
6. Construct bipartite graph G(S, GB, E, C)
7. Ψ=Hungarian(G)
8. Compute MS and DP with Match Ψ
9. for each $ms_i \in MS$
   Send the message with the destination position (dx_i, dy_i) to ms_i
   end for

end

Figure 5: Algorithm AHGB

5.3 AHGB Algorithm

The AHGB Algorithm is summarized briefly as follows. First, we get the position information of all sensors(SP) and the position information of N grids (GP), and find the *horizontal grid barrier* by HGBS approach. Then we construct a weighted bipartite graph with SPand GB

which denotes the set of grids in the selected horizontal grid barrier, and find the optimal matching by Hungarian Method to identify the sensors required to move(MS) and their destination positions(TP). At last,we send destination positions to identified sensors and move them to grids in horizontal grid barrier. The pseudocode of algorithm AHGB is shown in Figure 5.

The computation complexity of the step3 and step7 is $O(N)$ and $O(n_l^2 M)$ [9], respectively. Since $n_l > n_w$, we can obtain $n_l^2 > N$. Therefore, the computation complexity of AHGB is $O(n_l^2 M)$. The communication overhead of AHGB is $O(M)$, because we need to get the positions of all M sensors at the initial phase as shown in step1. In section 7, we show that the performance of AHGB is close to the optimal solution.

Algorithm AHGB can only construct one sensor barrier energy-efficiently, and it is not applicable for large scale sensor networks because it is centralized. In the next section, we present an efficient algorithm to achieve k-barrier coverage for large scale sensor networks.

6 Divide-and-Conquer Algorithm

k-barrier coverage is often required to guarantee to detect the moving target. According to [2], one sensor cannot locally determines whether the surveillance field is k-barrier covered or not. It is difficult to devise a distributed and energy-efficient algorithm for k-barrier coverage in large scale sensor networks. [8] presented a distributed algorithm DBarrier to construct 1-barrier coverage, which cannot provide the solution for k-barrier coverage. Furthermore, DBarrier is not energy-efficient since all sensors in the network are required to move iteratively. In this section, we propose a Divide-and-Conquer algorithm to achieve k-barrier coverage energy-efficiently for large scale sensor networks. The main idea of Divide-and-Conquer algorithm is described as follows:

1)A is divided into $k \times v$ equal-sized subregions, the length of each subregion is $l_s = l/v$,the width is $w_s = w/k$.

2) In each subregion, algorithm AHVGB(Approximate to Horizontal and Vertical Grid Barrier) independently constructs a horizontal grid barrier and a vertical grid barrier.

3) We obtain *k-barrier coverage* in A when all AHVGB are finished. Figure. 6 illustrates that A is 3-barrier covered when the Divide-and-Conquer algorithm is finished.

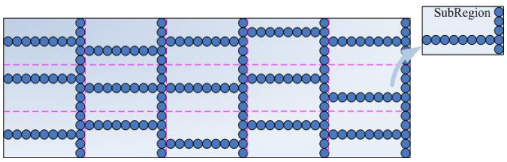


Figure 6: 3-barrier coverage when the Divide-and-Conquer algorithm is finished

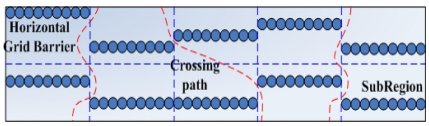


Figure 7: Gaps may exist in sensor networks

6.1 AHVGB Algorithm

As shown in Figure 7, there maybe exist a gap between two horizontal neighboring subregions.Thus, additional sensors are required to connect the two adjacent *horizontal grid barriers* to eliminate gaps. This section proposes a method which adds a *vertical grid barrier* between two adjacent subregions. Since the width of subregion is small, the sensors required to move for constructing *vertical grid barrier* is few.

We use AHVGB algorithm to construct a *horizontal grid barrier* and a *vertical grid barrier* in each subregion. In AHVGB algorithm, every subregion can independently construct barriers by

local information in its own area without communication with other subregions. The AHVGB is improved from AHGB algorithm. Specifically, when horizontal grid barrier has been selected, AHVGB constructs the weighted bipartite graph with $NewGB = GB \bigcup G_r$, but AHGB uses GB, as shown in step6 in Figure 5. GB is the set of grids in the selected horizontal grid barrier, G_r denotes the set of grids in the rightmost vertical column of A.

6.2 Low Communication Overhead and Computation Cost

In our Divide-and-Conquer algorithm, each subregion is only $\frac{1}{k \times v}$ of A. By dividing the large surveillance strip into small subregions, the message delay, communication overhead, and computation cost can be significantly reduced. The position information of a sensor node only need to be transmitted within the small subregion where the node is located, resulting in smaller delay and communication overhead compared with the whole network. The computation cost is also much lower since the number of nodes in each subregion is much less than that in the whole area.

7 Performance Evaluation

We conduct simulations by MATLAB7.0 to evaluate the performance of our algorithms. The sensing radius $R_s = 2.5m$. The mobile sensors are randomly deployed according to a Poisson point process.

In this section, we evaluate the performance of AHGB compared with the optimal solution computed by ILP model, which is polynomial unsolvable as the network size increases. In Figure 8, sensors are initially deployed in a $15m \times 500m$ strip, it shows the comparison of the sum of the moving distance between AHGB and the optimal solution, as the number of sensors increases from 100 to 800. The performance of AHGB is close to the optimal solution. In particular, the difference is only 4% when the sensor number is 300.

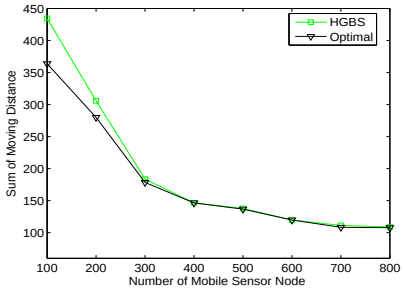


Figure 8: Sum of Moving Distance vs. Node number

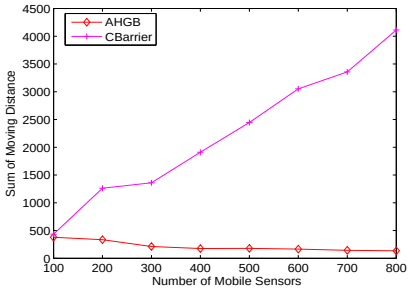


Figure 9: AHGB vs. CBarrier

We compare AHGB with CBarrier [8]which also achieves 1-barrier coverage. In Figure 9, the sum of moving distance by CBarrier increases fast as the number of sensors increases, but that of AHGB does not change much. The reason is that all sensors in the network are required to move in CBarrier, and only part of sensors is selected for relocation in AHGB .The experiment parameters of Figure9 are the same as Figure 8.

Figure 10 shows the effectiveness of Divide-and-Conquer algorithm. There are many gaps in the network after initial deployment, and 3-barrier coverage is achieved, when the algorithm is finished. Sensors are initially deployed in a $60m \times 160m$ strip according to Poisson point process with the density 0.02.

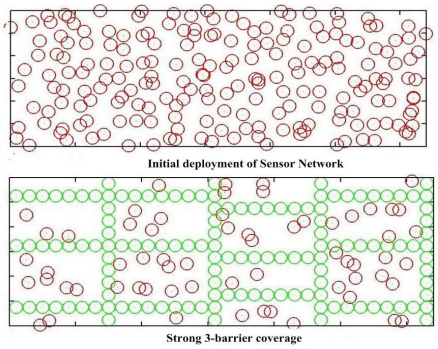


Figure 10: Effectiveness of Divide-and-Conquer algorithm

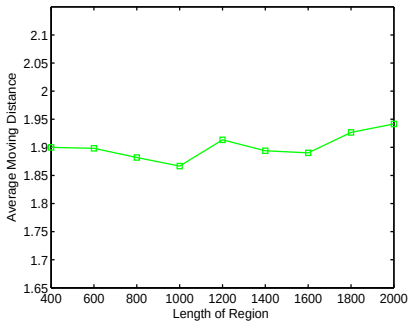


Figure 11: Average Moving Distance vs. Length of Region

As shown in Figure 11, the average moving distance of sensors does not change much, as we increase the length of strip with the constant deployment density 0.02. It implies that our Divide-and-Conquer algorithm can be applicable to large scale sensor networks. The initial experiment parameters of Figure 11 are the same as Figure 10.

8 Conclusions

In this paper, we first formulate the problem how to construct 1-barrier coverage with the minimum energy consumption in mobile sensor network as 1-BCMS problem, and analyze it in detail. Then an energy-efficient algorithm for 1-barrier coverage and an energy-efficient Divide-and-Conquer algorithm for k-barrier coverage are presented. At last, simulations demonstrate the effectiveness and energy-efficiency of the proposed algorithms. In the future, we will study how to achieve k-barrier coverage in hybrid sensor networks.

9 Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant No. 60603061, No.60903223

Bibliography

[1] Chen, S. Kumar and T. H. Lai. Designing Localized Algorithms for Barrier Coverage. *In Proc. of Mobicom07*, ACM, pp:63-74, 2007.

[2] S. Kumar, T. H. Lai and A. Arora. Barrier Coverage With Wireless Sensors. *In Proc. of Mobicom05*, ACM, pp:284-298, 2005.

[3] Benyuan Liu, Olivier Dousse and Jie Wang. Strong Barrier Coverage of Wireless Sensor Networks. *In Proc. of Mobihoc08*, ACM, pp:411-420, 2008.

[4] Anwar Saipulla, Benyuan Liu and Jie Wang. Barrier Coverage of Line-Based Deployed Wireless Sensor Networks *In Proc. of INFOCOM09*, IEEE, pp:127-135, 2009.

[5] Guanqun Yang and Daji Qiao. Barrier Information Coverage with Wireless Sensors.*In Proc.of INFOCOM09*, IEEE, pp:918-926, 2009

[6] B. Bhattacharya, M. Burmester, Y. Hu, E. Kranakis and Q. Shi. Optimal Movement of Mobile Sensors for Barrier Coverage of a Planar Region. *In Proc.of COCOA*, pp:103-115, 2008.

[7] Guanqun Yang, Wei Zhou and Daji Qiao. Defending Against Barrier Intrusions with Mobile Sensors. *In Proc.of WASA07*, pp: 113-120, 2007

[8] CX Shen, WF Cheng, XK Liao and SL Peng. Barrier coverage with mobile sensor. *In Proc. of I-SPAN08*, IEEE, pp.99-104, 2008.

[9] Eugene Lawler. *Combinatorial Optimization:Networks and Matroids*. Holt, Renehart and Winston Press, NewYork, 1976.

Int. J. of Computers, Communications & Control, ISSN 1841-9836, E-ISSN 1841-9844
Vol. V (2010), No. 5, pp. 625-633

Generic Multimodal Ontologies for Human-Agent Interaction

A. Braşoveanu, A. Manolescu, M.N. Spînu

Adrian Braşoveanu
Lucian Blaga Univeristy of Sibiu, Romania
E-mail: adrian.brasoveanu@gmail.com

Adriana Manolescu
Agora University, Oradea and R&D Agora Ltd.
Cercetare Dezvoltare Agora Oradea, Romania
E-mail: adrianamanolescu@gmail.com

Marian Nicu Spînu
Aurel Vlaicu University of Arad,
Faculty of Exact Sciences
Department of Mathematics-Informatics
Romania, 310330 Arad, 2 Elena Drăgoi

Abstract: Watching the evolution of the Semantic Web (SW) from its inception to these days we can easily observe that the main task the developers face while building it is to encode the human knowledge into ontologies and the human reasoning into dedicated reasoning engines. Now, the SW needs to have efficient mechanisms to access information by both humans and artificial agents. The most important tools in this context are ontologies. The last years have been dedicated to solving the infrastructure problems related to ontologies: ontology management, ontology matching, ontology adoption, but as time goes by and these problems are better understood the research interests in this area will surely shift towards the way in which agents will use them to communicate between them and with humans. Despite the fact that interface agents could be bilingual, it would be more efficient, safe and swift that they should use the same language to communicate with humans and with their peers. Since anthropocentric systems entail nowadays multimodal interfaces, it seems suitable to build multimodal ontologies. Generic ontologies are needed when dealing with uncertainty. Multimodal ontologies should be designed taking into account our way of thinking (mind maps, visual thinking, feedback, logic, emotions, etc.) and also the processes in which they would be involved (multimodal fusion and integration, error reduction, natural language processing, multimodal fission, etc.). By doing this it would be easier for us (and also fun) to use ontologies, but in the same time the communication with agents (and also agent to agent talk) would be enhanced. This is just one of our conclusions related to why building generic multimodal ontologies is very important for future semantic web applications.
Keywords: multimodal ontology, ontology matching, interface agents, Semantic Web, human-agent interaction

1 Introduction

The Knowledge Society (KS) is a society where information is the primary resource which can be consumed by both humans and machines. If we want to build such a society in a proper way we need different kinds of infrastructure: hardware, software, organizational, etc. SW and

agents represent only a small part of the large infrastructure needed in order to build the true KS.

SW ([1], [2], [3], and [4]) is one of those disruptive technologies which tend to be talked about years before their coming of age. One of the visions presented in [1] was that of agents replacing humans for simple everyday tasks like buying tickets for a concert or making appointments to the doctor. The main reason why this vision hasn't yet come to life is one that is now well understood and also explained in the article's revision [2]: encoding the human knowledge into ontologies and the human reasoning into dedicated reasoning engines is not an easy task. This process requires trans-disciplinary knowledge, dedicated tools and repositories, and advanced techniques from mathematics, logics and software. It is in fact an extremely difficult procedure which relies entirely on the cooperation between hundreds or thousands of organizations and different standards. Since the standardization processes take a long time even in these days and the time of adoption for new technologies is sometimes around 2-3 years at least, we should not be surprised that it will take a while until the SW reaches the critical mass.

Ontologies represent the key to a successful communication between human and agents if they are done right. We are only beginning to understand the implications of using the ontologies for the great tasks we assigned for them, but some problems like ontology management (versioning, change, tools and standards), ontology matching (finding correspondences between different ontologies) and the adoption of ontologies on large scale by developers and users proved to be quite challenging. Ontology dynamics is definitely a field on which we should keep an eye on. According to [30] there is still no clear winner in the process of ontology matching (in other words: a standard or a methodology with clear rules to match almost everything automatically or semi-automatically since sometimes humans will need to check the results). Therefore we should not be surprised, when reading a journal or conference proceeding, that most of the articles refer to these tasks rather than to the desired using of ontologies which is to give agents a way of understanding our world and reason about it. It is the way things should be: in order to build a functional system we always need to have its parts figured out. We should however not lose sight of the system we need to build and this is one of the purposes of this paper: to look at the current state of the art in several fields of study and see if we are heading in the right direction. In this context we will especially examine some problems related to the multimodal communication between human and agent and try to see how they are solved by using ontologies.

2 Rationale and Approach: Why Complicate Things and Use Generic Multimodal Ontologies?

First we need to clear one question: what is an ontology? Some answers to this (and also some examples of how to use ontologies) can be found in [12], [15], [16], [17], [22], [23] and [31]. The classic definition proposed by Gruber tells us that an ontology is "explicit specification of a conceptualization" [12]. This definition is examined and extended by many papers, most recently by Guarino, Oberle and Staab in [16] which also focuses on the importance of "shared explicit specifications" because without commiting to ontologies every agent would understand something else (they also take the opportunity to revise the semiotic triangle). Ontologies are us, Mika's thesis [23] is a simple yet powerful statement. It tells us that since we are the ones who design the ontologies they will only express what we want them to express and will sometimes be useless without the context in which they have been created.

The main problem when designing ontologies is to carefully choose the concepts within a domain and the relationships between them in such a way as the ontology to be well founded because "any ontology will always be less complete and less formal than it would be desirable in

theory" [16]. In the light of this statement it should become quite clear why we sometimes need to use generic ontologies: there is simply no other way to address the problem of uncertainty when developing ontologies than genericity.

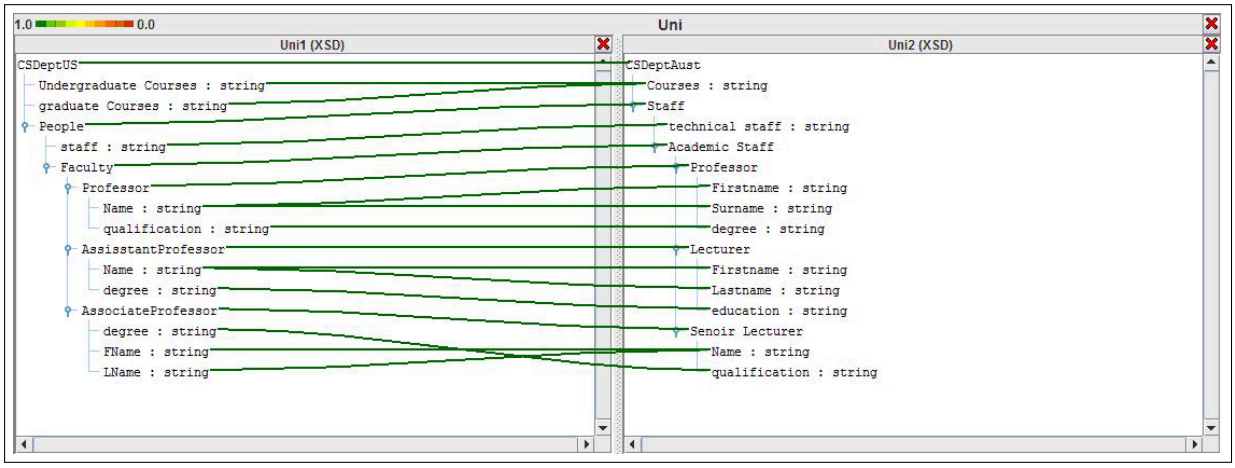


Figure 1: One of the most popular programs for ontology matching: COMA++, developed at the University of Leipzig. In this screenshot we can see how we can establish some correspondences between two ontologies representing a Computer Science Department

Nowadays there are probably thousands of ontologies in use, but if the SW will ever look like Berners-Lee's visions then ontologies will be common place for every designer, developer or user. Usually an ontology only addresses the problems from a narrow field of knowledge (domain ontology) so it is not uncommon that applications may use many ontologies for different purposes. In some of these cases it is useful to also use upper level ontologies which are general ontologies that represent concepts that are the same across all domains. A unique upper level ontology which should encompass all the human knowledge is not feasible and will never be built because of practical reasons (each society has its concepts, every field of knowledge has a certain language to protect itself, etc.), but upper level ontologies are used for mediation mainly in the idea that universal agreement between different ontologies will be/is possible. In other cases in order to use different ontologies the applications will use ontology matching schemes like those discussed in [10]. Since ontologies are the building blocks of SW, any application from this area must use them, even if that means adding layers of complexity because of the matching process, APIs, uncertainty. For everybody working in the IT industry these days it should be clear that the medium in which we work is becoming more and more like OHDUE (Open Heterogenous Dynamic Uncertain Environment) [8] and ontologies are part of this medium. These issues are addressed in articles and books like [10], [19], [30] (ontology matching), [26] (automatic generation of ontology APIs), [8] (OHDUE, agents). Because the field of ontology engineering is becoming more popular we should not be surprised that we will also hear a lot about the ontology driven software engineering. Ontology Driven Information Systems (ODIS) [36] is just one of the recent examples which fell into this category.

Given all these complications that appear when designing and working with ontologies it is interesting to ask a new question: why would we want to complicate our life even more by using multimodal ontologies? It is not enough that the ontology management or ontology matching problems still pose so many challenges? Are these new breed of ontologies even feasible?

Certainly from a user's perspective multiple modalities to enter input into a system (touch, voice, mouse, pen, etc.) can only mean increased usability (do we need to remember how touch screens became the norm in the mobile phones industry after iPhone was launched?), while from

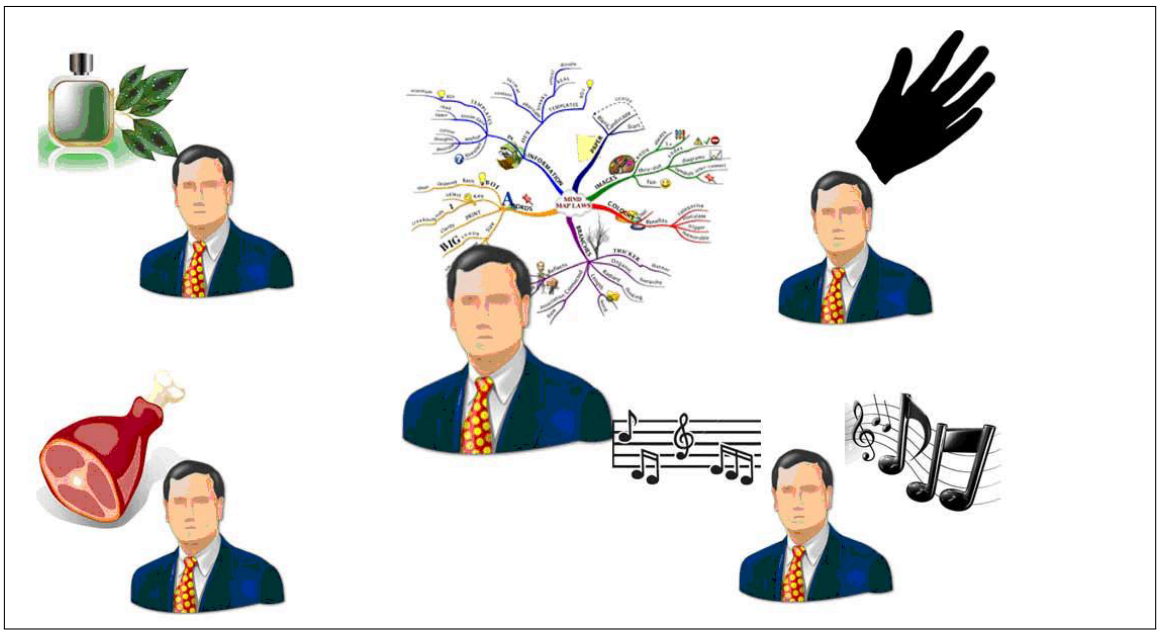


Figure 2: The multimodal communication dream: to use all the five senses (smell, sight, touch, taste, sound) during the process of communication.

a developer's perspective this means that software gets even more complicated than it is now. This is the right moment for such a development since for the multiple streams of data that come with multimodal communication we need distributed systems. Since multi-core processors are now luckily the norm in desktop computing we should have no problem (at least not hardware) dealing with the huge flux of data. In the past 40 years scientists have developed different mechanisms for getting audio, video and touch input, but the integration of all five senses in the communication between man and machine remains a dream. It is enough however to use one sense in different ways (for example for seeing we have images, text, video) to be able to speak about multimodal communication. In this respect different research groups (most notably [29]) started to develop also multimodal ontologies, but most of them took the approach of developing different ontologies for text, images, video or voice and then use ontology alignment to match them (multimodal integration through ontology matching [29]). A multimodal ontology gets us all the benefits of having such different ontologies. Like all things in life, multimodal ontologies do not come without bad parts (even harder to design, mantain and match), but they are definitely closer to our way of thinking. Is this a sufficient reason to try it? It might not be, but it is not the only one. The usage of multimodal ontologies will allow us to give a more natural, even realistic, feeling during communication between agents and humans, enhanced usability, the possibility to model mechanisms that are closer to the way we understand the world (diagrams, mind maps, feedback, brainstorming, slides, visual thinking, and others). It should be clear that it's not just art for art's sake, but rather art for a better life in the future.

3 Generic Multimodal Ontologies for Human-Agent Interaction

The process of multimodal ontology modeling is still open to exploratory research because ontologies are not everywhere. Without ontologies for all possible fields, and tools to match these ontologies it is debatable whether we will achieve an efficient semantic web, but rather the illusion of a semantic web maintained by few successful applications in certain areas (like social

networking, language translation or medicine). Since multimodal communication is difficult to process it is clear that in the first phase of any research regarding this subject, the communication between agents and humans will not be efficient. The question we need to ask ourselves in this situation is: If it is not efficient why should we bother at all to try something like this? The answer is simple and is typical for exploratory research: It takes time to find the best way to integrate multiple streams of data in an efficient manner and it also takes time to develop efficient ontology matching processes for such tasks. The role of exploratory research is to discover niches. The task of creating efficient mechanisms is one best suited for incremental research. Since this area of research is relatively new there is enough room for exploratory research and for breakthroughs.

Generic ontologies are rarely used by developers. Most of the articles present different ontologies and clearly state that they do not use generic ontologies because the problem's domain was well understood. Generic ontologies are best suited for modelling as we can see from [17], and [13]. It is easier to say you have an ontology with few concepts and not define all of them when doing modelling. The task of defining all the concepts and relationships between them is one that remains to the ontology engineer or to the developer. When dealing with models that are related to multimodal communication it makes sense to use generic multimodal ontologies. It also makes sense to use a generic ontology whenever dealing with uncertainty as suggested by [8] [28].

The agents of tomorrow will be built taking into account recent findings like the requirements-driven self-reconfiguration [6], multi-party, multi-issue, multi-strategy negotiation [35], natural language [18], and controlled natural language [32]. If we are to follow Berners-Lee vision from [1] we absolutely need to integrate such findings into our work. In fact according to [18] ontologies are the "common ground for virtual humans". Their architecture suggests using multimodal communication, but this is not clearly stated in the article since the ontology is not multimodal. If we look at [6] and [35] we can envision agents that dynamically change their strategies according to the environment and the context of conversations. This requires designing flexible ontologies, another reason to make them generic.

The agents must use ontologies if they are to understand something from this world. They also need to share them and commit to them if we want them to be able to talk between them. The multimodal ontology helps in some of the phases of multimodal communication: fusion and integration (getting the input from different channels), natural language processing, disambiguation, error reduction and fission (preparing the output). When designing a multimodal ontology one must also take into account the problems related to designing multimodal systems as described in [25], and also the medium in which these agents will evolve because an agent that needs to evolve in the urban computing environment [34] will have different needs than an agent that just surfs the web. The focus of research is usually on multimodal fusion, but a recent survey [9] shows that the interest in multimedia fission is increasing. Designing a multimodal ontology thus requires taking into account all these findings because the agent must be able to give us a response not only to understand our requirements. Probably one of the big challenges ahead is to annotate the multimodal content in real-time. This is particularly hard to do for video content, but not impossible, as [27] suggests. M3O (Multimedia Metadata Ontology) allows us to annotate the multimedia content from a page to retrieve it easier. If such ontologies will be improved then the road to the visions from [1] will be shorter.

4 Related Work

The current state of the art in multimodal HCI is presented in [7] and [20]. One of the conclusions from [7] leaves further space for improvements: "most researchers process each channel (visual, audio) independently, and multimodal fusion is still in its infancy". The same can be

rendered as true for the multimodal ontologies too. Since [7] is more recent we will use it as a basis for further investigation in this field.

Since there are only few interesting articles related to multimodal ontologies every year, we have selected a few of them to be used as basis for future research.

When searching for definitions related to ontologies and trends in the field of ontology development /matching some of the best research groups in the world are the ones from Trento (LOA and University of Trento), and Koblenz-Landau. Many of the articles cited in this paper come from some of the members of the Trento group: [6], [10], [16], [17], [30]. These are related to definitions of ontology, ontology matching, and modelling with ontologies. We have also used articles from the Koblenz-Landau group: [6], [26], [27] related to definitions, automatic generations of ontology APIs and M3O.

One interesting idea is that of multimodal context-aware interaction presented by Cearreta and his team in [5]. If we have to model emotions there might be no other solution than to use multimodal ontologies combined with special reasoners. Another article related to our subject is [29]. Their approach of using different ontologies for text and images and then use ontology matching can definitely be improved on the long term. They clearly state that for the moment multimodal ontology do not offer fast communication, but that in time speed might be improved. Also [24], [32] and [33] study the relationships between Natural Language Processing (NLP) and SW. The work of these research groups must be studied. One of them [32] is from Southampton, one of the workplaces of Timothy Berners-Lee.

When it comes to generic ontologies and tools for working with ontologies, one of the best research groups that needs to be followed is Stanford's [11], [28]. Their work on biomedicine ontologies and Protégé is fundamental.

5 Conclusions and Future Work

The SW tools are now an important part of the IT industry, the main clients coming from the fields of biomedicine, aeronautics, automotive, government and local administrations, and media. This sudden interest might be related to the success of social media [14], [21] and means that developers are starting to tap into the potential promises of the field. Even so there is a lot of work to be done regarding multimodal ontologies. The reason is one that was mentioned several times during this paper: the task of designing such ontologies is still difficult. As we do not have yet universal methods for ontology matching we do not have a clear methodology of designing multimodal ontologies (regardless of the fact that they are generic or not).

The main advantages of using generic multimodal ontologies should be better understood now: they offer us a modality to design the process of communication with agents as close to our way of thinking as possible and also play a very important role in several phases of the multimodal communication (multimodal fusion and integration, disambiguation, NLP, error reduction, multimodal fission, etc.). The main disadvantage will probably be efficiency for the next years, but given the exploratory nature of the research this is normal.

The future work of our group will consider implementing new mechanisms for linking the generic multimodal ontologies and affective interfaces with recent research in Semantic Web and

HCI in a 3 years interval (during the PhD studies of the first author). The objectives are to be fulfilled involving European teams of researchers interested in this kind of projects.

Acknowledgements

This work was partially supported by the strategic grant POSDRU/88/1.5/S/60370(2009) on "Doctoral Scholarships" of the Ministry of Labour, Family and Social Protection, Romania, co-financed by the European Social Fund - Investing in People.

Bibliography

[1] T. Berners-Lee, J. Hendler, O. Lassila. *The Semantic Web*. Scientific American, May 2001, 34-43.

[2] N. Shadbolt, W. Hall, T. Berners-Lee. *The Semantic Web revisited*. IEEE Intelligent Systems, pages 96- 101, May/June 2006.

[3] T. Berners-Lee, W. Hall, J.A. Hendler, K. O'Hara, N. Shadbolt, D.J. Weitzner. *A Framework for Web Science*. Foundations and Trends in Web Science, 1 (1), pages 1-130, 2006.

[4] C. Bizer, T. Heath, T. Berners-Lee. *Linked Data - The Story So Far*. International Journal on Semantic Web and Information Systems, Volume 5, Issue 3.

[5] I. Cearreta, J. M. Lopez, N. Garay-Vitoria. *Modelling multimodal context-aware affective interaction*. Proceedings of the Doctoral Consortium of the Second international conference on ACII'07. Lisbon, Portugal, 57-64, 2007.

[6] F. Dalpiaz, P. Giorgini, J. Mylopoulos. *An Architecture for Requirements-driven Self-Reconfiguration*. Proc. of the 21st Int. Conf. on Advanced Information Systems Engineering, LNCS 5565, Springer, 246- 260, http://www.disi.unitn.it/ pgiorgio/papers/caise09-b.pdf, 2009.

[7] B. Dumas, D. Lalanne, S. Oviatt. *Multimodal Interfaces: A Survey of Principles, Models and Frameworks*. In D. Lalame, J. Kohlas, editors, Human Machine Interaction Research Results of the MMI Program, Springer, 3-27, 2009.

[8] I. Dzitac, B.E. Barbat. Artificial Intelligence + Distributed Systems = Agents. *International Journal of Computers, Communications & Control*, ISSN 1841-9836, 4(1):17-26, 2009.

[9] D.W. Embley, A. Zitzelberger. Theoretical Foundations for Enabling a Web of Knowledge. Retrieved from: http://dithers.cs.byu.edu/tango/papers/formalWoK.pdf, 2009.

[10] J. Euzenat, P. Shvaiko. *Ontology Matching*. Springer, 2007

[11] A. Ghazvinian, N. F. Noy, C. Jonquet, N. H. Shah, M. A. Musen. *What Four Million Mappings Can Tell You about Two Hundred Ontologies*. International Semantic Web Conference 2009: 229-242

[12] T. R. Gruber. *A Translation Approach to Portable Ontologies*. Knowledge Acquisition, 5(2):199- 220, 1993.

[13] M. Gruninger. Designing and Evaluating Generic Ontologies. In 'ECAI96's workshop on Ontological Engineering'.

[14] T. Gruber. *Collective knowledge systems: Where the social web meets the semantic web*. Journal of Web Semantics, 6(1):4-13, 2008.

[15] N. Guarino. *The Ontological Level: Revisiting 30 Years of Knowledge Representation*. In A. Borgida, V. Chaudhri, P. Giorgini, E. Yu (eds.), Conceptual Modelling: Foundations and Applications, Springer Verlag 2009: 52-67.

[16] N. Guarino, D. Oberle, S. Staab. *What is an Ontology?* In S. Staab and R. Studer (eds.), Handbook on Ontologies, Second Edition. International handbooks on information systems. Springer Verlag: 1-17, 2009.

[17] G. Guizzardi, T. Halpin. *Ontological foundations for conceptual modeling.* Applied Ontology 3, 1- 12, 2008.

[18] A. Hartholt, T. Russ, D. Traum, E. Hovy, S. Robinson. *A common ground for virtual humans: Using an ontology in a natural language oriented virtual human architecture.* In: Language Resources and Evaluation Conference (LREC). (May 2008)

[19] W. Hu, Y. Qu. Falcon-AO: *A practical ontology matching system.* Web Semantics: Science, Services and Agents on the World Wide Web 6 (2008) 237-239.

[20] A. Jaimez, N. Sebe. *Multimodal human-computer interaction: A survey.* Computer Vision and Image Understanding, Volume 108, Issues 1-2, October-November 2007, 116-134, Special Issue on Vision for Human-Computer Interaction, 2007.

[21] F. Limpens, F.Gandon, and M. Buffa. *Linking folksonomies and ontologies for supporting knowledge sharing: a state of the art.* Technical report, EU Project, ISICIL, 2009.

[22] D. Lonsdale, D. W. Embley, Y. Ding, L. Xu, M. Hepp. *Reusing Ontologies and Language Components for Ontology Generation, accepted for publication in Data and Knowledge Engineering.* Retrieved from: http://www.heppnetz.de/files/dke2008.pdf, 2010.

[23] P. Mika. *Social Networks and The Semantic Web*, Springer, 2007

[24] J. Niekrasz and M. Purver. *A multimodal discourse ontology for meeting understanding.* In The 2nd Joint Workshop on Multimodal Interaction and Related, 2005.

[25] L. Nigay, J. Coutaz. *A design space for multimodal systems: Concurrent processing and data fusion.* ACM Conf. Human Factors in Computing Systems (CHI), 1993.

[26] F. S. Parreiras, C. Saathoff, T. Walter, T. Franz, S. Staab. APIs a gogo: Automatic Generation of Ontology APIs. icsc, 342-348, 2009 IEEE International Conference on Semantic Computing, 2009

[27] C. Saathoff, A. Scherp. M3O: *The Multimedia Metadata Ontology.* Proceedings of the Workshop on Semantic Multimedia Database Technologies, 10th International Workshop of the Multimedia Metadata Community (SeMuDaTe 2009), Graz, Austria, 2009.

[28] Abraham Sebastian, Natalya Fridman Noy, Tania Tudorache, and Mark A. Musen. *A generic ontology for collaborative ontology-development workflows.* In Aldo Gangemi and Jérôme Euzenat, editors, EKAW, volume 5268 of Lecture Notes in Computer Science, 318-328. Springer, 2008.

[29] A.A.A. Shareha, M. Rajeswari, D. Ramachandram. *Multimodal Integration (Image and Text) Using Ontology Alignment.* American Journal of Applied Sciences 6 (6): 1217-1224, 2009.

[30] P. Shvaiko, J. Euzenat. Ten challenges for ontology matching. In Proceedings of the 7th International Conference on Ontologies, DataBases, and Applications of Semantics (ODBASE), pages 1164-1182, Monterrey (MX), 2008.

[31] W. V. Siricharoen. *Ontology Modeling and Object Modeling in Software Engineering.* International Journal of Software Engineering and Its Applications, Vol. 3, No. 1, January, 2009, 43-59, 2009.

[32] P. Smart, J. Bao, D. Braines, N. Shadbolt. *Development of a Controlled Natural Language Interface for Semantic MediaWiki*. In: Proceedings of the Workshop on Controlled Natural Language, Springer- Verlag, Heidelberg, Germany.

[33] D. Sonntag, M. Romanelli. *A multimodal result ontology for integrated semantic web dialogue applications*. In Proceedings of the 5th Conference on Language Resources and Evaluation (LREC 2006), Genova, Italy, May 24-26.

[34] A. Tenschert, M. Assel, A. Cheptsov, G. Gallizo, E. Della Valle, I. Celino. *Parallelization and Distribution Techniques for Ontology Matching in Urban Computing Environments*. OM 2009

[35] D. Traum, S. Marsella, J. Gratch, J. Lee, and A. Hartholt.. *Multi-party, multi-issue, multi-strategy negotiation for multi-modal virtual agents*. In Proc. of Intelligent Virtual Agents Conference IVA-2008.

[36] M. Uschold. *Ontology-Driven Information Systems: Past, Present and Future*. In Proceedings of the 5thInternational Conference on Formal Ontology in Information Systems (FOIS2008), Saarbrücken, Germany, (Oct 31st - Nov 3rd), 2008.

Int. J. of Computers, Communications & Control, ISSN 1841-9836, E-ISSN 1841-9844
Vol. V (2010), No. 5, pp. 634-641

Parallel Simulation of Quantum Search

S. Caraiman, V. Manta

Simona Caraiman, Vasile Manta
"Gheorghe Asachi" Technical University of Iasi
Romania, 700050 Iasi, 27 Dimitrie Mangeron
E-mail: {sarustei,vmanta}@cs.tuiasi.ro

Abstract: Simulation of quantum computers using classical computers is a computationally hard problem, requiring a huge amount of operations and storage. Parallelization can alleviate this problem, allowing the simulation of more qubits at the same time or the same number of qubits to be simulated in less time. A promising approach is represented by executing these simulators in Grid systems that can provide access to high performance resources. In this paper we present a parallel implementation of the QC-lib quantum computer simulator deployed as a Grid service. Using a specific scheme for partitioning the terms describing quantum states and efficient parallelization of the general singe qubit operator and of the controlled operators, very good speed-ups were obtained for the simulation of the quantum search problem.
Keywords: quantum computer simulation, parallel computing, quantum search.

1 Introduction

The research in quantum informatics has gained an immense interest due to the remarkable results obtained for the factorization [11] and search [6] problems. These results prove the huge computational power of a quantum machine with respect to the classical computers. However, building quantum computers represents an immense technological challenge and, at present, the quantum hardware is only available in research labs. Under these circumstances quantum simulators have become valuable instruments in developing and testing quantum algorithms and in the simulation of physical models used in the implementation of a quantum processor.

According to Feynman's paper [3], classical computers will never be able to simulate quantum systems in polynomial time. The simulation of 29 qubits (quantum bits) uses 32 GB of memory [1] and any additional qubit doubles the resources needed: time, memory, computational power and space.

In this paper we present a solution based on Grid computing for the quantum simulation problem. Our simulator relies on paralel processing for storing quantum states and applying quantum operators. The deployment of this solution in Grid systems provides access to high performance computing devices for simulation and availability in the context of collaboration through the means of Virtual Organizations.

Our quantum simulator, GQCL, partitions the terms coresponding to a quantum state between several procesing nodes using a scheme that minimizes communication between nodes during the application of quantum operators. In a previous paper [1] we describe the development of a grid service that provides this functionality to client applications by enabling the Quantum Computation Language [9] through a parallel implementation of the QC-lib simulator [8]. The results recorded for the application of the Hadamard transform illustrate the performances of this approach [1]. In the following we present the parallelization of the general single qubit operator, the conditional operators and of the measurement process. This allows us to study the performance of our simulator regarding the quantum search problem.

2 Basic Concepts in Quantum Computing

The quantum analogous of the classical bit is the qubit. A qubit is a quantum system whose states can be completely described by the superposition of two orthonormal basis states, labeled $|0\rangle$ and $|1\rangle$ (in a Hilbert space $\mathcal{H} = C^2$, $|0\rangle = (1 \quad 0)^T$, $|1\rangle = (0 \quad 1)^T$). Any state $|\Psi\rangle$ can be described by:

$$|\Psi\rangle = \alpha|0\rangle + \beta|1\rangle, \quad |\alpha|^2 + |\beta|^2 = 1, \tag{1}$$

where α and β are complex numbers. Thus, unlike the classical bit, the qubit can also be in a state different from $|0\rangle$ and $|1\rangle$: linear combinations of states can be formed, called superpositions (eq. 1). When measuring a qubit either the result 0 is obtained, with probability $|\alpha|^2$, or 1 with probability $|\beta|^2$. The sum of the probabilities must be 1, so the state of a qubit represents a unit vector in a complex bi-dimensional vector space.

A collection of n qubits is called a quantum register with dimension n. The general state of a n-qubit register is

$$|\Psi\rangle = \sum_{i=0}^{2^n-1} a_i|i\rangle, \tag{2}$$

where $a_i \in \mathcal{C}$, $\sum_{i=0}^{2^n-1} |a_i|^2 = 1$. This means that the state of a n-qubit register is represented by a complex unit vector in Hilbert space $\mathcal{H}_{2^n}$.

The quantum analogous of the classical NOT gate is labeled X and can be defined such that $X|0\rangle = |1\rangle$ and $X|1\rangle = |0\rangle$. The quantum NOT gate acts similarly with its classical counterpart, although, unlike in the classical case, its action is linear: state $\alpha|0\rangle + \beta|1\rangle$ is transformed in a coresponding state $\beta|0\rangle + \alpha|1\rangle$. A convenient way of representing the action of the quantum NOT gate is in matrix form:

$$X = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}. \tag{3}$$

Controlled gates are quantum logical gates acting on more than one qubit. The notion of controlled gate allows the implementation of the $\mathtt{if-else}$ constructs. Quantum controlled gates use a control qubit to determine whether a specific unitary action is applied to a target qubit.

The controlled-NOT operator (CNOT) is the prototypical multi-qubit gate. The first parameter of a CNOT gate is the control qubit. If this qubit is in state $|0\rangle$, the target qubit is left unchanged and if the control qubit is in state $|1\rangle$, the target qubit is flipped:

$$|00\rangle \to |00\rangle; \quad |01\rangle \to |01\rangle; \quad |10\rangle \to |11\rangle; \quad |11\rangle \to |10\rangle.$$

The CNOT operator is a generalization of the classical XOR, since its action can be summarized as $|x, y\rangle \to |x, x \oplus y\rangle$, where $\oplus$ is addition modulo two. The matrix representation of CNOT is:

$$\mathrm{CNOT} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{pmatrix}. \tag{4}$$

There are several other multi-qubit gates, Nevertheless, the controlled-NOT gate and the single qubit gates represent the prototypes for any other quantum gate because of the folowing remarkable universality result: any multi-qubit gate can be built out of CNOT gates and single qubit gates. The proof of this statement represents the quantum analogous to the universality of the classical NAND gate.

3 Parallel Simulation of Quantum Computation

The state of a n-qubit register is represented by a complex unit vector in Hilbert space $\mathcal{H}_{2^n}$. Storing a complex number $a = x + iy$ on a classical computer requires storing the pair of real numbers (x, y) for which the 8 byte representation is prefered. Thus, in order to store an n-qubit quantum register using a conventional (classical) computer, at least 2^{n+4} bytes are required. The memory needed for simulating a n-qubit quantm computer grows exponentially with respect to the number n. For example, when $n = 24$ ($n = 36$) at least 256 MB (1 TB) of memory is required to store a single arbitrary state $|\Psi\rangle$. The time evolution of a n-qubit quantum register is determined by a unitary operator defined in the $\mathcal{H}_{2^n}$ space. The matrix dimension is 2^n x 2^n. In general, 2^n x 2^n space and $2^n(2^{n+1} - 1)$ arithmetic operations are needed to execute such an evolution step.

Thus, the simulation of a quantum computer using a classical device represents a computationally hard problem and the memory and processor generate drastic limitations on the size of the quantum computer that can be simulated. Because of the exponential behavior of quantum systems, simulation using classical computers enforces the use of exponential memory space and the execution of an exponential number of operations. It is obvious that the simulation of quantum problems of interesting sizes enforces the use of high performance computing devices. Parallel computing can represent a solution to this problem [5, 7, 10, 13].

Nevertheless, the development of a quantum simulator must consider another important aspect: it has to be easily accessible. But this contradicts the first requirement, that it is parallel, which deeply restricts the group of potential users. A solution based on the concept of Grid systems ca be used to solve this contradiction and to provide the scientific community with an useful and easily accessible instrument. The Grid concept adresses the problem of coordinated resource sharing and problem solving in dinamic, multi-institutional virtual organisations [4].

A Grid enabled quantum computer simulator is GQCL [1]. This simulator allows the use of the QCL [9] quantum programming language to implement quantum algorithms and the quantum programs are executed using a parallel version of the QC-lib simulation library [8]. Using a specific data partitioning scheme and efficient storing of quantum states allowed very good speedups and efficiency of this parallel implementation which will be discussed in the following.

3.1 Overview of GQCL Grid Service for Quantum Simulation

Our quantum computer simulator is based on the QC-lib simulation library which provides a framework to execute programs written in a quantum programming language, QCL, in the absence of quantum hardware. The reasons that lead to this choice for a quantum programming language are detailed in [1] and mainly consider the representation of the quantum state using complex numbers, the possibility to write complex quantum operators, the classical extension and its universality. QCL was conceived by Ömer [9] and the first version appeared in 1998 and the last one in 2004. It is open-source running under Linux operating system and it is a procedural high level language with a C like syntax. QC-lib is a C++ library for the simulation of quantum computers at an abstract functional level [8], and it is used as the back end interpreter for QCL.

For the execution of quantum programs written in QCL we developed a parallel version of the QC-lib simulator in which the terms representing quantum states are distributed across multiple processing nodes. We have chosen to expose the parallel implementation of QC-lib through a Globus Toolkit 4 (GT4 for short) Grid service. Parallelization has been achieved through the use of LAM 7.1.2/7.1.4 implementation of the MPI-2 standard. In GQCL, the execution of

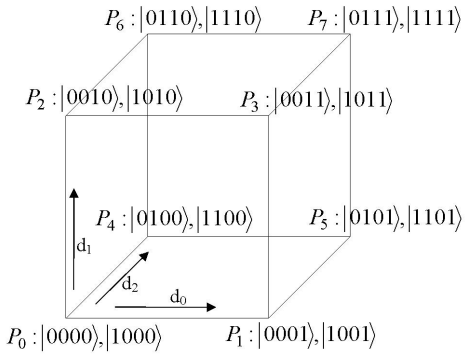


Figure 1: Distribution of the 16 basis states of a 4-qubit register using $2^3 = 8$ proccesing nodes. The processing nodes represent the corners of a 3-dimensional hypercube.

the MPI implementation of QC-lib is enabled through the means of a wrapper service based on the Factory/Instance architecture [12]. In GQCL, the instance service, is responsible for actually managing the quantum simulation as a Grid job. Through the use of a *WS-Resource*, the instance service starts and monitors the job state, notifying the client application on any relevant changes. Also, file staging is automated and when the job finishes, the client is given access to the results of the simulation.

The architectural details of GQCL and its advantages are presented in [1]. In the following we discuss the complete parallel implementation of the simulation library, addressing the issues regarding the representation of quantum states using multiple processing nodes, the application of single- and multi-qubit quantum operators and measurements.

3.2 Parallel Implementation of the QC Library

A quantum register in QCL contains a number of basis vectors, each with a corresponding amplitude. When the qubits forming the quantum register are in a superposition of states, the number of vectors grows exponentially. In QC-lib, the superposition state of a 1-qubit register is represented by two basis vectors (terms) for which the correponding complex amplitudes must be stored: $2 \text{ x } \texttt{complex} < \texttt{double} >= 32$ bytes. When appplying a quantum operator, two term lists are created: one for storing the terms in the current state and one to accumulate the result of an operation on the state. This gives a total of 64 bytes/term which for a n-qubit register requires the use of 2^{n+6} bytes. Thus for $n = 25$ ($n = 29$) qubits 2 GB (32 GB) of memory is necessary.

In order to provide an efficient parallel execution of QC-lib we take advantage of the specific form of the quantum computation process and distribute the 2^n basis states of a quantum register to 2^p processors based on the p least significant bits. In this representation, the processing nodes and the basis vectors are actually considered the coordinates of a n-dimensional hypercube (Figure 1). Each processing node applies quantum operators only to local terms and communicates the generated terms to corresponding processing nodes if necessary. Another feature of our implementation is that for a quantum state only non-zero amplitude terms are stored thus diminishing the communication costs in early stages of some operator execution and the space required to store a quantum state.

The General Single Qubit Operator. Communication between processing nodes is only necessary when applying the operator to a qubit determining the data distribution. Applying a

general single qubit operator $U = \begin{pmatrix} u_{11} & u_{12} \\ u_{21} & u_{22} \end{pmatrix}$ on a single qubit with state $|\Psi\rangle = \alpha|0\rangle + \beta|1\rangle$ yelds

$$U|\Psi\rangle = (\alpha u_{11} + \beta u_{12})|0\rangle + (\alpha u_{21} + \beta u_{22})|1\rangle \tag{5}$$

If 2 processors are used, then each processor holds the amplitude of one basis state. Applying operator U locally on each processor, terms are created that are not owned by the processor, and so communication of these terms is needed:

$$\begin{array}{l} P_0 : \alpha|0\rangle \xrightarrow{U} u_{11}\alpha|0\rangle + u_{21}\alpha|1\rangle \\ P_1 : \beta|1\rangle \xrightarrow{U} u_{12}\beta|0\rangle + u_{22}\beta|1\rangle \end{array} \xrightarrow{\text{Comm}} \begin{array}{l} u_{11}\alpha|0\rangle + u_{12}\beta|0\rangle \\ u_{21}\alpha|1\rangle + u_{22}\beta|1\rangle \end{array}$$

For each term in the initial state at most two terms are created, out of which at most one needs to be communicated. If working, for example, with an n-qubit register and 2^p processors, communication is necessary only when applying a single qubit operator on any of the qubits that form the distribution key. For the rest of the qubits, all the terms needed for computing the amplitude of the resulting state are locally owned by each processor. Moreover, in the first case, for each processor, all the remotely owned terms are owned by the same other processor as a single bit is flipped in the distribution key.

The parallel implementation of the general single qubit operator allows the parallel execution of NOT, Hadamard, phase shift of the amplitude and exponentiation gates.

Controlled Gates. CNOT operator (controlled-NOT) In the parallel implementation of QC-lib, when the control qubit is in state $|0\rangle$, the state of the target qubit doesn't change, so no new terms are generated and the amplitudes of existing terms are left unchanged. When the control qubit is in state $|1\rangle$, the state of the target qubit is flipped. In this case new terms are generated that need to be communicated to another processing node if the target qubit is part of the distribution key. For example, working with 4 processing nodes and applying CNOT to the least significant qubit in a 3-qubit register initially in the state $\alpha|100\rangle + \beta|011\rangle$, and the control qubit is qubit 2, the following evolution is obtained:

$$\begin{array}{l} P_0 : \alpha|100\rangle \xrightarrow{\text{CNOT}} \alpha|101\rangle \\ P_1 : - \\ P_2 : - \\ P_3 : \beta|011\rangle \xrightarrow{\text{CNOT}} \beta|011\rangle \end{array} \xrightarrow{\text{Comm}} \begin{array}{c} - \\ \alpha|101\rangle \\ - \\ \beta|011\rangle \end{array}$$

In QC-lib the CNOT operator can act on two registers: the control register and the target register. These registers can represent substates of the quantum basis state (sub-registers). In this case, the CNOT gate inverts the state of the target (sub-)register if the control (sub-)register is in the state $|1\rangle_c$, where c is the number of qubits in the control register. For example, let a be a 1-qubit register in state $|0\rangle$, b a 2-qubit register in state $\alpha_0|01\rangle + \beta_0|10\rangle$ and c a 2-qubit register in state $\alpha_1|10\rangle + \beta_1|11\rangle$. Applying CNOT to target register b with control register c, the state of the entire quantum memory distributed to 8 processing nodes will be:

$$\begin{array}{l} P_{\mathbf{2}} : |10010\rangle, |11010\rangle \xrightarrow{\text{CNOT}} |10010\rangle, |11\mathbf{1}00\rangle \\ P_{\mathbf{4}} : |10100\rangle, |11100\rangle \xrightarrow{\text{CNOT}} |10100\rangle, |11\mathbf{0}10\rangle \end{array} \xrightarrow{\text{Comm}} \begin{array}{l} |10010\rangle, |11010\rangle \\ |10100\rangle, |11100\rangle \end{array}$$

Similar to the case of the general single qubit operator, each processing node communicates with at most one other process node. The index of the process involved in communication is determined by the qubits of the target (sub-)register that make part of the distribution key of the whole quantum memory.

The CPhase operator is another example of 2-qubit quantum gate implemented in QC-lib and allows for a controlled phase shift of the amplitudes. Its inputs are a rotation angle, θ, and a control qubit that acts in the same manner as in the CNOT case. The amplitudes of the basis states where the control qubit is $|0\rangle$ are left unchanged, an if the control qubit is in state $|1\rangle$, the phase of the amplitudes of the basis states are multiplied by $e^{i\theta}$. The matrix form of the CPhase operator is:

$$CPhase = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & e^{i\theta} \end{pmatrix}. \tag{6}$$

Operator CPhase can also act on (sub-)registers of the quantum memory and its parallel implementation is analogous to that of the CNOT operator. One important difference between the two implementations is that when applying the CPhase operator communication between processing nodes is not necessary as the action of this operator doesn't generate new terms, and only the amplitudes of the local terms are modified.

Measurement of Quantum States. In QC-lib, the measurement of a n-qubit quantum register is simulated in $O(2^n)$ time. Let $|\Psi\rangle = \sum_{j=0}^{2^n-1} \alpha_j |j\rangle$ be the state of a n-qubit quantum register. The measurement step is simulated in the following manner:

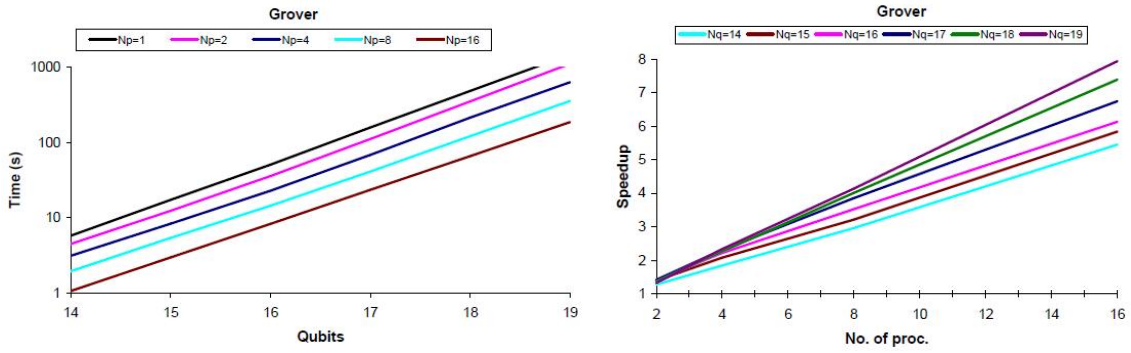
1. Rrandomly generate a number p, $0 \leq p < 1$,
2. Randomly generate a positive integer x, smaller than the number of terms with non-zero amplitude,
3. Determine an integer i, $0 \leq i < 2^n - 1$, such that $\sum_{j=x}^{i-1} |\alpha_j|^2 \leq p < \sum_{j=x}^{i} |\alpha_j|^2$.

Integer i is the representation of the measured state. After measurement, the state of the register becomes $|i\rangle$. Because the terms of the quantum register are distributed across processing nodes, the measurement operation requires communication between these nodes in order to correctly select the term i, but also to colapse the register in state $|i\rangle$. Thus, a master process is responsible with the random generation of numbers p and i and with computing the sum. Sincronization between processing nodes is achieved using MPI_Bcast operations such that the master process could receive the norm of the amplitude of a term j from the owning processing node. After selecting number i, all processing nodes know this number and can pass the state of the quantum register in $|i\rangle$.

4 Simulation of Quantum Search in GQCL

Many problems in classical computing can be reformulated to express the search of a unique element that satisfies a certain predefined condition [2]. If there is no additional information about the search condition, the best classical algorithm is a brute-force search, meaning that the elements are sequentially tested against the search condition. For a list of N elements, this algorithm executes an average of $N/2$ comparisons. By exploiting the advantages of quantum parallelism and interference of quantum states, Grover formulated a quantum algorithm that can find the searched element in an unstructured database in only $O(\sqrt{N})$ steps [6].

Grover's algorithm is based on the concept of amplitude amplification and its principle is to encode the elements in the data set as quantum states of a quantum register and to apply an operator, G, whose effect is to raise the probability that the system finds itself in the marked state (the state encoding the solution of the search problem). Because only unitary transformations are used to act upon the system, the probability conservation takes place. This allows that as the probability that the system finds itself in the desired state grows, the probability of all



(a) (b)

Figure 2: *(a)* Execution time for Grover's algorithm; (b) Speed-up for Grover's algorithm.

other (unmarked) states are correspondingly diminished. Applying Grover's operator G a certain number of times will determine the probability of the marked state to be very close to $\mathsf{1}$. In order to acheive this behavior of the quantum system, a Grover iteration first inverts the phase of the amplitude of the marked state and then inverts the phase of the amplitude of all states around the mean. The inversion of the solution state can be obtained using a so-called "black box" function (known as a quantum oracle) which must be able only to identify wheather a certain record is member of the solution set and thus the mechanism is very general.

After one application of Grover operator, the amplitude of the marked state grows with a factor of $\mathsf{O}(\frac{1}{\sqrt{\mathsf{N}}})$, while the amplitudes of the unmarked states lower correspondingly. To obtain $\mathsf{O}(1)$ probability for the solution state, Grover iteration must be applied $\mathsf{O}(\sqrt{\mathsf{N}})$ times. There is a finite probability that the search operation doesn't end in success, in which case, Grover's algorithm must be repeated.

The performance of our GQCL quantum simulator with respect to the quantum search problem is evaluated. In order to eloquently compare the running times for different problem sizes, we only measure the execution time for one application of Grover's algorithm. In figure 2 we present the results obtained for different problem sizes on various numbers of processing nodes. It can be observed from figure 2(a) that the run time grows with an average factor of about 2.8 for each additional qubit. This is due to the fact that each extra qubit represents a doubling of the problem size and that the number of applied Grover iterations grows with a factor of $\sqrt{2}$ for an additional qubit. Variation of the speedup with the number of processing nodes is presented in figure 2(b). For only 19 qubits we obtain a speed-up of 7.9 and the measurements reveal a growing trend of the speed-up with the increase of the problem size.

5 Conclusions

Classical sequential computers enforce drastic limitations over the quantum computation simulation process. Quantum computer simulators have become an attractive alternative for experimentation with quantum algorithms, but their purpose cannot be achieved without significant computing resources. A promising approach is represented by executing these simulators in Grid systems that can provide high performance resources. The quantum computer simulator described in this paper relies on parallel processing implemented in QC-lib. Besides the parallelization of the general single qubit operator, we also described the parallelization of the control gates (CNOT, CPhase) and of the measurement process. The efficient representation and partitioning of the quantum states using the distributed memory of a computer cluster allowed

very good speed-ups to be recorded at the execution of Grover's search algorithm.

Bibliography

[1] S. Caraiman, A. Archip, and V. Manta. A grid enabled quantum computer simulator. In *Proc. of SYNASC'09*. IEEE Computer Society, 2009.

[2] S. Caraiman and V. Manta. New applications of quantum algorithms to computer graphics: the Quantum Random Sample Consensus algorithm. In *Proc. of ACM CF '09*, pages 81–88, New York, NY, USA, 2009. ACM.

[3] R. Feynman. Simulating physics with computers. *Int. J. Theor. Phys.*, 21(6):467–488, 1982.

[4] I. Foster, C. Kesselman, and S. Tuecke. The anatomy of the grid: Enabling scalable virtual organizations. *Int. J. High Perform. Comput. Appl.*, 15(3):200–222, 2001.

[5] I. Glendinning and B. Omer. Parallelization of the QC-lib quantum computer simulator library. In *Proc. of PPAM 2003*, volume 3019 of *LNCS*, pages 461–468. Springer, 2004.

[6] L. Grover. A fast quantum mechanical algorithm for database search. In *Proc. of 28th ACM Annual STOC*, pages 212–219, 1996.

[7] J. Niwa, K. Matsumoto, and H. Imai. General-purpose parallel simulator for quantum computing. In *Proc. of UMC '02*, pages 230–251, London, UK, 2002. Springer-Verlag.

[8] B. Ömer. Simulation of quantum computers, 1996. http://tph.tuwien.ac.at/ oemer/doc/qcsim.ps.

[9] B. Ömer. *Structured Quantum Programming*. PhD thesis, TU Vienna, 2003.

[10] K. D. Raedt, K. Michielsen, H. D. Raedt, B. Trieuc, G. Arnoldc, M. Richterc, T. Lippertc, H. Watanabed, and N. Itoe. Massively parallel quantum computer simulator. *Computer Physics Communications*, 176(2):121–136, 2007.

[11] P. Shor. Algorithms for quantum computation: discrete logarithms and factoring. In *Proc. of SFCS '94*, pages 124–134. IEEE Computer Society, 1994.

[12] B. Sotomayor. The Globus Toolkit 4 programmer's tutorial, 2005. http://gdp.globus.org/gt4-tutorial/multiplehtml/index.html.

[13] F. Tabakin and B. Juliá-Díaz. QcMpi: A parallel environment for quantum computing. *Computer Physics Communications*, 180(6):948–964, 2009.

Int. J. of Computers, Communications & Control, ISSN 1841-9836, E-ISSN 1841-9844
Vol. V (2010), No. 5, pp. 642-653

Tense θ-valued Moisil propositional logic

C. Chiriţă

Carmen Chiriţă
University of Bucharest
Faculty of Mathematics and Computer Science
Romania, 010014 Bucharest, 4 Academiei
E-mail: stama@funinf.cs.unibuc.ro

Abstract: In this paper we study the tense θ-valued Moisil propositional calculus, a logical system obtained from the θ-valued Moisil propositional logic by adding two tense operators. The main result is a completeness theorem for tense θ-valued Moisil propositional logic. The proof of this theorem is based on the representation theorem of tense θ-valued Łukasiewicz-Moisil algebras, developed in a previous paper.
Keywords: Łukasiewicz-Moisil algebras, tense Moisil logic.

1 Introduction

The first contribution to the algebraic logic of finite-valued Łukasiewicz propositional calculus is Moisil's paper [18], where n-valued Łukasiewicz algebras (named today Łukasiewicz-Moisil algebras) were introduced. According to an example given by A. Rose (1957), for $n \geq 5$ the Łukasiewicz implication cannot be defined in an n-valued Łukasiewicz-Moisil algebra. Hence, Moisil discovered a new many-valued logical system (named today Moisil logic), whose algebraic models are n-valued Łukasiewicz-Moisil algebras.

In 1969, Moisil defined the θ-valued Łukasiewicz algebras, where θ is the order type of a bounded chain. These structures extend a part of the definition of n-valued Łukasiewicz algebras, but they differ from these by accepting many negation operations ([3], [10], [16], [23]). The logic corresponding to the θ-valued Łukasiewicz-Moisil algebras was developed by Boicescu [1] and Filipoiu [10] (see also [2]). This logical system is called the θ-valued Moisil propositional logic. The chrysippian endomorphisms of θ-valued Łukasiewicz-Moisil algebras are reflected in the syntax of the θ-valued Moisil propositional logic by chrysippian operations.

This paper is devoted to the tense θ-valued Moisil propositional calculus, a logical system obtained from the θ-valued Moisil propositional calculus by adding the tense operators G and H. The algebraic basis of this logic consists of tense θ-valued Łukasiewicz-Moisil algebras (tense LM_θ-algebras), algebraic structures studied in our paper [7]. We extend some of the results of [8], where a tense n-valued propositional logic was studied. The tense θ-valued Moisil propositional calculus unifies two logical systems: the classical tense logic and the θ-valued Moisil logic. The connection between these logics is realized by axioms that express the behaviour of the tense operators with respect to the chrysippian operations.

The paper is organized as follows.

In Section 2 we recall some definitions and basic facts on θ-valued Łukasiewicz-Moisil algebras and θ-valued Moisil logic, with emphasis on the connectives $\rightarrow_k$ and $\leftrightarrow_k$ and their algebraic counterparts. Section 3 deals with tense θ-valued Łukasiewicz-Moisil algebras (tense LM_θ-algebras), algebraic structures obtained from θ-valued Łukasiewicz-Moisil algebras by adding the two tense operators G and H. Section 4 contains the syntactical construction of the tense θ-valued Moisil propositional calculus. We establish some properties regarding the inferential structure of this logical system.

The Lindenbaum-Tarski algebra associated with the tense θ-valued Moisil propositional calculus is studied in Section 5. We obtain the structure of tense LM_θ-algebra. The syntactical properties of the tense θ-valued Moisil logic are reflected in this tense LM_θ-algebra, thus we use the algebraic framework in order to obtain results for the logical system.

In section 6 we define the interpretations of tense θ-valued Moisil propositional calculus and the k-tautologies of this logic. Our main result is the completeness theorem proved in this section (Theorem 26). Its proof uses the representation theorem of tense LM_θ-algebras applied to the Lindenbaum-Tarski algebra constructed in the previous section.

2 θ-valued Moisil logic and θ-valued Łukasiewicz-Moisil algebras

Let $(I, \leq)$ be a totally ordered set, with first and last element, denoted by 0 and 1 respectively, and of order type θ, through this paper.

We fix an element $k \in I$, through this paper.

In this section, we recall the θ-valued Moisil logic $\mathcal{M}_\theta$ described in [2]. The axiomatization of θ-valued Moisil propositional calculus uses the system of axioms of θ-valued calculus introduced by Boicescu [4] and Filipoiu [10]. The basic results are taken from Filipoiu [10](see also [2]). The alphabet of $\mathcal{M}_\theta$ has the following primitive symbols: an infinite set V of propositional variables; the logical connectives $\vee, \wedge, \varphi_i, \overline{\varphi}_i$ for all $i \in I$ and the parantheses (,). The set $\mathsf{Prop}(V)$ of propositions of $\mathcal{M}_\theta$ is defined by canonical induction. For each $i \in I$, we shall use the following abbreviations: $p \to_i q = \overline{\varphi}_i p \vee \varphi_i q$ and $p \leftrightarrow_i q = (p \to_i q) \wedge (q \to_i p)$. The θ-valued propositional calculus has the following *k-axioms*:

(2.1) $p \to_k (q \to_k p)$,

(2.2) $(p \to_k (q \to_k r)) \to_k ((p \to_k q) \to_k (p \to_k r))$,

(2.3) $p \wedge q \to_k p$,

(2.4) $p \wedge q \to_k q$,

(2.5) $(p \to_k q) \to_k ((p \to_k r) \to_k (p \to_k q \wedge r))$,

(2.6) $p \to_k p \vee q$,

(2.7) $q \to_k p \vee q$,

(2.8) $(p \to_k q) \to_k ((r \to_k q) \to_k (p \vee r \to_k q))$,

(2.9) $\varphi_i(p \wedge q) \leftrightarrow_k \varphi_i p \wedge \varphi_i q$, for every $i \in I$,

(2.10) $\overline{\varphi}_i(p \vee q) \leftrightarrow_k \overline{\varphi}_i p \wedge \overline{\varphi}_i q$, for every $i \in I$,

(2.11) $\varphi_j p \leftrightarrow_k \varphi_i \varphi_j p$, for every $i, j \in I$,

(2.12) $\varphi_j p \leftrightarrow_k \overline{\varphi}_i \overline{\varphi}_j p$, for every $i, j \in I$,

(2.13) $\overline{\varphi}_j p \leftrightarrow_k \varphi_i \overline{\varphi}_j p$, for every $i, j \in I$,

(2.14) $\overline{\varphi}_j p \leftrightarrow_k \overline{\varphi}_i \varphi_j p$, for every $i, j \in I$,

(2.15) $\varphi_i p \to_k \varphi_j p$, for every $i, j \in I$, $i \leq j$.

The notion of formal proof in $\mathcal{M}_\theta$ is defined in terms of the above k-axioms and the k-*modus ponens* inference rule: $\dfrac{p,\ p \to_k q}{q}$.
For briefness, we will say "modus ponens" (m.p) instead of "k-modus ponens" from now on.
We shall denote by $\vdash_k p$ that p is a *k-theorem*.
We remind some k-theorems of $\mathcal{M}_\theta$, which will be used in our proofs.

Proposition 1. *([2], p. 491, Example 3.12) The following propositions are* k*-theorems of* $\mathcal{M}_\theta$*:*

(2.16) $\vdash_k p \to_k p$,

(2.17) $\vdash_k p \leftrightarrow_k \varphi_k p$,

(2.18) $\vdash_k (\varphi_i p \vee \overline{\varphi}_i p)$, *for every* $i \in I$, $j \in I$,

(2.19) $\vdash_k (\varphi_j(p \vee q) \leftrightarrow_k \varphi_j p \vee \varphi_j q)$, $j \in I$,

(2.20) $\vdash_k (\overline{\varphi}_j(p \wedge q) \leftrightarrow_k \overline{\varphi}_j p \vee \overline{\varphi}_j q)$, $j \in I$,

(2.21) $\vdash_k ((p \to_k q) \to_k (\overline{\varphi}_k q \to_k \overline{\varphi}_k p))$,

(2.22) $\dfrac{p}{\varphi_j p}$, $j \geq k$,

(2.23) $\dfrac{\varphi_k p \to_k \varphi_k q}{p \to_k q}$.

Proposition 2. *The following propositions are* k*-theorems of* $\mathcal{M}_\theta$*:*

(2.24) $\vdash_k p \to_k (q \to_k (p \wedge q))$,

(2.25) $\vdash_k (p \wedge q \to_k r) \to_k (p \to_k (q \to_k r))$,

(2.26) $\vdash_k (p \to_k (q \to_k r)) \to_k ((p \wedge q) \to_k r)$,

(2.27) $\vdash_k (p \to_k q) \to_k ((q \to_k r) \to_k (p \to_k r))$,

(2.28) $\vdash_k (p \to_k q) \to ((r \to_k t) \to_k (p \wedge r \to_k q \wedge t))$.

Proof: *We shall establish only the* k*-theorems (2.24), (2.25) and (2.28).*

(2.24) We shall use (2.5), (2.16), (2.1), modus ponens and the Deduction Theorem (see [2], p. 495, Proposition 3.17).

$\{p, q\} \vdash_k (p \to_k p) \to_k ((p \to_k q) \to_k (p \to_k p \wedge q))$ *(2.5)*
$\{p, q\} \vdash_k p \to_k p$ *(2.16)*
$\{p, q\} \vdash_k (p \to_k q) \to_k (p \to_k p \wedge q))$ *(m.p)*
$\{p, q\} \vdash_k q \to_k (p \to_k q)$ *(2.1)*
$\{p, q\} \vdash_k q$
$\{p, q\} \vdash_k p \to_k q$ *(m.p)*
$\{p, q\} \vdash_k p \to_k (p \wedge q)$ *(m.p)*
$\{p, q\} \vdash_k p$
$\{p, q\} \vdash_k p \wedge q$ *(m.p)*
$\{p\} \vdash_k q \to_k (p \wedge q)$ *(Deduction Theorem)*
$\vdash_k p \to_k (q \to_k (p \wedge q))$ *(Deduction Theorem)*

(2.25) We shall apply (2.24), modus ponens and the Deduction Theorem.

$\{p \wedge q \to_k r, p, q\} \vdash_k p \to_k (q \to_k (p \wedge q))$ *(2.24)*
$\{p \wedge q \to_k r, p, q\} \vdash_k p$
$\{p \wedge q \to_k r, p, q\} \vdash_k q \to_k (p \wedge q)$ *(m.p)*
$\{p \wedge q \to_k r, p, q\} \vdash_k q$
$\{p \wedge q \to_k r, p, q\} \vdash_k p \wedge q$ *(m.p)*
$\{p \wedge q \to_k r, p, q\} \vdash_k p \wedge q \to_k r$
$\{p \wedge q \to_k r, p, q\} \vdash_k r$ *(m.p)*
$\{p \wedge q \to_k r, p\} \vdash_k q \to_k r$ *(Deduction Theorem)*
$\{p \wedge q \to_k r\} \vdash_k p \to_k (q \to_k r)$ *(Deduction Theorem)*
$\vdash_k (p \wedge q \to_k r) \to_k (p \to_k (q \to_k r))$ *(Deduction Theorem)*

(2.28) We shall use k*-axioms (2.3), (2.4), modus ponens,* k*-theorem (2.24) and the Deduction Theorem.*

$\{p \to_k q, r \to_k t, p \wedge r\} \vdash_k p \wedge r$
$\{p \to_k q, r \to_k t, p \wedge r\} \vdash_k p \wedge r \to_k p$ *(2.3)*
$\{p \to_k q, r \to_k t, p \wedge r\} \vdash_k p$ *(m.p)*
$\{p \to_k q, r \to_k t, p \wedge r\} \vdash_k p \to_k q$
$\{p \to_k q, r \to_k t, p \wedge r\} \vdash_k q$ *(m.p)*
$\{p \to_k q, r \to_k t, p \wedge r\} \vdash_k p \wedge r \to_k r$ *(2.4)*
$\{p \to_k q, r \to_k t, p \wedge r\} \vdash_k r$ *(m.p)*
$\{p \to_k q, r \to_k t, p \wedge r\} \vdash_k r \to_k t$
$\{p \to_k q, r \to_k t, p \wedge r\} \vdash_k t$ *(m.p)*
$\{p \to_k q, r \to_k t, p \wedge r\} \vdash_k q \to_k (t \to_k (q \wedge t))$ *(2.24)*
$\{p \to_k q, r \to_k t, p \wedge r\} \vdash_k t \to_k (q \wedge t)$ *(m.p)*
$\{p \to_k q, r \to_k t, p \wedge r\} \vdash_k q \wedge t$ *(m.p)*
Applying the Deduction Theorem three times we obtain that
$\vdash_k (p \to_k q) \to ((r \to_k t) \to_k (p \wedge r \to_k q \wedge t))$.

The rest of the proof is straightforward. □

The θ-valued Łukasiewicz-Moisil algebras constitute the algebraic counterpart of the θ-valued Moisil logic. The Lindenbaum-Tarski algebra of the θ-valued Moisil propositional calculus is an θ-valued Łukasiewicz-Moisil algebra (see [2], p. 500, Theorem 3.30).

We shall recall the definition of θ-valued Łukasiewicz-Moisil algebras.

Definition 3. *A* θ*-valued Łukasiewicz-Moisil algebra (*LM_θ*-algebra) is an algebra* $\mathcal{L} = (L, \wedge, \vee, \{\varphi_i\}_{i\in I}, \{\overline{\varphi}_i\}_{i\in I}, 0_L, 1_L)$ *of type* $(2, 2, \{1\}_{i\in I}, \{1\}_{i\in I}, 0, 0)$ *such that for all* $x, y \in L$,

(2.29) $(L, \wedge, \vee, 0_L, 1_L)$ *is a bounded distributive lattice,*

(2.30) φ_i *is a bounded distributive lattice endomorphism for all* $i \in I$,

(2.31) $\varphi_i x \wedge \overline{\varphi}_i x = 0_L$; $\varphi_i x \vee \overline{\varphi}_i x = 1_L$ *for all* $i \in I$,

(2.32) $\varphi_i \circ \varphi_j = \varphi_j$ *for all* $i, j \in I$,

(2.33) If $i \leq j$ *then* $\varphi_i \leq \varphi_j$ *for all* $i, j \in I$,

(2.34) If $\varphi_i x = \varphi_i y$ *for all* $i \in I$, *then* $x = y$ *(this is known as Moisil's determination principle).*

Let $\mathcal{L} = (L, \wedge, \vee, \{\varphi_i\}_{i\in I}, \{\overline{\varphi}_i\}_{i\in I}, 0_L, 1_L)$ be an LM_θ-algebra.
We say that $\mathcal{L}$ is *complete* if the lattice $(L, \wedge, \vee, 0_L, 1_L)$ is complete. $\mathcal{L}$ is *completely chrysippian* if for every $\{x_k\}_{k\in K}$ ($x_k \in L$ for all $k \in K$) such that $\bigwedge\limits_{k\in K} x_k$ and $\bigvee\limits_{k\in K} x_k$ exist, the following properties hold: $\varphi_i(\bigwedge\limits_{k\in K} x_k) = \bigwedge\limits_{k\in K} \varphi_i x_k$, $\varphi_i(\bigvee\limits_{k\in K} x_k) = \bigvee\limits_{k\in K} \varphi_i x_k$ $(\forall i \in I)$.

Example 4. *Let $\mathcal{B} = (B, \wedge, \vee, ^-, 0_B, 1_B)$ be a Boolean algebra.*
The set $D(B) = B^{[I]} = \{f | f : I \to B, i \leq j \Rightarrow f(i) \leq f(j)\}$ of all increasing functions from I to B can be made into a LM_θ-algebra $D(\mathcal{B}) = (D(B), \wedge, \vee, \{\varphi_i\}_{i\in I}, \{\overline{\varphi}_i\}_{i\in I}, 0_{D(B)}, 1_{D(B)})$ where $0_{D(B)}$, $1_{D(B)} : I \to B$ are defined by $0_{D(B)}(i) = 0_B$ and $1_{D(B)}(i) = 1_B$ for every $i \in I$, the operations of the lattice $(D(B), \wedge, \vee, 0_{D(B)}, 1_{D(B)})$ are defined pointwise (cf. [2], p.6, Example 1.10) and $(\varphi_i f)(j) = f(i)$, $(\overline{\varphi}_i f)(j) = (f(i))^-$ $(\forall j \in I)$ $(\forall i \in I)$.

Let $\mathcal{L} = (L, \wedge, \vee, \{\varphi_i\}_{i\in I}, \{\overline{\varphi}_i\}_{i\in I}, 0_L, 1_L)$ be an LM_θ-algebra. For each $j \in I$ we consider the binary operation $\to_j$ on $\mathcal{L}$ defined by (2.35) $a \to_j b = \bar{\varphi}_j a \vee \varphi_j b = (\varphi_j a \wedge \bar{\varphi}_j b)^-$ for all $a, b \in L$. This implication is associated to $\wedge$ (like for Boolean algebras), but like for Boolean algebras also, there exists the following implication: $a \rightsquigarrow_j b = \bar{\varphi}_j a \wedge \varphi_j b$, associated to $\vee$.
The notion of morphism of LM_θ-algebras is defined as usual ([2]). Of course, a morphism of LM_θ-algebras preserves the operation $\to_j$.

3 Tense θ-valued Łukasiewicz-Moisil algebras

In this section we shall recall some definitions and basic results on tense θ-valued Łukasiewicz-Moisil algebras from [7].

Definition 5. *A tense LM_θ-algebra is a triple $\mathcal{A}_t = (\mathcal{A}, G, H)$, where $\mathcal{A} = (A, \wedge, \vee, \{\varphi_i\}_{i\in I}, \{\overline{\varphi}_i\}_{i\in I}, 0_A, 1_A)$ is an LM_θ-algebra and G, $H : A \to A$ are two unary operations on A such that for all $x, y \in A$,*

(3.1) $G(1_A) = 1_A, H(1_A) = 1_A$,

(3.2) $G(x \wedge y) = G(x) \wedge G(y)$, $H(x \wedge y) = H(x) \wedge H(y)$,

(3.3) $G \circ \varphi_i = \varphi_i \circ G$, $H \circ \varphi_i = \varphi_i \circ H$, *for any* $i \in I$,

(3.4) $G(x) \vee y = 1_A$ *iff* $x \vee H(y) = 1_A$.

Definition 6. *Let $(\mathcal{A}, G, H)$ be a tense LM_θ-algebra. For any $i \in I$, let us consider the unary operations P_i, F_i defined by $P_i x = \overline{\varphi}_i H \overline{\varphi}_i x$ and $F_i x = \overline{\varphi}_i G \overline{\varphi}_i x$, for any $x \in A$.*

Proposition 7. *Let $\mathcal{A} = (A, \wedge, \vee, \{\varphi_i\}_{i\in I}, \{\overline{\varphi}_i\}_{i\in I}, 0_A, 1_A)$ be an LM_θ-algebra and G, H be two unary operations on A that satisfy conditions (3.1), (3.2) and (3.3). Then, the condition (3.4) is equivalent with (3.4′) $\varphi_i \leq G \circ P_i$ and $\varphi_i \leq H \circ F_i$ for all $i \in I$.*

Thus, if we replace in Definition 5 the axiom (3.4) with the condition (3.4′), we obtain an equivalent definition of tense LM_θ-algebra.

Proposition 8. *Let $\mathcal{A} = (A, \wedge, \vee, \{\varphi_i\}_{i\in I}, \{\overline{\varphi}_i\}_{i\in I}, 0_A, 1_A)$ be an LM_θ-algebra and G, H be two unary operations on A that satisfy conditions (3.1) and (3.3). Then, the condition (3.2) is equivalent to (3.2′) $G(a \to_k b) \leq G(a) \to_k G(b)$; $H(a \to_k b) \leq H(a) \to_k H(b)$ for all $k \in I$ where $\to_k$ is defined by (2.35).*

Thus, if in Definition 5 we replace the axiom (3.2) by (3.2'), we obtain an equivalent definition for tense LM_θ-algebra.

Definition 9. *A frame is a pair* (X, R)*, where* X *is a nonempty set and* R *is a binary relation on* X*.*

Let (X, R) be a frame and $\mathcal{L} = (L, \wedge, \vee, \{\varphi_i\}_{i\in I}, \{\bar{\varphi}_i\}_{i\in I}, 0_L, 1_L)$ be a complete and completely chrysippian LM_θ-algebra. L^X has a canonical structure of LM_θ-algebra. Let's us define for all $p \in L^X$ and $x \in X$: $G^*(p)(x) = \bigwedge\{p(y) | y \in X, xRy\}$, $H^*(p)(x) = \bigwedge\{p(y) | y \in X, yRx\}$.

Proposition 10. *For any frame* (X, R)*,* $(\mathcal{L}^X, G^*, H^*)$ *is a tense* LM_θ*-algebra.*

Let $(\mathcal{B}, G, H)$ be a tense Boolean algebra. We define on $D(B)$ the unary operations $D(G)$ and $D(H)$ by: $D(G)(f) = G \circ f$, $D(H)(f) = H \circ f$ for all $f \in D(B)$.

Lemma 11. *If* $(\mathcal{B}, G, H)$ *is a tense Boolean algebra then* $(D(\mathcal{B}), D(G), D(H))$ *is a tense* LM_θ*-algebra.*

Theorem 12. *(The representation theorem for tense* LM_θ*-algebras) For every tense* LM_θ*-algebra* $(\mathcal{A}, G, H)$ *there exist a frame* (X, R) *and an injective morphism of tense* LM_θ*-algebras* $\alpha : A \to (D(L_2))^X$*, where* $L_2 = \{0, 1\}$*, the standard Boolean algebra.*

4 Tense θ-valued Moisil logic (the syntax)

In this section we introduce the tense θ-valued Moisil propositional calculus $\mathcal{TM}_\theta$, a logical system obtained from the θ-valued propositional calculus (see [2]) by adding the two tense operators G and H. We define the notion of k-theorem and k-deduction then we establish some syntactical properties of $\mathcal{TM}_\theta$.
The alphabet of $\mathcal{TM}_\theta$ has the following primitive symbols: an infinite set V of propositional variables; the logical connectives $\vee, \wedge, \varphi_i, \overline{\varphi}_i$ for all $i \in I$; the tense operators G and H and parantheses $(,)$. The set E of propositions of $\mathcal{TM}_\theta$ is defined by canonical induction.

Definition 13. *We shall use the following abbreviations: for all* $\alpha, \beta \in E$ *and* $i \in I$*, we define* $\alpha \to_i \beta = \overline{\varphi}_i\alpha \vee \varphi_i\beta$*;* $\alpha \leftrightarrow_i \beta = (\alpha \to_i \beta) \wedge (\beta \to_i \alpha)$*;* $F_i\alpha = \overline{\varphi}_i G\overline{\varphi}_i\alpha$*;* $P_i\alpha = \overline{\varphi}_i H\overline{\varphi}_i\alpha$*.*

Definition 14. *We call a k-axiom of tense* θ*-valued Moisil propositional calculus a proposition of one of the following forms:*

(4.1) The k-axioms of θ*-valued Moisil propositional calculus ((2.1)-(2.15) in Section 2);*

(4.2) $G(\alpha \to_k \beta) \to_k (G\alpha \to_k G\beta)$*;* $H(\alpha \to_k \beta) \to_k (H\alpha \to_k H\beta)$*;*

(4.3) $G\varphi_i\alpha \leftrightarrow_k \varphi_i G\alpha$*;* $H\varphi_i\alpha \leftrightarrow_k \varphi_i H\alpha$*, for all* $i \in I$*;*

(4.4) $\varphi_i\alpha \to_k GP_i\alpha$*;* $\varphi_i\alpha \to_k HF_i\alpha$*, for all* $i \in I$*.*

The notion of formal *k-proof* in $\mathcal{TM}_\theta$ is defined in terms of the above axioms and the following inference rules: $\dfrac{\alpha, \alpha \to_k \beta}{\beta}$ (modus ponens); $\dfrac{\alpha}{G\alpha}\ \dfrac{\alpha}{H\alpha}$ (Temporal Generalizations)

Definition 15. *We say that a proposition* α *is a k-theorem of* $\mathcal{TM}_\theta$ *if there exists a k-proof of it. We will denote by* $\vdash_k \alpha$ *the fact that* α *is a k-theorem of* $\mathcal{TM}_\theta$*.*

Definition 16. *Let* $\Gamma \subseteq E$ *and* $\alpha \in E$*. We say that* α *is a k-deduction from* Γ *and write* $\Gamma \vdash_k \alpha$ *if there exist* $n \in \mathbb{N} = \{0, 1, 2, ...\}$ *and* $\alpha_1, ..., \alpha_n \in \Gamma$ *such that* $\vdash_k \bigwedge_{i=1}^{n} \alpha_i \to_k \alpha$*.*

We remark that the logical structure of $\mathcal{IM}_\theta$ (k-theorems and k-deduction) combines the logical stuctures of two logical systems: the θ-valued Moisil logic and tense classical logic. Further we shall prove some syntactical properties.

Lemma 17. *Let $\Gamma \subseteq E$ and $\alpha \in E$. Then $\Gamma \vdash_k \alpha$ iff there exist $n \in \mathbb{N}$ and $\alpha_1, ..., \alpha_n \in \Gamma$ such that $\vdash_k \alpha_1 \rightarrow_k (\alpha_2 \rightarrow_k ...(\alpha_n \rightarrow_k \alpha)...)$.*

Proof: *By Definition 16 and k-theorems (2.25) and (2.26).* □

Lemma 18. *Let $\Gamma \subseteq E$ and $\alpha \in E$. Then $\Gamma \vdash_k \alpha$ iff there exists $\Gamma' \subseteq \Gamma$, Γ' finite, such that $\Gamma' \vdash_k \alpha$.*

Proof: *By Definition 16 and Lemma 17.* □

Proposition 19. *Let $\Gamma, \Sigma \subseteq E$ and $\alpha, \beta \in E$. The following properties hold:*
(i) If $\vdash_k \alpha$ then $\Gamma \vdash_k \alpha$;
(ii) If $\Gamma \subseteq \Sigma$ and $\Gamma \vdash_k \alpha$ then $\Sigma \vdash_k \alpha$;
(iii) If $\alpha \in \Gamma$ then $\Gamma \vdash_k \alpha$;
(iv) $\{\alpha\} \vdash_k \beta$ iff $\vdash_k \alpha \rightarrow_k \beta$;
(v) If $\Gamma \vdash_k \alpha$ and $\{\alpha\} \vdash_k \beta$ then $\Gamma \vdash_k \beta$;
(vi) If $\Gamma \vdash_k \alpha$ and $\Gamma \vdash_k \alpha \rightarrow_k \beta$ then $\Gamma \vdash_k \beta$;
(vii) $\Gamma \vdash_k \alpha \wedge \beta$ iff $\Gamma \vdash_k \alpha$ and $\Gamma \vdash_k \beta$.

Proof: *(i) Using Definition 16 for $n = 0$. (ii) By applying Definition 16.*
(iii) Using k-theorem (2.16) and Definition 16.
(iv) We assume that $\vdash_k \alpha \rightarrow_k \beta$. Then, by Definition 16, we obtain that $\{\alpha\} \vdash_k \beta$. Conversely, if $\{\alpha\} \vdash_k \beta$ then there exists $n \in \mathbb{N}$ such that $\vdash_k (\underbrace{\alpha \wedge ... \wedge \alpha}_{n}) \rightarrow_k \beta$. By using k-axioms (2.4) and (2.5), we get that $\vdash_k (\underbrace{\alpha \wedge ... \wedge \alpha}_{n}) \leftrightarrow_k \alpha$, so $\vdash_k \alpha \rightarrow_k \beta$.
(v) We suppose that $\Gamma \vdash_k \alpha$ and $\{\alpha\} \vdash_k \beta$. Then there exist $n \in \mathbb{N}$ and $\alpha_1, .., \alpha_n \in \Gamma$ such that $\vdash_k \bigwedge_{i=1}^{n} \alpha_i \rightarrow_k \alpha$. Using (iv), it follows that $\vdash_k \alpha \rightarrow_k \beta$ and by applying k-theorem (2.27) and modus ponens, we obtain that $\vdash_k \bigwedge_{i=1}^{n} \alpha_i \rightarrow_k \beta$, so $\Gamma \vdash_k \beta$.
(vi) Let $\Gamma \vdash_k \alpha$ and $\Gamma \vdash_k \alpha \rightarrow_k \beta$. By applying Lemma 18, there exist Γ_1, $\Gamma_2 \subseteq \Gamma$ such that $\Gamma_1 \vdash_k \alpha$ and $\Gamma_2 \vdash_k \alpha \rightarrow_k \beta$. By (ii), it follows that $\Gamma_1 \cup \Gamma_2 \vdash_k \alpha$ and $\Gamma_1 \cup \Gamma_2 \vdash_k \alpha \rightarrow_k \beta$. If we consider $\Gamma_1 \cup \Gamma_2 = \{\gamma_1, ..., \gamma_n\}$, we obtain that $\vdash_k \bigwedge_{i=1}^{n} \gamma_i \rightarrow_k \alpha$ and $\vdash_k \bigwedge_{i=1}^{n} \gamma_i \rightarrow_k (\alpha \rightarrow_k \beta)$. By applying k-axiom (2.2) and modus ponens, we get that $\vdash_k \bigwedge_{i=1}^{n} \gamma_i \rightarrow_k \beta$, so $\Gamma \vdash_k \beta$.
(vii) We assume that $\Gamma \vdash_k \alpha \wedge \beta$. By using k-axioms (2.3) and (2.4) and applying (i) and (vi), we obtain that $\Gamma \vdash_k \alpha$ and $\Gamma \vdash_k \beta$. Conversely, we assume that $\Gamma \vdash_k \alpha$ and $\Gamma \vdash_k \beta$. By using k-theorem (2.24) and (i), we obtain that $\Gamma \vdash_k \alpha \rightarrow_k (\beta \rightarrow_k \alpha \wedge \beta)$. By applying twice (vi), we get $\Gamma \vdash_k \alpha \wedge \beta$. □

Theorem 20. *(The deduction theorem) Let $\Gamma \subseteq E$ and $\alpha, \beta \in E$. Then $\Gamma \cup \{\alpha\} \vdash_k \beta$ iff $\Gamma \vdash_k \alpha \rightarrow_k \beta$.*

Proof: *We assume that $\Gamma \cup \{\alpha\} \vdash_k \beta$. Then there exist $n \in \mathbb{N}$ and $\alpha_1, ..., \alpha_n \in \Gamma$ such that $\vdash_k (\bigwedge_{i=1}^{n} \alpha_i \wedge \alpha) \rightarrow_k \beta$. By applying k-theorem (2.25) and modus ponens, it follows that*

$\vdash_k \bigwedge_{i=1}^{n} \alpha_i \to_k (\alpha \to_k \beta)$. *Using Definition 16, we obtain that* $\Gamma \vdash_k \alpha \to_k \beta$. *Conversely, we suppose that* $\Gamma \vdash_k \alpha \to_k \beta$. *Thus, by Proposition 4.1 (ii), we get* $\Gamma \cup \{\alpha\} \vdash_k \alpha \to_k \beta$. *Also, by Proposition 4.1 (iii), we have that* $\Gamma \cup \{\alpha\} \vdash_k \alpha$, *hence by applying Proposition 4.1 (vi), it results that* $\Gamma \cup \{\alpha\} \vdash_k \beta$. □

Proposition 21. *In* $\mathcal{TM}_\theta$, *the following properties hold:*

(4.5) If $\vdash_k \alpha \leftrightarrow_k \beta$, *then* $\vdash_k G\alpha \leftrightarrow_k G\beta$,

(4.6) $\vdash_k G(\alpha \wedge \beta) \leftrightarrow_k (G\alpha \wedge G\beta)$.

Proof: (4.5) *By using* k*-axioms* (2.3), (2.4), k*-theorem* (2.24) *and modus ponens, we obtain that:* $\vdash_k \alpha \leftrightarrow_k \beta$ *iff* $\vdash_k \alpha \to_k \beta$ *and* $\vdash_k \beta \to_k \alpha$. *Applying the temporal generalization rule* G, *we get that* $\vdash_k G(\alpha \to_k \beta)$ *and* $\vdash_k G(\beta \to_k \alpha)$. *Then, by* k*-axiom* (4.2) *and modus ponens, it follows that* $\vdash_k G\alpha \to_k G\beta$ *and* $\vdash_k G\beta \to_k G\alpha$, *hence* $\vdash_k G\alpha \leftrightarrow_k G\beta$.

(4.6) *We shall prove that* $\vdash_k G(\alpha \wedge \beta) \to_k (G\alpha \wedge G\beta)$ *and* $\vdash_k (G\alpha \wedge G\beta) \to_k G(\alpha \wedge \beta)$. *By applying Proposition 21 (4.5) for* k*-axioms (2.3), (2.4), we obtain that* $\vdash_k G(\alpha \wedge \beta) \to_k G\alpha$ *and* $\vdash_k G(\alpha \wedge \beta) \to_k G\beta$. *Using* k*-axiom (2.5) and modus ponens, it results that* $\vdash_k G(\alpha \wedge \beta) \to_k (G\alpha \wedge G\beta)$. *By* k*-teorem (2.24) and the temporal generalization rule* G, *we obtain that* $\vdash_k G(\alpha \to_k (\beta \to_k \alpha \wedge \beta))$. *Applying* k*-axiom (4.2), modus ponens and* k*-theorem (2.27), it follows that* $\vdash_k G\alpha \to_k (G\beta \to_k G(\alpha \wedge \beta))$. *Using* k*-theorem (2.26) and modus ponens, we get that* $\vdash_k (G\alpha \wedge G\beta) \to_k G(\alpha \wedge \beta)$. *Thus* $\vdash_k G(\alpha \wedge \beta) \leftrightarrow_k (G\alpha \wedge G\beta)$. □

We remark that there exists a similar Proposition concerning H.

5 The k-Lindenbaum-Tarski algebra of tense θ-valued Moisil logic

In this section we shall prove that the k-Lindenbaum-Tarski algebra of $\mathcal{TM}_\theta$ is a tense θ-valued Łukasiewicz-Moisil algebra. Therefore, the tense θ-valued Łukasiewicz-Moisil algebras constitute the algebraic structures of $\mathcal{TM}_\theta$ and the properties of tense LM_θ-algebras reflect the syntactical properties of $\mathcal{TM}_\theta$.
We consider the binary relation $\sim_k$ on the set of all propositions E, defined by: $\alpha \sim_k \beta$ iff $\vdash_k \varphi_i\alpha \leftrightarrow_k \varphi_i\beta$ for all $i \in I$.

Lemma 22. $\sim_k$ *is an equivalence relation on* E.

For any proposition $\alpha \in E$, we denote by $[\alpha]_k$ the equivalence class of α. We can define the following operations on the set $E/_{\sim_k}$: $[\alpha]_k \vee [\beta]_k = [\alpha \vee \beta]_k$; $[\alpha]_k \wedge [\beta]_k = [\alpha \wedge \beta]_k$; $\varphi_i[\alpha]_k = [\varphi_i\alpha]_k$; $\overline{\varphi}_i[\alpha]_k = [\overline{\varphi}_i\alpha]_k$ for all $i \in I$; $G([\alpha]_k) = [G\alpha]_k$; $H([\alpha]_k) = [H\alpha]_k$; $0_k = [\overline{\varphi}_k\alpha]_k$, $1_k = [\varphi_k\alpha]_k$, where α is a k-theorem of $\mathcal{TM}_\theta$.

Proposition 23. $(E/_{\sim_k}, \wedge, \vee, \{\varphi_i\}_{i\in I}, \{\overline{\varphi}_i\}_{i\in I}, 0_k, 1_k, G, H)$, *the* k*-Lindenbaum-Tarski algebra of* $\mathcal{TM}_\theta$, *is a tense* LM_θ*-algebra.*

Proof: *By ([2], p.500, Theorem 3.30), we have that* $(E/_{\sim_k}, \wedge, \vee, \{\varphi_i\}_{i\in I}, \{\overline{\varphi}_i\}_{i\in I}, 0_k, 1_k)$ *is an* LM_θ*-algebra. What is left to prove is that the operations* G *and* H *are well defined and the conditions (3.1)-(3.4) are satisfied. Due to the symmetrical position of* G *and* H *we shall only include the proofs for* G. *Let* $\alpha, \beta \in E$ *such that* $\alpha \sim_k \beta$. *Thus,* $\vdash_k \varphi_i\alpha \leftrightarrow_k \varphi_i\beta$ *for all* $i \in I$. *Applying Proposition 21 (4.5), we obtain that* $\vdash_k G\varphi_i\alpha \leftrightarrow_k G\varphi_i\beta$ *for all* $i \in I$. *Using* k*-axiom (4.3), it follows that* $\vdash_k \varphi_i G\alpha \leftrightarrow_k \varphi_i G\beta$ *for all* $i \in I$, *so* $G\alpha \sim_k G\beta$.

(3.1) *We have to prove that* $G([\varphi_k\alpha]_k) = [\varphi_k\alpha]_k$ *i.e. by definition of* $\sim_k$ *that* $\vdash_k \varphi_i G\varphi_k\alpha \leftrightarrow_k \varphi_i\varphi_k\alpha$ *for every* α *such that* $\vdash_k \alpha$ *and for all* $i \in I$. *Let* $\alpha \in E$ *such that* $\vdash_k \alpha$ *and* $i \in I$. *By* k*-theorem (2.22), we obtain that* $\vdash_k \varphi_k\alpha$ *and by applying the temporal generalization rule G, we obtain that* $\vdash_k G\varphi_k\alpha$. *Using* k*-axiom (2.1) and modus ponens, it results that* $\vdash_k \varphi_k\alpha \rightarrow_k G\varphi_k\alpha$ *and* $\vdash_k G\varphi_k\alpha \rightarrow_k \varphi_k\alpha$. *Thus, we get that (i)* $\vdash_k \varphi_k\alpha \leftrightarrow_k G\varphi_k\alpha$. *Using* k*-axiom (2.11), we have that (ii)* $\vdash_k \varphi_i\varphi_k\alpha \leftrightarrow_k \varphi_k\alpha$ *and by using Proposition 21(4.5), we obtain that (iii)* $\vdash_k G\varphi_i\varphi_k\alpha \leftrightarrow_k G\varphi_k\alpha$. *Using* k*-axiom (4.3) and the conditions (i),(ii), (iii), it results that* $\vdash_k \varphi_i G\varphi_k\alpha \leftrightarrow_k \varphi_i\varphi_k\alpha$.

(3.2) *Let* $\alpha, \beta \in E$. *We must prove that* $G([\alpha]_k \wedge [\beta]_k) = G([\alpha]_k) \wedge G([\beta]_k)$ *i.e.* $G(\alpha \wedge \beta) \sim_k G\alpha \wedge G\beta$ *which is equivalent with* $\vdash_k \varphi_i G(\alpha \wedge \beta) \leftrightarrow_k \varphi_i(G\alpha \wedge G\beta)$ *for all* $i \in I$. *Let* $i \in I$. *By using Proposition 21(4.6) for* $\alpha = \varphi_i\alpha$ *and* $\beta = \varphi_i\beta$, *we obtain that (i)* $\vdash_k G(\varphi_i\alpha \wedge \wedge\varphi_i\beta) \leftrightarrow_k (G\varphi_i\alpha \wedge G\varphi_i\beta)$. *By using* k*-axiom (2.9) and Proposition 21(4.5), we get that (ii)* $\vdash_k G\varphi_i(\alpha \wedge \beta) \leftrightarrow_k G(\varphi_i\alpha \wedge \varphi_i\beta)$. *By conditions (i) and (ii), we obtain that (a)* $\vdash_k G\varphi_i(\alpha\wedge\beta) \leftrightarrow_k (G\varphi_i\alpha\wedge G\varphi_i\beta)$. *By* k*-axiom (4.3), we have:* $\vdash_k G\varphi_i\alpha \leftrightarrow_k \varphi_i G\alpha$ *and* $\vdash_k G\varphi_i\beta \leftrightarrow_k \varphi_i G\beta$. *Applying* k*-theorem (2.28), it follows that (b)* $\vdash_k (G\varphi_i\alpha \wedge G\varphi_i\beta) \leftrightarrow_k \leftrightarrow_k (\varphi_i G\alpha \wedge \varphi_i G\beta)$. *By conditions (a), (b) and* k*-axiom (4.3), we obtain that* $\vdash_k \varphi_i G(\alpha \wedge \wedge\beta) \leftrightarrow_k \varphi_i(G\alpha \wedge G\beta)$.

(3.3) *We have to prove that* $\vdash_k \varphi_j G\varphi_i\alpha \leftrightarrow_k \varphi_j\varphi_i G\alpha$ *for all* $i, j \in I$. *Let* $i, j \in I$. *By* k*-axiom (2.11), we obtain that (a)* $\vdash_k \varphi_j\varphi_i G\alpha \leftrightarrow_k \varphi_i G\alpha$. *Using* k*-axiom (4.3), we have that (b)* $\vdash_k \varphi_j G\varphi_i\alpha \leftrightarrow_k G\varphi_j\varphi_i\alpha$. *By* k*-axioms (2.11) and Proposition 21(4.5), it follows that (c)* $\vdash_k G\varphi_j\varphi_i\alpha \leftrightarrow_k G\varphi_i\alpha$. *By (a), (b), (c) and k-axiom (4.3), we get that* $\vdash_k \varphi_j G\varphi_i\alpha \leftrightarrow_k \varphi_j\varphi_i G\alpha$.

(3.4) *Since by Proposition 7, the condition (3.4) is equivalent with (3.4'), we shall prove that* $[\varphi_i\alpha]_k \leq [GP_i\alpha]_k$ *for all* $i \in I$, *i.e.* $\vdash_k \varphi_j\varphi_i\alpha \rightarrow_k \varphi_j GP_i\alpha$ *for all* $i, j \in I$. *Let* $i, j \in I$. *By* k*-axiom (2.13), we have that* $\vdash_k P_i\alpha \leftrightarrow_k \varphi_j P_i\alpha$. *Applying Proposition 21 (4.5), it follows that* $\vdash_k GP_i\alpha \leftrightarrow_k G\varphi_j P_i\alpha$. *Using* k*-axiom (4.3), it results that (1)* $\vdash_k GP_i\alpha \leftrightarrow_k \varphi_j GP_i\alpha$. *Also, by* k*-axiom (2.11), we have that (2)* $\vdash_k \varphi_i\alpha \leftrightarrow_k \varphi_j\varphi_i\alpha$. *By (1), (2) and* k*-axiom (4.4), we get that* $\vdash_k \varphi_j\varphi_i\alpha \rightarrow_k \varphi_j GP_i\alpha$.

□

6 Semantics and completeness theorem of tense θ-valued Moisil logic

This section concernes with the semantics of $\mathcal{TM}_\theta$, which combines the properties of Kripke semantics for $\mathcal{T}$ and the algebraic semantics for $\mathcal{M}_\theta$. We establish a completeness theorem for $\mathcal{TM}_\theta$ by using the representation theorem of tense θ-valued Łukasiewicz-Moisil algebras [7].

Definition 24. *Let* (X, R) *be a frame. A valuation of* $\mathcal{TM}_\theta$ *is a function* $\nu : E \times X \rightarrow L_2^{[I]}$ *such that for all* $\alpha, \beta \in E$ *and* $x \in X$, *the following equalities hold:* $\nu(\alpha \rightarrow_k \beta, x) = \nu(\alpha, x) \rightarrow_k \nu(\beta, x)$; $\nu(\alpha \wedge \beta, x) = \nu(\alpha, x) \wedge \nu(\beta, x)$; $\nu(\alpha \vee \beta, x) = \nu(\alpha, x) \vee \nu(\beta, x)$; $\nu(\varphi_i\alpha, x) = \varphi_i\nu(\alpha, x)$ *for any* $i \in I$; $\nu(\overline{\varphi}_i\alpha, x) = \overline{\varphi}_i\nu(\alpha, x)$ *for any* $i \in I$; $\nu(Gp, x) = \bigwedge\{\nu(p, y) | xRy\}$; $\nu(Hp, x) = \bigwedge\{\nu(p, y) | yRx\}$.

The first five conditions of the previous definition reflect "the many-valued past" of $\mathcal{TM}_\theta$ (see [2], p.487) and the last two conditions correspond to "the tense past" of $\mathcal{TM}_\theta$ (see [5], p.93).

Definition 25. *We say that a proposition* α *is a k-tautology and we write* $\models_k \alpha$ *if for every frame* (X, R), *for any valuation* $\nu : E \times X \rightarrow L_2^{[I]}$ *and for all* $x \in X$, *we have* $\nu(\alpha, x)(k) = 1$.

The following result establishes the equivalence between the k-theorems and the k-tautologies of $\mathcal{TM}_\theta$. The proof of the main implication is based on the representation theorem for tense θ-valued Łukasiewicz-Moisil algebras (Theorem 12).

Theorem 26. *(Completeness theorem). For any proposition α of $\mathcal{TM}_\theta$, we have: $\vdash_k \alpha$ iff $\models_k \alpha$.*

Proof: ($\Rightarrow$). *We shall prove by induction on the definition of $\vdash_k \alpha$ that for every frame (X, R) and for any valuation $\nu : E \times X \to L_2^{[I]}$, we have $\nu(\alpha, x)(k) = 1$, for all $x \in X$.*
Let (X, R) be a frame, $\nu : E \times X \to L_2^{[I]}$ be a valuation and $x \in X$.
• *We suppose that α is a k-axiom.*

(a) Let α be $G(p \to_k q) \to_k (Gp \to_k Gq)$ with $p, q \in E$. It is known that $a \to_k (b \to_k c) = (a \wedge b) \to_k c$ ([7], p.6, Proposition 2.1 (l)). We have: $\nu(\alpha, x)(k) = \nu(G(p \to_k q) \to_k (Gp \to_k Gq), x)(k) = [\nu(G(p \to_k q), x) \to_k (\nu(Gp, x) \to_k \nu(Gq, x))](k) = [(\nu(G(p \to_k q), x) \wedge \nu(Gp, x)) \to_k \nu(Gq, x)](k) = [\bigwedge_{xRy}((\nu(p, y) \to_k \nu(q, y)) \wedge \nu(p, y)) \to_k \to_k \bigwedge_{xRy} \nu(q, y)](k) = [\overline{\varphi}_k \bigwedge_{xRy}((\nu(p, y) \to_k \nu(q, y)) \wedge \nu(p, y)) \vee \varphi_k \bigwedge_{xRy} \nu(q, y)](k) = [(\bigwedge_{xRy}(\nu(p, y) \to_k \nu(q, y)) \wedge \nu(p, y))(k)]^- \vee (\bigwedge_{xRy} \nu(q, y))(k) = [\bigwedge_{xRy}((\nu(p, y)(k))^- \vee \nu(q, y)(k)) \wedge \wedge \nu(p, y)(k)]^- \vee (\bigwedge_{xRy} \nu(q, y)(k)) = [\bigwedge_{xRy}(\nu(q, y)(k) \wedge \nu(p, y)(k))]^- \vee (\bigwedge_{xRy} \nu(q, y)(k))$.
Since $\nu(q, y)(k), \nu(p, y)(k) \in L_2$ and $\nu(q, y)(k) \wedge \nu(p, y)(k) \leq \nu(q, y)(k)$, we obtain that $\bigwedge_{xRy}(\nu(q, y)(k) \wedge \nu(p, y)(k)) \leq \bigwedge_{xRy} \nu(q, y)(k)$. Since in a Boolean algebra we have $a \leq b$ iff $\bar{a} \vee b = 1$, we get that $[\bigwedge_{xRy}(\nu(q, y)(k) \wedge \nu(p, y)(k))]^- \vee (\bigwedge_{xRy} \nu(q, y)(k)) = 1$.

(b) Let α be $G\varphi_i p \leftrightarrow_k \varphi_i Gp$ with $p \in E$ and $i \in I$. Then $\nu(\alpha, x)(k) = \nu(G\varphi_i p \leftrightarrow_k \varphi_i Gp, x)(k) = \nu((G\varphi_i p \to_k \varphi_i Gp) \wedge (\varphi_i Gp \to_k G\varphi_i p), x)(k) = [(\nu(G\varphi_i p, x) \to_k \nu(\varphi_i Gp, x)) \wedge (\nu(\varphi_i Gp, x) \to_k \nu(G\varphi_i p, x))](k)$. *Since $L_2^{[I]}$ is complete and completely chrysippian, it follows that $\nu(G\varphi_i p, x) = \bigwedge_{xRy} \varphi_i \nu(p, y) = \varphi_i(\bigwedge_{xRy} \nu(p, y)) = \nu(\varphi_i Gp, x)$. We know that $a \to_k a = 1$ ([7], p.6, Proposition 2.1 (f)), hence $\nu(\alpha, x)(k) = 1$.*

(c) Let α be $\varphi_i p \to_k GP_i p$ with $i \in I$. We have: $\nu(\alpha, x)(k) = \nu(\varphi_i p \to_k GP_i p, x)(k) = (\nu(\varphi_i p, x) \to_k \nu(GP_i p, x))(k) = (\varphi_i \nu(p, x) \to_k \bigwedge_{xRy} \nu(P_i p, y))(k) = (\varphi_i \nu(p, x) \to_k \bigwedge_{xRy} \bigvee_{zRy} \varphi_i \nu(p, z))(k) = \overline{\varphi}_k(\varphi_i \nu(p, x))(k) \vee \varphi_k(\bigwedge_{xRy} \bigvee_{zRy} \varphi_i \nu(p, z))(k) = [\nu(p, x)(i)]^- \vee \bigwedge_{xRy} \bigvee_{zRy} \nu(p, z)(i)$. *Let $y \in X$ such that xRy. Then $\nu(p, x)(i) \leq \bigvee_{zRy} \nu(p, z)(i)$, hence $\nu(p, x)(i) \leq \bigwedge_{xRy} \bigvee_{zRy} \nu(p, z)(i)$. We obtain that $[\nu(p, x)(i)]^- \vee \bigwedge_{xRy} \bigvee_{zRy} \nu(p, z)(i) = 1$.*

• *We assume that α was obtained by applying the modus ponens rule. We have that $\nu(\beta, x)(k) = 1$ and $\nu(\beta \to_k \alpha, x)(k) = 1$. But* $\nu(\beta \to_k \alpha, x)(k) = (\nu(\beta, x) \to_k \nu(\alpha, x))(k) = (\overline{\varphi}_k \nu(\beta, x) \vee \varphi_k \nu(\alpha, x))(k) = \overline{\varphi}_k(\nu(\beta, x))(k) \vee \varphi_k(\nu(\alpha, x))(k) = [\nu(\beta, x)(k)]^- \vee \nu(\alpha, x)(k)$.
We deduce that $\nu(\alpha, x)(k) = 1$.

• *We suppose that* $\alpha = G\beta$ *such that* $\vdash_k \beta$. *We have that* $\nu(\beta, x)(k) = 1$, *for every* $x \in X$. *Then* $\nu(G\beta, x)(k) = (\bigwedge_{xRy} \nu(\beta, y))(k) = \bigwedge_{xRy} \nu(\beta, y)(k) = 1$.

($\Leftarrow$). *We shall prove that if* $\nvdash_k \alpha$ *then* $\not\models_k \alpha$. *Assume that* $\nvdash_k \alpha$, *so* $[\alpha]_k \neq 1_k$. *By using Proposition 23, we have that the* k*-Lindenbaum-Tarski algebra* $(E/_{\sim_k}, G, H)$ *of* $\mathcal{TM}_\theta$ *is a tense* LM_θ*-algebra. Applying the representation theorem for tense* LM_θ*-algebras (Theorem 12), there exist a frame* (X, R) *and an injective morphism of tense* LM_θ*-algebras* $d : (E/_{\sim_k}, G, H) \to \to (D(L_2)^X, G^*, H^*)$. *Let us consider the function* $\nu : E \times X \to L_2^{[I]}$ *defined by* $\nu(\alpha, x) = d([\alpha]_k)(x)$, *for all* $\alpha \in E$ *and* $x \in X$. *It is straightforward to prove that* ν *is a valuation. Since* d *is injective and* $[\alpha]_k \neq 1_k$, *we obtain that* $d([\alpha]_k) \neq 1_{D(L_2)^X}$, *hence there exists* $a \in X$ *such that* $\nu(\alpha, a) = d([\alpha]_k)(a) \neq 1_{D(L_2)}$. *Thus* α *is not a* k*-tautology.* □

7 Concluding Remarks

The tense θ-valued Moisil propositional calculus $\mathcal{TM}_\theta$ can be viewed as a common generalization of the θ-valued Moisil propositional logic $\mathcal{M}_\theta$ and the classical tense logic $\mathcal{T}$.
$\mathcal{TM}_\theta$ combines the logical structures of these logical systems and its semantic is inspired from the semantics of $\mathcal{T}$ and $\mathcal{M}_\theta$. The main result of this paper is a completeness theorem for $\mathcal{TM}_\theta$. Its proof is derived from the representation theorem of tense θ-valued Łukasiewicz-Moisil algebras [7].

An open problem is to obtain a proof of the representation theorem for tense θ-valued Łukasiewicz-Moisil algebras by using Theorem 26.

The next step in the study of tense aspects of Moisil logic is to define the tense θ-valued predicate logic (the syntax and the semantic) and the algebras corresponding of this logic (polyadic tense θ-valued Łukasiewicz-Moisil algebras). We hope to prove a completeness theorem for tense θ-valued Moisil predicate logic and a representation theorem for the corresponding algebras. The tense logics corresponding to the LM_θ-algebras with negations [16] will be the subject of another paper.

Bibliography

[1] V. Boicescu, *Sur les systèmes déductifs dans la logique* θ*-valente*, Publ. Dép. Math. Lyon, 8, 123-133, 1971.

[2] V. Boicescu, A. Filipoiu, G. Georgescu and S. Rudeanu, *Łukasiewicz-Moisil algebras*, North-Holland, 1991.

[3] V. Boicescu, *Contributions to the study of Łukasiewicz-Moisil algebras* (Romanian), Ph.D. Thesis, University of Bucharest, 1984.

[4] V. Boicescu, *Sur une logique polyvalente*, Rev. Roumaine Sci. Soc., sér. Philos. et Logique, 17, 393-405, 1973b.

[5] J. P. Burgess. Basic tense logic. In: Dov Gabbay and F. Guenthner, Eds., *Handbook of philosophical logic*, chapter II.2, Reidel, 89-134, 1984.

[6] R. Cignoli, I.M.L. D'Ottaviano and D. Mundici, *Algebraic Foundations of Many-valued Reasoning*, Kluwer, 2000.

[7] C. Chiriţă, *Tense* θ*-valued Łukasiewicz -Moisil algebras*, to appear in Journal of Multiple-Valued Logic and Soft Computing.

[8] D. Diaconescu, G. Georgescu, *Tense operators on MV-algebras and Łukasiewicz-Moisil algebras*, Fundamenta Informaticae XX: 1-30, 2007.

[9] I. Dziţac, L. Andrei, *65 Years from Birth of Prof. Gheorghe S. Nadiu (1941-1998)*, International Journal of Computers, Communications & Control Vol. I, No. 3, pp. 93-98, 2006.

[10] A. Filipoiu, θ-*valued Łukasiewicz-Moisil algebras and logics* (Romanian), Ph.D.Thesis, University of Bucharest, 1981.

[11] G. Georgescu, A. Iorgulescu, S. Rudeanu, *Grigore C. Moisil (1906-1973) and his School in Algebraic Logic*, Int. Journal of Computers, Communications & Control, Vol. I, No. 1, pp. 81-99, 2006.

[12] G. Georgescu, A. Iorgulescu, S. Rudeanu, *Some Romanian researches in algebra of logic*, In: Grigore C.Moisil and his followers, Editura Academiei Romane, 86-120, 2007.

[13] G. Georgescu, A. Iorgulescu, I. Leuştean, *Monadic and closure MV-algebras*, Multiple-Valued Logic, 3, 235-257, 1998.

[14] R. Goldblatt, *Logics of Time and Computation*, CSLI Lecture Notes No. 7, 1992.

[15] P. Hájek, *Metamathematics of fuzzy logic*, Kluwer Acad.Publ., Dordrecht, 1998

[16] A. Iorgulescu, $1 + \theta$ -*valued Łukasiewicz-Moisil algebras with negation* (Romanian). Ph.D. Thesis, University of Bucharest, 1984.

[17] J. Łukasiewicz, *On three-valued logic*, Ruch Filozoficzny (Polish), 5, 60-171, 1920.

[18] Gr. C. Moisil, *Recherches sur les logiques non-chrysippiennes*, Ann. Sci. Univ. Jassy, 26, 431-466, 1940.

[19] Gr. C. Moisil, *Notes sur les logiques non-chrysippiennes*, Ann. Sci. Univ. Jassy, 27, 86-98, 1941.

[20] Gr. C. Moisil, *Logique modale*, Disquis. Math. Phys, 2, 1942.

[21] Gr. C. Moisil, *Łukasiewiczian algebras*, Computing Center, University of Bucharest (preprint), 311-324, 1968.

[22] Gr. C. Moisil, *Essais sur les logiques non-chrysippiennes*, Ed. Academiei, Bucharest, 1972.

[23] Gh. S. Nadiu, *Cercetări asupra logicilor necryssipiene/Research about Necryssipiene Logics*, 1972 (PhD Thesis, supervisor Grigore C. Moisil).

[24] H. Rasiowa, *An algebraic approach to non-classical logics*, North-Holland Publ., Amsterdam, Polish Scientific Publ., Warszawa, 1974.

Int. J. of Computers, Communications & Control, ISSN 1841-9836, E-ISSN 1841-9844
Vol. V (2010), No. 5, pp. 654-663

Driving Style Analysis Using Data Mining Techniques

Z. Constantinescu, C. Marinoiu, M. Vladoiu

Zoran Constantinescu
ZealSoft Ltd., Bucharest, Romania
E-mail: zoran@zealsoft.ro

Cristian Marinoiu, Monica Vladoiu
PG University of Ploiesti, Romania
E-mail: cmarinoiu@upg-ploiesti.ro, monica@unde.ro

Abstract: This paper investigates the modeling of the personal driving style of various vehicle drivers based on several driving parameters. The purpose of such an endeavor is to classify the drivers according to their risk-proneness within the larger context of increasing traffic safety, which is a major concern worldwide. This information is valuable especially for those involved in fleet management and it can be used to improve and to make safer the driving style of various individuals who serve within that fleet. Equally important, such information could help any driver to see the danger within his or her driving style. Cluster and principal component analyses from exploratory statistics have been used to identify and explain drivers grouping according to their driving behavior. The driving parameters (behavioral indices) are collected from urban traffic by an in-house developed GPS-based device and sent to a data server for analysis.
Keywords: driving style, driving parameters, real time vehicle tracking, data mining.

1 Introduction and Related Work

Our society is changing at an amazing rate and there is no domain of human activity that is not affected by this process. Nowadays, people are overwhelmed with information and are under continuous pressure of being "in time" with some processing of that information. Consequently, everything happens or goes faster and faster each and every day. If we imagine having a look from above at the life pace in the 19th Century and if we compare this image with the current view, the difference is striking. This is also true for people trying to reach their destinations by using motorized vehicles that rush in various directions showing a throng-like view. In the last decades the number of such means of transportation has increased dramatically, their performances have improved at a fantastic rate and therefore, traffic conditions have worsened, especially in major cities. A direct effect of these changes can be seen on the driving behavior of city drivers, which becomes increasingly aggressive and incident-prone, reducing therefore the traffic safety. In this larger context, there is a major interest in categorizing the driving style of city drivers based on their driving behavior, which can be abstracted by means of various driving parameters.

In spite of this interest, the undertaking of extracting "standard behaviors from raw data of real human drivers has not yet been tackled and will be an area of future research" [1]. Even in cases where there have been made field tests that involved real drivers in real world experiences, "the driver's performance in terms of driving style was defined in each test through the subjective judgment of experts present at test" corroborated with fuel consumption [3]. The need for an objective method to understand daily driving behavior which derives from the driving style is emphasized in many works [1, 2, 4, 5]. There are some works that tries to determine the

driving style, seen as "the attitude, orientation and way of thinking for daily driving", based on questionnaires' surveys [4,5]. More recent works use a virtual driving simulator to collect realistic driving data from human drivers and to model human driving behavior [18], or classify driving style by combining objective rank method with recurrent learning based on Elman's type neural network [20]. There are also related works on modelling traffic flow, driving course decisions, or drivers behavior in emergency situations [19]. However, very few studies explore the modeling of personal driving style of various vehicle drivers based on several driving parameters, especially in urban traffic [7–9].

This work is about modeling the personal driving style of various vehicle drivers based on several driving parameters (behavioral indices). The purpose of such an endeavor is to classify the drivers according to their risk-proneness within the larger context of increasing traffic safety, which is a major concern worldwide. This information is valuable especially for those involved in fleet management and it can be used to improve and to make safer the driving style of various individuals who serve within that fleet. Equally important, such information could help any driver to see the danger within his or her driving style.

The paper is organized as follows: section 2 describes shortly the Gipix system and the GPS-based tracking device used for collecting the raw data, section 3 presents the driving parameters extracted from the raw data from the device, section 4 details the used methods, the analysis of the data and the interpretation of the results, and section 5 gives some concluding remarks and possible future work.

2 The Gipix System for Vehicle Tracking

Gipix is a system for real-time vehicle tracking, which offers very accurate positioning that is based both on state-of-the-arte GPS technology and GSM/GPRS data transmission. The system can be used both for individual vehicles and fleets. The core of the system consists of a data server that processes the maps for the main Romanian cities and roads, the monitored vehicles, the tracks for those vehicles, the drivers, critical events, predefined tracks, specific reports etc. Gipix collects the data of interest by using a GPS-based device, called Gipix-102B that must be installed on each monitored vehicle. This device automatically transmits the vehicle's positions and signals various critical events, both to the system and to the interested users. The Gipix System has been developed in-house to overcome some of the limitations of the commercially available GPS-based solutions [10].

The main advantages of our solution are one-second acquisition interval, high sensibility (the antenna works in difficult conditions), integrability with other applications, adaptability to users' needs, local storage for aproximately 100 hours of data for areas which are not GSM covered and positioning without GPS signal if the antenna fails, based on the GSM cells. From all these benefits, the most important for the current work has been the one-second acquisition interval, which is crucial for the success of the statistical methods that we have used.

A screenshot of the Gipix System is depicted in Figure 1, in which the cars of some of the drivers who have been involved in the analysis can be seen. In this screen is displayed also a sample track (the purple curved line). An upside down drop-like cursor can be moved along the track path and the given information will be tailored accordingly. A small info box, which contains data about instantaneous speed (v), altitude(h), acceleration (a), GPS error (e) can be seen on the right side of the screen. At the lower part of this screen the speed-acceleration plot is available for this track (red for speed and blue for acceleration). Several options can be selected from the left upper main menu: positioning, tracks, routes, personal or general points of interest, settings, configuration of users, car devices, vehicles and so on [10].

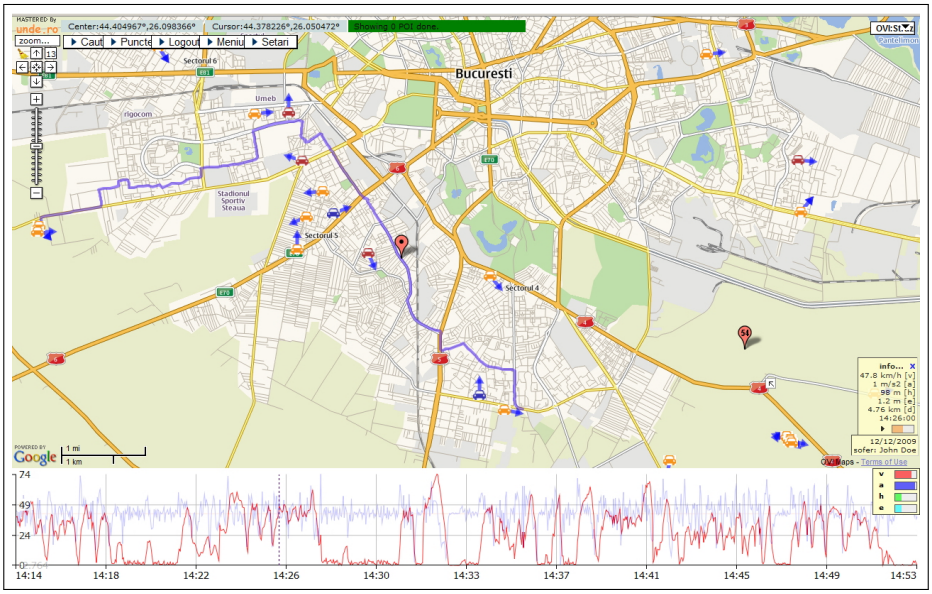


Figure 1: Gipix tracking system - driver's track analysis

3 Raw Data and the Driving Parameters

The raw data that has been considered for analysis is obtained from the tracking device in real time, by using GPRS and Internet as communication media. All the data is collected in the central server, where it is analyzed. It consists of GPS positions, time and speed values. Data is sampled at 1 second interval. From the speed values we calculate longitudinal acceleration at each time step (using numerical derivation), as well as the mechanical work (as the energy required to increase the speed over the time). A sample speed and acceleration diagram for one random track and driver, and for a small time interval is shown in Figure 2. Values for speed and acceleration are measured in [km/h] and [km/hs], respectively.

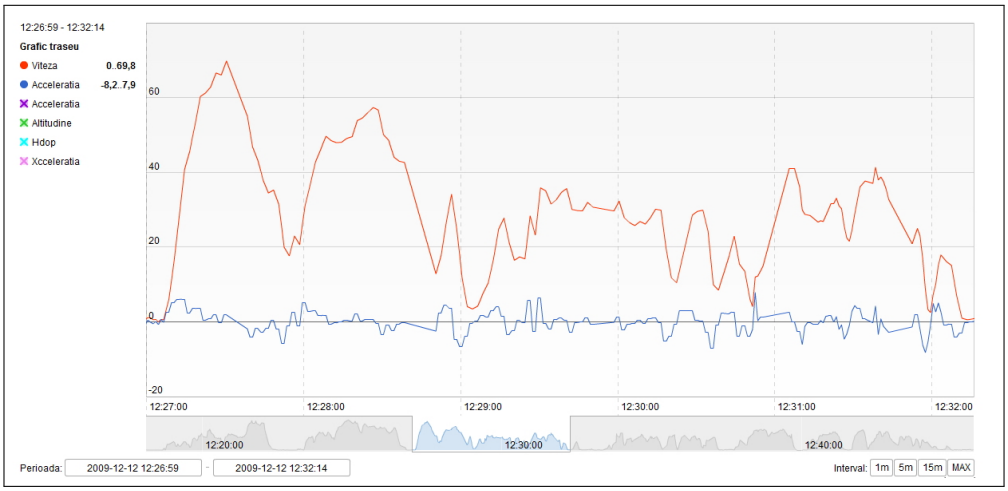


Figure 2: Speed and acceleration plot over time

For the statistical analysis we have been using the following driving parameters (extracted from the raw data):

- *speed over 60 km/h*: percent of time (V_{60}) - represents the percentage of time, from the

total time the vehicle is moving (thus excluding all stops of the vehicle), in which the speed is larger than 60 km/h - considered as the speed limit to be fined;

- *speed*: mean value (V_{mn}) and standard deviation (V_{sd}) - statistical values for the driving speed, the vehicle is also considered as moving;

- *acceleration*: standard deviation (A_{sd}) - statistical value for the acceleration, the vehicle is considered as moving as well;

- *positive acceleration*: mean value ($A+_{mn}$) and standard deviation ($A+_{sd}$) - statistical values for all positive accelerations, when the vehicle is considered to increase its speed; this is an indirect measure of the acceleration pedal position;

- *braking*: mean value (Br_{mn}) and standard deviation (Br_{sd}) - statistical values for negative accelerations, when the vehicle is considered to decrease its speed, and ignoring free decelerations (decrease in speed without breaking); because we don't have any indication of the actual brake pedal position, this is done by setting a threshold for the negative acceleration, above which all values are considered to be free decelerations;

- *mechanical work*: (W) - this is calculated as the sum, over the time, of all positive kinetic energy values required to increase the vehicle speed.

Data is collected for a total of 23 different drivers, with two additional controlled test drives. For each driver we collected several tracks (an average of 9 tracks per driver), over a short period of time (2-5 working days), and in similar conditions (same city: Bucharest). The total number of tracks analyzed is 200. The additional two test drives are done with extreme driving styles: a very aggressive one (D91) and a slow, non-aggressive, more economical one (D94). The resulting data is presented in Table 1.

sample	driver	V_{60}	V_{mn}	V_{sd}	A_{sd}	$A+_{mn}$	$A+_{sd}$	Br_{mn}	Br_{sd}	W
		%	km/h	km/hs	km/h	km/hs	km/hs	km/hs	km/hs	xJ
1	D1	5.5	25.6	17.5	2.45	2.23	1.46	3.07	1.57	58.42
2	D2	6.2	31.5	17.5	3.24	2.67	1.81	3.93	2.07	81.08
3	D3	14.4	34.7	19.5	2.78	2.30	1.59	3.89	2.23	64.43
4	D4	10.0	32.8	19.4	3.08	2.65	1.72	3.72	2.06	76.67
5	D5	6.4	29.1	19.1	2.81	2.48	1.68	3.34	1.90	70.55
6	D6	0.6	24.2	14.9	3.20	2.64	1.72	4.02	2.30	78.99
7	D7	8.0	27.9	18.3	3.32	2.72	1.91	3.89	2.15	82.17
8	D8	7.4	27.5	18.6	2.89	2.61	1.70	3.31	1.71	69.58
9	D9	2.8	24.7	17.6	2.99	2.68	1.75	3.36	2.03	74.64
10	D10	6.1	26.9	14.1	2.78	2.49	1.70	3.33	1.61	63.82
11	D11	4.6	27.9	17.5	3.27	2.67	1.71	4.18	2.33	76.33
12	D12	3.6	26.0	17.9	2.88	2.48	1.76	3.45	1.82	74.65
13	D13	6.3	30.0	18.5	2.74	2.39	1.61	3.66	1.87	64.96
14	D14	10.2	35.2	18.0	2.79	2.39	1.98	3.16	1.78	61.55
15	D15	6.6	27.0	18.9	2.88	2.50	1.75	3.52	2.15	69.77
16	D16	5.2	33.4	17.6	1.98	1.66	1.08	3.13	1.54	42.42
17	D17	11.3	29.7	21.5	2.45	2.21	1.67	3.01	1.69	53.39
18	D18	3.6	26.3	17.0	3.13	2.74	2.02	3.57	1.84	74.26
19	D19	2.3	25.3	16.6	2.97	2.39	1.99	3.47	2.35	67.71
20	D20	8.0	28.6	20.5	2.66	2.22	1.57	3.32	1.92	65.34
21	D21	18.4	37.8	21.7	3.80	3.02	1.97	4.25	2.76	95.35
22	D22	1.2	23.0	15.1	2.36	2.19	1.48	2.93	1.45	60.92
23	D23	7.2	27.1	17.6	3.37	2.71	1.79	4.15	2.49	87.87
24	D91	14.9	36.8	21.7	3.81	3.27	2.06	4.49	2.53	94.25
25	D94	2.1	27.1	13.9	2.09	1.81	1.14	3.30	1.74	45.16

Table 1: Driving Parameters

4 Data Analysis

4.1 Analysis Methods

The matrix that represents the driving parameters has been used to conduct the multivariate analysis (Table 1). We have approached two widely used methods: Hierarchical Cluster Analysis (HCA) and Principal Component Analysis (PCA). They allow the evaluation of sample similarities according to determined variables. HCA classifies the drivers according to some variables so that homogeneity within and heterogeneity among groups are obtained. PCA linearly transforms the original variables, and thus composes a new set of independent variables (components), which can be used to identify the important variables that explain sample grouping.

Cluster Analysis (CA) is a method of unsupervised learning, and a statistical methodology used to categorize individual objects into groups with similar meanings (homogeneous). CA is typically used when the researcher does not know the number of groups in advance and wishes to establish groups and to analyze group membership. It seeks to identify groups that both minimize inter-group variation and maximize outer-group variation. We have performed a hierarchical cluster analysis, starting with individual points as clusters, then successively merging two clusters until only one cluster remains. We have used Ward's method and Euclidean distance [11]. Compared to other hierarchical methods, it uses an analysis of variance approach to evaluate the distances between clusters, and it is regarded as very efficient in creating clusters [12]. The method is described more formally in Algorithm 1.

Algorithm 1 - HCA - basic Hierarchical Clustering Algorithm

Compute the proximity matrix.
repeat
  Merge the closest two clusters.
  Update proximity matrix.
until only one cluster remains.

Principal Component Analysis is a statistical method for arranging large arrays of data into interpretable patterning match. It transforms a number of (possibly) correlated variables into a (smaller) number of uncorrelated variables called *principal components*. The first principal component accounts for as much of the variability in the data as possible, and each succeeding component accounts for as much of the remaining variability as possible. The principal components are computed from the matrix of correlations between the variables, outputting their eigenvalues (the amount of variance accounted for by each component) and the component loadings (how the variables correlate with the principal component) [13–15]. This analysis attempts to plot and arrange these variables in a lower-dimensional space, where more closely related items are plotted closer to each other than the less closely related items. A simple formal description of PCA is shown in Algorithm 2.

Algorithm 2 - PCA - Principal Component Analysis

1: Organize the data as a $\mathfrak{m} * \mathfrak{n}$ matrix, where $\mathfrak{m}$ is the number of measurement types ans $\mathfrak{n}$ is the number of trials.
2: Substract off the mean for each measurement type or row in the matrix.
3: Calculate the covariance matrix.
4: Calculate the eigenvectors and eigenvalues of the covariance matrix.
5: Rearrange the eigenvectors and eigenvalues in order of decreasing eigenvalue.

4.2 Data Analysis

Using our input data, the result of cluster analysis is visualized as a tree-diagram or dendrogram in Figure 3. The dendrogram [12–14] represents the pattern of clustering among the drivers; the longer the connecting lines further to the right the more distance is between the clusters and/or the drivers. Based on the distance, we can distinguish between 6 major clusters (labeled 1 to 6 and colored differently). We can clearly see the two "extreme" clusters (no. 2 and no. 6), associated with the two controlled test drives: D91 (sample 24, cluster no. 6) and D94 (sample 25, cluster no. 2). This is also clear in some of the two-variable plots of the data in Figure 4. However, the other clusters need to be further investigated for a better interpretation.

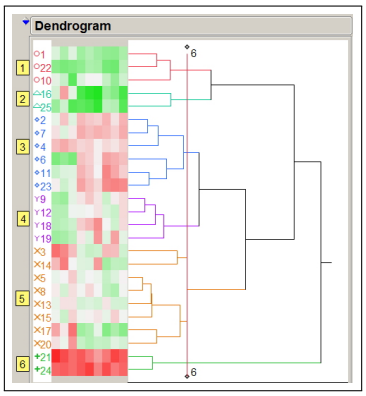


Figure 3: Hierarchical Cluster Analysis

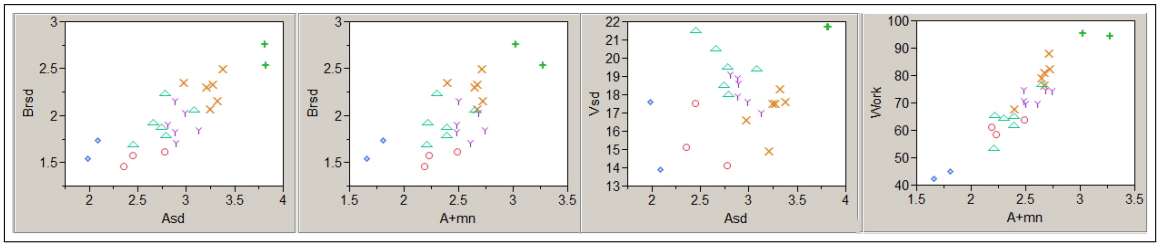


Figure 4: Different plots for Cluster Analysis

Using PCA on our data, we have obtained the principal components by computing the eigenvalues and eigenvectors of data correlation matrix as shown in Figure 5. There are a few common criteria for deciding how many components to keep: a) visual interpretation of the scree plot for the "elbow", b) the number of eigenvalues larger than 1.0, c) required meaningful percentage of variance (80-90%), and d) how many components are interpretable. Using the criteria a) and c), we can account for 2-3 components, with 2 components accounting for 84.5% of the data, and 3 components for more than 92% of the data. Based on the second criteria (b), a number of 2 components would seem more appropriate.

We will decide how many components are significant by further analyzing the correlation between the variables and the principal components, the 2- and 3-component biplots and the clustering of the data.

The correlations between the original variables and the principal components (PC1, PC2, PC3) are presented in Table 2. We can see that the first principal component have a good correlation to all acceleration- (A_{sd}, $A{+}_{mn}$, $A{+}_{sd}$) and braking- (Br, Br_{mn}, Br_{sd}) related variables, but also to the mechanical work (W), whereas the second component has a strong positive correlation to the speed-related variables (V_{60}, V_{mn}, V_{sd}). The third principal component has

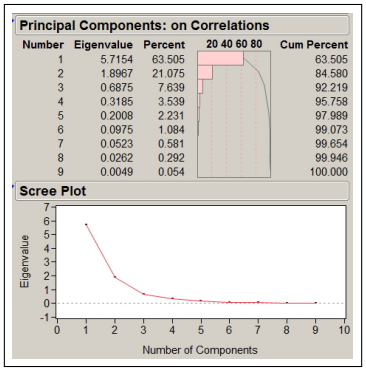


Figure 5: Principal Component Analysis

	PC1	PC2	PC3	RC1	RC2	RC3
V_{60}	0.6388762	**0.7239720**	0.0500097	0.1770218	0.1951046	**0.9302670**
V_{mn}	0.5083209	**0.7715614**	-0.1699350	-0.0693270	0.2716954	**0.8966329**
V_{sd}	0.5800321	**0.6531136**	0.2707885	0.3075579	0.0069074	**0.8612093**
A_{sd}	**0.9620060**	-0.2584930	0.0190901	**0.7478074**	**0.6315966**	0.1857643
$A+_{mn}$	**0.9105883**	-0.2841430	0.1992828	**0.8397556**	0.4688754	0.1568073
$A+_{sd}$	**0.7888608**	-0.2625120	**0.4604541**	**0.9192378**	0.1912318	0.1471884
Br_{mn}	**0.8581642**	-0.1517880	**-0.4571810**	0.3312685	**0.9062566**	0.1935466
Br_{sd}	**0.8658366**	-0.1163580	-0.3475490	0.3933960	**0.8199581**	0.2385597
W	**0.9216296**	-0.3125500	0.0299027	**0.7488098**	**0.6104728**	0.1207866

Table 2: Correlation between variables and components (factor loadings)

correlations (positive and negative) only to the acceleration and braking. Apparently, the first two principal components would give a good correlation with all the variables, however they cannot provide any indication on differentiating the use of the acceleration or of the braking pedal. By calculating the rotation of the principal components using the varimax method, we have obtained the rotated components RC1, RC2, RC3, presented also in Table 2 [15, 16]. We can now see a clear distinction between RC1 and RC2 in explaining the acceleration and breaking. The third rotated component, RC3 will have a very good correlation with the speed variables, similarly to PC2.

We further look in Figure 5 at the 2D and 3D scatterplots of the two principal components and of the first two and three rotated components. In the first, 2D plot, we can clearly see the drivers grouped in the six main groups by the cluster analysis. However, there is small overlap between clusters 3 and 4, which is difficult to separate using only the first two components. The second and third plots are made by using the rotated components: the 2D plot of the first 2 components does not separate easily the clusters; however, the 3D plot of all 3 components does indeed separate the clusters well.

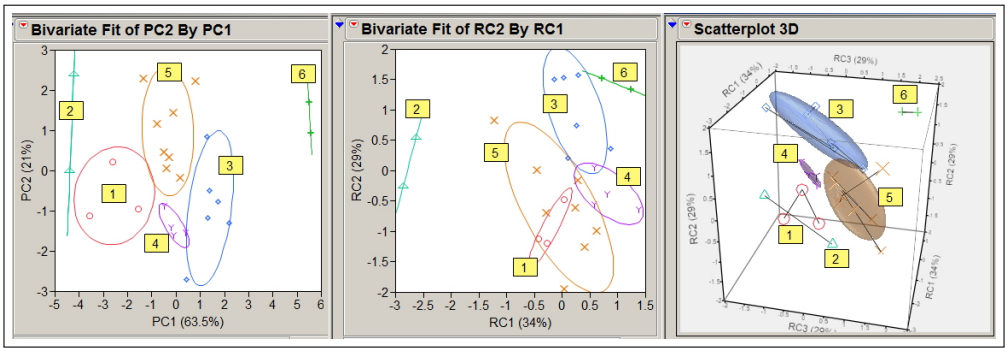


Figure 6: Scatterplots for data: a) PC1-PC2, b) RC1-RC2, c) RC1-RC2-RC3

4.3 Data Interpretation

Based on the analysis of the first two principal components (PC1, PC2) and of their values and correlations to the clusters, we can derive an interpretation (Table 3). Based on first principal component (PC1), we can suggest 5 categories of "aggressiveness": from non-aggressive (drive test D94) to a very aggressive (drive test D91). Using the second principal component (PC2), we can further obtain two categories: a tendency to drive with high speed in the city, and a more moderate speed driving.

By analyzing the three rotated components (RC1, RC2, RC3) and their values and correlation to the clusters, we can suggest three intervals for each of the factors: (very) small (below 1.0 or 1.5), near zero-one (between -1.0 and 1.0), and large (above 1.0). Their interpretation is given in Table 3. We combine the interpretations of the clusters, based on principal components and rotated components into Table 4.

Component	Values	Interpretation (driving style)	Clusters
PC1 (63.5%)	Very small (< -5)	Non-aggressive	2
	Small (-5 < -1)	Somewhat non-aggressive	1
	Between (-1 < 1)	Neutral	4 5
	Large (1 < 5)	Moderately aggressive	3
	Very large (> 5)	Very aggressive	6
PC2 (21.0%)	Negative	Low-moderate speed	1 2 4 5 3
	Pozitive	Tendency to high speed	2 5 6
RC1 (34%)	Very small (< -2)	Lower acceleration usage	2 5
	Between (-1 < 1)	Moderate acceleration usage	1 2 3 5
	Large (> 1)	Higher acceleration usage	4 6
RC2 (29%)	Very small (< -1.5)	Smooth braking	1 5
	Between (-1.5 < 1)	Moderate braking	1 2 3 4 5
	Large (> 1)	Sudden braking	3 6
RC3 (29%)	Small (< -1)	Tendency to lower speed	1 2
	Between (-1 < 1)	Moderate speed	1 2 3 4 5
	Large (> 1)	Tendency to high speed	5 6

Table 3: Interpretation by Principal Components and by Rotated Components

Cluster	Test	Aggressivity (PC1)	Speed (PC2,RC3)	Accelerating (RC1)	Braking (RC2)
1		Moderately low	Low-Moderate	Moderate	Smooth-Moderate
2	D94	Very low	Low-Moderate	Low-Moderate	Smooth-Moderate
3		Moderately high	Moderate	Moderate	Sudden
4		Neutral	Moderate	High	Moderate
5		Neutral	Moderate-High	Low-Moderate	Moderate-Sudden
6	D91	High	High	High	Sudden

Table 4: Interpretation of the Clusters

5 Conclusions and Future Work

Driving quality can be thought as the "combination of an energy saving driving style with a behavior that is respectful of the environment (noise, pollution and safety) and of the vehicle and is comfortable for the passengers as well" [3].

We are aware of the fact that many things affect drivers' behavior, such as driving environment (weather, road condition, day/night etc.), traffic context, driver's particular condition (upset, ill, tired, sleepless, distracted etc.), and, of course, driver's individual characteristics (gender, age, driving experience and frequency, annual mileage, familiarity and confidence with respect to driving, personal attitude, etc.). Unfortunately, for the time being we have not been able to access all that information for our set of drivers. The results of this work can be significantly improved by taking into consideration, at least, some of these factors, which can be easily provided (individual characteristics). To tackle other factors, such us the drivers state of mind is more difficult, but can be done with the right technical solution (real-time computer vision analysis of video streams from the driver and other sensors attached to the driver). Finally, there are some issues, as

anxiety related to possible traffic accidents, which are very hard to assess objectively, from an observer's viewpoint, without having the drivers interviewed or questioned explicitly about those matters. Though, the questionnaires-based studies may fall into the pitfall of answers from the perspective of socially desirable behavior. Anyway, when the external factors such as traffic situation or road environment have a smaller effect, "the driving behavior will be regulated more by the driving style, which is one of the internal factors" [4].

The data analysis for this paper was generated using Version 9 of the SAS System. Copyright 2008 SAS Institute Inc., Cary, NC, USA [17].

Bibliography

[1] Rigolli, M., Brady, M., Towards a Behavioural Traffic Monitoring System, International Conference on Autonomous Agents, *Proceedings of the 4th International Joint Conference on Autonomous Agents and Multiagent Systems*, pp. 449-454, 2005.

[2] Rygula, A., Driving Style Identification Method Based on Speed Graph Analysis, *International Conference on Biometrics and Kansei Engineering*, pp. 76-79, 2009.

[3] Vangi, D., Virga, A., Evaluation of Energy-Saving Driving Styles for Bus Drivers, Proc. Instn Mech. Engrs, Vol. 217 Part D: *J. Automobile Engineering*, pp. 299-305, 2003.

[4] Ishibashi, M., Okuwa, M., Doi, S., Akamatsu, M., Indices for Characterizing Driving Style and their Relevance to Car Following Behavior, *SICE Annual Conf.*, pp. 1132-1137, 2007.

[5] O. Taubman-Ben-Ari, M. Mikulincer and O. Gillath, The multidimensional driving style inventory-scale construct and validation, *Accident Analysis and Prevention*, Vol. 36, pp. 323-332, 2004.

[6] D. J. French, R. J. West, J. Elander and J. M. Wilding, Decision-making style, driving style, and self-reported involvement in road traffic accidents, *Ergonomics*, Vol. 36, No. 6, pp. 627-664, 1993.

[7] Bonsall, P., Liu, R., Young, W., Modelling Safety-related Driving Behaviour-Impact of Parameter Values, *Transportation Research Part A* 39, pp. 425-444, 2005.

[8] Mierlo, J., Maggetto, G., Burgwal, E., Gense, E., Driving Style and Traffic Measures-Influence on Vehicle Emissions and Fuel Consumption, Proc. Instn Mech. Engrs Vol. 218 Part D: *J. Automobile Engineering*, pp. 43-50, 2004.

[9] Cherrett, T., Pitfield, D., Extracting Driving Characteristics from Heavy Goods Vehicle Tachograph Charts, *Transportation Planning and Technology*, Vol. 24, No. 4, pp. 349-363, 2001.

[10] M. Vladoiu, Z. Constantinescu, Toward Location-based Services using GPS-based Devices, *Proceedings of ICWN 2008 - the 2008 International Conference of Wireless Networks*, a Conference of World Congress on Engineering 2008 (WCE 2008), July, London, UK, Vol. I, pp. 799-804, 2008.

[11] Ward, J.H., Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, Vol. 58, pp. 236-244, 1963.

[12] Anderberg, M. R., *Cluster Analysis for Applications*. Academic Press, New York, 1973.

[13] Sokal, P.H.A., Sneath, R.R., *NUMERICAL TAXONOMY: The Principles and Practice of Numerical Classification*. Freeman, San Francisco, 1973

[14] MacQueen, J.B, Some Methods for classification and Analysis of Multivariate Observations, *Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability*, Berkeley, University of California Press, Vol. 1, pp. 281-297, 1967.

[15] Jolliffe, I.T., *Principal Component Analysis*, Springer, 2002.

[16] Dauxois, J., Pousse, A. and Romain, Y., Asymptotic theory for the principal component analysis of a vector random function: some applications to statistical inference. *Journal of Multivariate Analysis*, Vol. 12, pp. 136-154, 1982.

[17] SAS Institute Inc., *SAS 9.1.3 Help and Documentation*. Cary, NC: SAS Institute Inc., 2000-2004.

[18] Hattori, Hiromitsu, Nakajima, Yuu and Ishida, Toru, Agent Modeling with Individual Human Behaviors, *Proc. of 8th Int'l. Conf. on Autonomous Agents and Multiagent Systems (AAMAS 2009)*, pp. 1369-1470, 2009.

[19] Plochl, Manfred and Edelmann, Johannes, Driver models in automobile dynamics application, *Vehicle System Dynamics*, Vol. 45, No. 7, pp. 699-741, 2007.

[20] Augustynowicz, A., Preliminary Classification of Driving Style with Objective Rank method, *International Journal of Automotive Technology*, Vol. 10, No. 5, pp. 607-610, 2009.

Int. J. of Computers, Communications & Control, ISSN 1841-9836, E-ISSN 1841-9844
Vol. V (2010), No. 5, pp. 664-664

A Fuzzy Control Heuristic Applied to Non-linear Dynamic System Using a Fuzzy Knowledge Representation

F.M. Cordova, G. Leyton

Felisa M. Cordova
University of Santiago of Chile
Ecuador 3769. Estacion Central
Chile, Santiago
E-mail: felisa.cordova@usach.cl

Guillermo Leyton
University of La Serena
Benavente 980
E-mail: gleyton@userena.cl

Abstract: This paper presents the design of a fuzzy control heuristic that can be applied for modeling nonlinear dynamic systems using a fuzzy knowledge representation. Nonlinear dynamic systems have been modeled traditionally on the basis of connections between the subsystems that compose it. Nevertheless, this model design does not consider some of the following problems: existing dynamics between the subsystems; order and priority of the connection between subsystems; degrees of influence or causality between subsystems; particular state of each subsystem and state of the system on the basis of the combination of the diverse states of the subsystems; positive or negative influences between subsystems. In this context, the main objective of this proposal is to manage the whole system state by managing the state combination of the subsystems involved. In the proposed design the diverse states of subsystems at different levels are represented by a knowledge base matrix of fuzzy intervals (KBMFI). This type of structure is a fuzzy hypercube that provides facilities operations like: insert, delete, and switching. It also allows Boolean operations between different KBMFI and inferences. Each subsystem in a specific level and its connectors are characterized by factors with fuzzy attributes represented by membership functions. Existing measures the degree of influence among the different levels are obtained (negatives, positives). In addition, the system state is determined based on the combination of the statements of the subsystems (stable, oscillatory, attractor, chaos). It allows introducing the dynamic effects in the calculation of each output level. The control and search of knowledge patterns are made by means of a fuzzy control heuristic. Finally, an application to the co-ordination of the activities among different levels of the operation of an underground mine is developed and discussed.
Keywords: Fuzzy Systems, Knowledge Representation, Heuristics, Nonlinear Dynamic Systems.

1 Introduction

Organizations can be visualized as complex systems composed of various subsystems that respond to different problems and have their own dynamics. This process in turn is recursive, so each subsystem has a particular dynamics. Such is the case of Managements, Business Areas, Departments, primary and support activities of the value chain, activities plans, besides many

other systems and subsystems existing in the organization. Each subsystem is characterized by its variables and by inputs that can alter its performance and its outputs, which are the inputs of other subsystems, whose dependent effects are known only approximately. This constitutes a situation of a set of highly dynamic subsystems and with clearly nonlinear characteristics. Usually, these factors are not considered in the decision making processes.

It is clear that a universe of this kind is quite heterogeneous, dynamic, and growing. Also, because of the nature of the stated problem, it must be considered that these subsystems represent inputs among themselves, giving the problem a high dose of parallelism. Insofar as these subsystems serve as inputs among themselves, feedback is taking place continuously, making the system's dynamics difficult to control, predict, manage and administer [1]. It is also necessary to take into account the increasing number of data, information and knowledge that current systems must administer, in particular their adequate representation [9]. If we consider that the problem of knowledge-based management and decision making must be carried out in organizations having these characteristics, then it is ever more important to support conceptual models and tools adequate for the planning, management and control processes of this dynamics.

On the other hand, the representation knowledge is a fundamental component in any intelligent system that allows coding knowledge, objects, objectives, actions, and processes. The scheme for the chosen representation of knowledge determines the reasoning process and its efficiency. Numerous studies on the representation of knowledge show that a representation can be more adequate than another one for a particular case or it can be capable of covering a greater number of cases [8]. The more traditional methods used are Semantic Networks, Frames, Production Rules, Trees, and Bits Matrices. Cazorla et al. [3] suggest that knowledge can be classified according to the specific application to be used that develops knowledge: procedural, declarative, meta-knowledge, heuristic, or structural. However, the theory of diffuse sets proposed by Zadeh [12], [13] allows the generation of knowledge representations that are closer to the nature itself of what it is desired to represent.

The conceptual models of systems, their representation based on knowledge, and the tools for supporting management and decision making must then consider in their design factors such as high dynamism, parallelism, feedback, incompleteness, handling of uncertainty, nonlinearity, vagueness, qualitative definitions and behaviors, personal opinions, etc. Along this line, some authors [1], [16], [15] make a profound development of various concepts such as fuzzy function approximations, chaos and fuzzy control, and processing of fuzzy signals. However, his greatest contribution refers to the calculation and representation of knowledge by means of fuzzy cubes and fuzzy cognitive maps. McNeill [6] also works with fuzzy theory as a means of representing environments with uncertainty usually characterized by their nonlinearity. Welstead, on the other hand, supported by one of Kosko's results [11] suggests that fuzzy rules can be represented by one or more fuzzy associative memory matrices (FAM); combining the above with genetic algorithms he proposes a model to approach prediction problems. They also use fuzzy representations centered mainly on the interaction of fuzzy theory, neural networks, and genetic algorithms, supporting a new line of work known as Computational Intelligence. Tsoukalas [10] is more centered on the interaction and creation of fuzzy theory and neural network hybrids. To approach these kinds of problems, models are designed making use mainly of causal diagrams or knowledge maps with a series of nodes that would represent the concepts that are relevant to the system, and links between them that show the causal relation (influence) between concepts. In this context, the objective of this paper is to make a study and analysis that will allow modeling some types of dynamic systems, representing knowledge by means of a knowledge base matrix of fuzzy intervals and fuzzy cognitive maps [4], [14] and [15] with the purpose of achieving their categorization and fuzzy weight, as well as the levels of incidence in other subsystems, in this way characterizing the complete system with its levels of fuzzy incidence [5], [10].

2 Modeling of the Diffuse Knowledge Base Matrix

In this proposal each of the map's concepts corresponds to a fuzzy set, and it is specifically a particular Knowledge Base Matrix of Fuzzy Intervals (KBMFI). The connections between concepts will have an associated value in the [-1,1] range that represents the degree of influence of a (KBMFI) node on another. If the value is positive, it indicates that an increase on the evidence of the origin concept increases the meaning, the evidence or the truth value of the destination concept. If it is negative, an increase of the evidence of the source causes a decrease of that of destination. If the value is 0, there is no connection, and no causal relation.

In this way it is possible to get blurred cognitive maps from the opinion of one or various experts on the relations between some aspects of the evaluation process of a hypothetical case. Also, the clear recursiveness involved in these types of systems is considered, and a vision of granularity is proposed that allows overcoming the various levels of abstraction subjacent in the dissimilar subsystems. On the other hand, internally each subsystem can be represented by KBMFIs, allowing their incidence weight to be obtained with respect to other subsystems and at the same time represent their particular behavior.

Definition 1. Let X be a classical set of objects, called the universe. Belonging to a subset A of X can be defined in terms of the characteristic function:

$$\mu_A : X \longrightarrow [0,1] \; x \longrightarrow \mu_A(x) \tag{1}$$

where:

$$\mu_A(x) = \begin{cases} 1 & x \in A \\ 0 & x \notin A \end{cases}$$

If the evaluation set $0,1$ is extended to the real interval [0,1], then it is possible to talk about the partial belonging in A, where $\mu_A(x)$ is the degree of belonging of x in A, and the values 0 and 1 are interpreted as "non-belonging" and "total belonging", respectively.

Clearly, A is a subset of x, which has no defined boundaries. This leads to the following definition.

Definition 2. Let X be an object's space. A fuzzy set A of X is characterized by the set of pairs:

$$A = \{(x, \mu_A(x))/x \in X\} \; where \; \mu_A : X \longrightarrow [0,1] \tag{2}$$

The fuzzy concept proposed by Zadeh [11] is based on the fact of allowing the partial belonging in a set for certain elements of a given universe.

Definition 3. A fuzzy hypercube can be considered as a unit hypercube, i.e., a hypercube $I^n = [0,1]^n$. The n fuzzy cube has two vertices or binary subsets.

A fuzzy cube contains all the fuzzy sets of a set X of n objects. The non-fuzzy sets are found at the vertices of the cube. The continuum of fuzzy sets is in the cube.

Definition 4. Knowledge Base Matrix of Fuzzy Intervals (KBMFI) means the hypercube that is constituted by the various knowledge $E_1, E_2, E_3, ..., E_n$, relative to a domain of knowledge, considering also the different weight or importance that each of them has in the particular domain.

The KBMFI is a fuzzy hypercube where $E_1, E_2, ..., E_n$, represent the various contingencies or characteristics of the area under discussion, according to the opinion of the experts. E_j, with $j = 1, 2, ..., n$, do not necessarily have the same relevance or weight, they can be in particular

fuzzy frames consisting of $S_1, ..., S_m$, where $S_1, S_2, ..., S_m$, are the possible factors, not necessarily disjoint, such that each characteristic E_i can be expressed by means of some particular union of $S_1, S_2, ..., S_m$ factors. Now the particular determination of each E_i through its particular factors $S_1, S_2, ..., S_m$, model systems composed of a range of nodes $N_1, N_2, ..., N_n$, continually influencing each other if and where the incidence of one with respect to others is completely dynamic. In particular, this outlines a vision of dynamic nonlinear systems which in similar but not equal versions are seen as causality maps.

If the map is adjusted to the opinions of several experts, one would have to get the assessments of all of them and therefore establish the definitive values associated with the causality relations. It must be noted that in general the causalities mentioned by the experts with respect to the various influences exerted by the nodes of the maps are more attributable to qualitative than quantitative concepts.

As already stated, nonlinear dynamic systems involve nonlinear and feedback behaviors. In these systems the output of a process or node is used as input for the following node or iteration, and the output of this can again be the input of the same previous node, i.e., self-recurrent behaviors. This behavior corresponds to the following equation:

$$f(x_0) = \begin{cases} X_{n-1} \\ X_n \\ X_{n+1} \end{cases}$$

Assuming that the following situation occurs when modeling the system: $x_1, x_2, x_3, ..., x_n$.

Definition 5. Let x_0 be an arbitrary starting node, then the above sequence is called the Trajectory.

Considering these definitions, several behaviors can occur, such as, for example: fixed points; periodic trajectories; behaviors given by attractor nodes, and chaos.

3 Case Study

The case study corresponds to the situation of an underground mine which has three levels: Production Level, Reduction Level, and Transport Level. The problem consists in "providing support to activity scheduling management". The problem consists in "providing support to activity scheduling management" [2]. The total system shown by Figure 1 consists of these three subsystems and the dynamics that exists between them. This situation is denoted as Level 1.

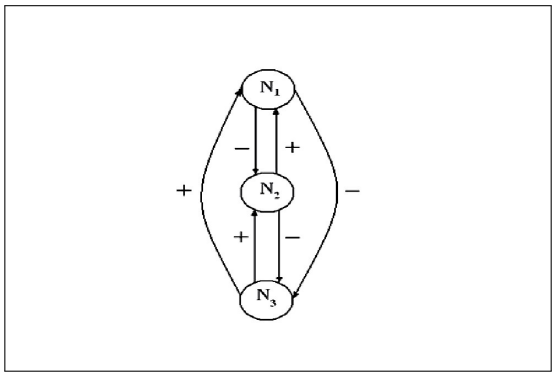


Figure 1: Production, Reduction and Transport Levels.

N_1: Production Level considers N_{11}, N_{12}, N_{13}, as subsystems; N_2: Reduction Level considers N_{21}, N_{22}, as subsystems; N_3: Transport Level considers no subsystems.

Looking at it at a more particular abstraction level, Level 2 appears, as shown in Figure 2. From the particular situation shown, in Figure 1 it is seen that: N_1 influences N_2 negatively and N_3 positively, N_2 influences N_1 and N_3 positively, N_3 influences N_2 negatively and N_3 positively.

However, Figure 2 shows that the information obtained at Level 1 of abstraction of the system does not have the sensitivity or reliability that is obtained at Level 2 of abstraction, whose granularity or disaggregation is slightly higher.

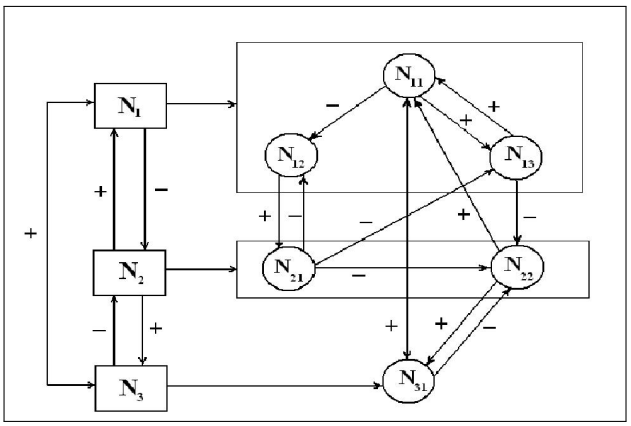


Figure 2: Diagram of influence at the different levels.

If both levels are confronted, it may be incorrectly deduced that apparently contradictory information is obtained. For example, if we look at Level 1 and Level 2 for the case of N_3 with N_2, at Level 1 it was stated that N_3 influences N_2 negatively, but at Level 2 it could be concluded that both have the opposite influence, N_{31} influences N_{22} negatively and N_{21} influences N_31 positively. This apparent contradiction can be explained, for example, by saying that when production at the Reduction Level decreases, there is less pressure on the demand for trains or cars, and on the other hand, if there is not sufficient transport from N_{31} there is an impact due to accumulation of material at the Reduction Level, which is considered a negative influence. Then the question is, which of the two situations has greater incidence weight? According to Figure 3, and only as an example, it can be stated that the negative impact from N_{31} to N_{22} is greater than the influence of N_{22} on N_{31}.

The main observations to the system are: it is clear that it is a Dynamic Fuzzy System. In turn, every N_i is a Dynamic Fuzzy Subsystem. The connections between the various N_i are fuzzy. These connections can be positive or negative. If positive, N_i influences positively on N_j. If negative, N_i influences negatively on N_j.

4 Design and Implementation of the KBMFI Matrix

Going more deeply into Table 1, the experts draw these KBMFI as causal tables. They do not state equations, but make links between subsystems. The KBMFI systems convert each pictograph into a Fuzzy Rules Weight Matrix. The nodes of the KBMFI can model the complex nonlinearities between the input and output nodes. The KBMFI can model the dynamics that occur in the multiple iterations that take place in these dynamic systems.

The KBMFIs with N nodes have N^n arcs. Since $N_i(t)$ nodes are fuzzy concepts, their values $\in [0, 1]$; a state of a KBMFI is the $N_i(t) = (N_1(t), N_2(t), ..., N_n(t))$ vector, so it is a point of the

hypercube $I^n = [0,1]^n$.

An inference in a KBMFI is a road or sequence of points in I^n, i.e., it is a fuzzy process or an indexed family of fuzzy sets $N(t)$. It is clearly seen that the KBMFIs can perform "forward chaining," and whether they can perform "backward chaining" (nonlinearity inverse causality) is an open question. The KBMFIs form, as nonlinear dynamic systems, Semantic Fuzzy Networks and act as neural networks. The KBMFIs can converge to a fixed point, to a limited cycle, that can be a stable or oscillating state or a chaotic attractor in the fuzzy cube In. In this context, one of the basic questions to be answered is: what happens if the input to the (KBMFI) system is known? In this sense, each KBMFI stores a set of global rules of the form:

$$\text{IF } N(0) \text{ THEN attractor } A \tag{3}$$

A KBMFI with a single fixed global point has only one global rule. The size of the attractor regions in the fuzzy cube governs the number of these global regions or hidden patterns. The KBMFIs can have large and small attractor regions in In, each of them with a different degree of complexity. Therefore an input state can lead to chaos and a relatively close input state can end up in a fixed point or limited cycle or a stable state. Since the KBMFIs correspond to a Semantic Fuzzy Network structure, it is possible to associate a matrix M. This matrix lists the causal links between N_i nodes. As an example, if it is considered again the case described by Figure 2, the corresponding KBMFI is presented where a row is the incidence of N_i on N_j; columns are nodes influence N_i and $\alpha, \beta, \gamma, \delta, \eta, \tau$, are values. Fuzzy function: [little, moreorless, much, etc.].

	N_{11}	N_{12}	N_{13}	N_{21}	N_{22}	N_{31}
N_{11}	0	$-\alpha\mu$	$+\alpha\mu$	0	0	$+\alpha\mu$
N_{12}	0	0	0	$+\beta\mu$	0	0
N_{13}	$+\gamma\mu$	0	0	0	$-\gamma\mu$	0
N_{21}	0	$-\eta\mu$	$-\eta\mu$	0	$-\eta\mu$	0
N_{22}	$+\delta\mu$	0	0	0	0	$+\delta\mu$
N_{31}	$+\tau\mu$	0	0	0	$-\tau\mu$	0

The proposed model is decomposed in diverse abstraction levels and at each level is represented by a corresponding KBMFI. Initially, observing Figure 2, the Abstraction Level 0 appears. Only the influence shapes are observed. A Node N_i can influence positively or negatively to the Node N_j. Abstraction Level 0 appears:

	N_1	N_2	N_3
N_1	0	-	+
N_2	+	0	-
N_3	+	-	0

Experts are asked to qualify the degree of influence between: μ = [nothing, irrelevant, few, influence, regular, alter, a lot, very much, so much] as shown in the Incidence Graphic of Figure 3.

Applying the incidence graphic, a second level of abstraction 01 is obtained:

	N_1	N_2	N_3
N_1	0	$-\mu$	$+\mu$
N_2	$+\mu$	0	$-\mu$
N_3	$+\mu$	$-\mu$	0

It is observed that the degree of incidence between a node N_i with a node N_j, this means that it exists a bigger degree of specificity (granulation) between them. This enhancement of

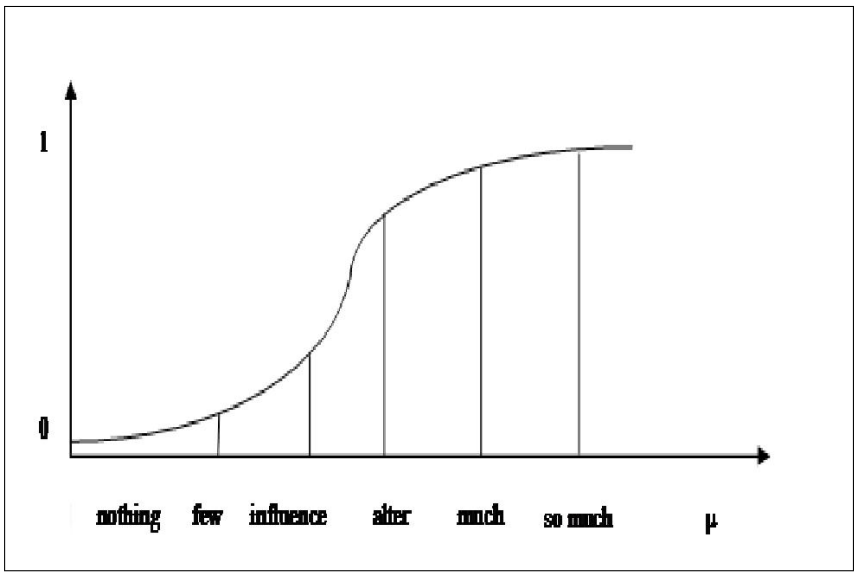


Figure 3: Incidence Graphic.

specificity is explicit in the following level, it exist a "slot" between N_i with N_j. In this case different situations are denoted: N_1 influences in a negative way to N_2; N_1 influences in a positive way to N_3; N_2 influences in a positive way to N_1; N_2 influences in a negative way to N_3; N_3 influences in a positive way to N_1; N_3 influences in a negative way to N_2.

If it is considered that a Node N_i can be decomposed in $N_{i1}, N_{i2}, ..., N_{ik}$, in where those N_{im}, $m = 1, 2, ..., k$, with a particular dynamic conforms a N_i, the situation in the analyzed case is as follows:

$N_1 = (N_{11}, N_{12}, N_{13})$; at Level 0; Node or Subsystem N_1.

N_1 at Level 01, Node or Subsystem N_{1i} is defined by:

	N_{11}	N_{12}	N_{13}
N_{11}	0	-	+
N_{12}	0	0	0
N_{13}	+	0	0

N_1 at Level 011 is defined by:

	N_{11}	N_{12}	N_{13}
N_{11}	0	$-\alpha\mu$	$+\alpha\mu$
N_{12}	0	0	0
N_{13}	$+\alpha\mu$	0	0

$N2 = N_{21}, N_{22}$; at Level 0; Node or Subsystem N_2.

N_2 at Level 01, Node or Subsystem N_{2i} is defined by:

	N_{21}	N_{22}
N_{21}	0	-
N_{22}	0	0

N_2 at Level 011 is defined by:

	N_{21}	N_{22}
N_{21}	0	$-\eta\mu$
N_{22}	0	0

Applying the same procedure to node N_3 and it is only characterized by N_{31}, N_3 at Level 011 is defined by:

	N_{31}
N_{31}	0

At this point only the fuzzy subsystem cohesion is developed. So, it is necessary to visualize what it happens with the external dynamic between subsystems, in order to obtain the fuzzy matching inter systems. Continuing with the fuzzy cohesion procedure, links between nodes N_1, N_2 and N_3, at Level 0 by Nodes are obtained:

	N_1	N_2	N_3
N_1	0	-	+

At Level 01 by Node N_1:

	N_1	N_2	N_3
N_1	0	$-\alpha\mu$	$+\alpha\mu$

At Level 011 by Node N_1:

	N_1	N_{21}	N_{22}	N_{31}
N_{11}	0	0	0	$+\alpha\mu$

At Level 02 by Node N_1:

	N_1	N_{21}	N_{22}	N_{31}
N_{11}	0	$+\beta\mu$	0	0

In this way, influences are obtained allowing the fuzzy matching.

5 Heuristic Control for the KBMFI

Each N_i level has F_{ij} factors that determine it, with $i = 1, 2, 3$; $j = 1, 2, ..., m$. Table 1 shows relevant characteristics, factors, attributes and fuzzy functions at Production Level.

Table 2 shows relevant factors, attributes and fuzzy functions at Production Level.

Each F_{ij} factor has $A_{ij}s$ attributes that determine it, where $i = 1, 2, 3$; $j = 1, 2, ..., m$; $s = 1, 2, ..., k$ (see Table 1).

Each $A_{ij}s$ has attribute metrics associated with its nature. These metrics are functions of fuzzy membership (see Table 2).

For the above points, it is possible to state that the degrees of influence (negative or positive) that exist between the various levels can be measured, allowing the calculation of the existing dynamics of the system to achieve an Intelligent Fuzzy Control with the purpose of keeping the system in a desirable state (stable).

CHARACTERISTICS OF LEVEL 1 (PRODUCTION)	ATTRIBUTES (Metrics or fuzzy functions Table 1.1)									
		Fuzzy Functions								
FACTORS 1	Rel. Card. (CRC)	1	2	3	4	5	6	7	8	9
1. Number of workmen present										
2. Drilling, agents, and resources										
3. Blasting, agents and resources										
4. Technologies involved										
5. Number of equipments										
6. Lectures										
Relative cardinality of Level 1 (CRN1)										

Table 1: Relevant characteristics of Level 1 at Production Level.

6 Heuristic

The proposed heuristic consists of the following stages:

Stage 1: Obtaining the F_{ij} factors of each level N_i.

Stage 2: Obtaining the A_{ij}s attributes of each F_{ij} factor.

Stage 3: Determining the metrics associated with each A_{ij}s attribute, i.e., determining the fuzzy membership functions for each A_{ij}s.

Stage 4: Determining the "formula" that corresponds to each F_{ij} from the A_{ij}s, for example:

$$F_{ij} = \lambda_1 A_{ij}1 \oplus \lambda_2 A_{ij}2 \oplus ... \oplus \lambda_k A_{ij}k \tag{4}$$

where $\oplus$ is the operator to be determined $(=>, \vee, \cup, \text{etc.})$ and $\sum \lambda_k = 1$.

Stage 5: Determining N_i from the F_{ij}, for example:

$$N_i = \lambda_1 F_{i1} \oplus \lambda_2 F_{i2} \oplus ... \text{ oplus} \lambda_m F_{im} \tag{5}$$

where $\oplus$ is the operator to be determined $(=>, \vee, \cup, \text{etc.})$ and $\sum \lambda_m = 1$.
Note that the output of all N_j must be between 0 and 1.

Stage 6: Determining whether the "influence" of the output of N_i to other levels is negative or positive.

Stage 7: Recalculating the N_t output, with its internal values, considering the influence exerted on it by the recursive dynamics of the nodes N_i at Stages 1, 2,..., 5.

Stage 8: Determining the output of N_t, input of N_l, and determining whether we feed N_l or N_i, and specifying the times. Note that in this step we distinguish between what influences what, or we make a push, we make a pull, or both at the same time, with a delay of one with respect to the other, etc.

FACTORS AND ATTRIBUTES	FUZZY FUNCTIONS
PROJECT SYSTEM	
1. Number of workmen present. (Decision making complexity).	
1.1 Number of engineers.	$\mu_1^1(x) = 1 - \frac{25-x}{25} \quad 10 \leq x \leq 25$
1.2 Number of technicians.	$\mu_2^1(x) = 1 - \left(\frac{75-x}{75}\right)^2 \quad 30 \leq x \leq 75$
1.3 Number of miners.	$\mu_3^1(x) = 1 - \left(\frac{150-x}{150}\right)^2 \quad 60 \leq x \leq 150$
1.4 Number of equipments	$\mu_4^1(x) = 1 - \frac{30-x}{30} \quad 12 \leq x \leq 30$
	$x =$ **amount of engineers, miners, ...**
2. Drilling, agents and resources.	For evaluating this characteristic, first the predominant factor must be identified and then the calculation can be made. For example, if x = 25 or 30 or 90 or 21, for respective:
2.1 Planned drillings.	$\mu_i^1 : \mu_1^1(25) = 1; \mu_2^1(30) = 0.64; \mu_3^1(90) = 0.84; \mu_4^1(21) = 0.91$
2.2 Direct agents involved.	$\mu_1^2(x) = 1 - \sqrt{\frac{50-x}{50}} \quad 20 \leq x \leq 50$
2.3 Indirect agents involved.	$\mu_{2,3}^2(x) = 1 - \left(\frac{15-x}{30}\right)^3 \quad 6 \leq x \leq 15$

Table 2: Factors, attributes and fuzzy functions at Production Level.

7 Conclusions and Future Works

The work done in the paper allows the characterization of a complex system through subsystems considering the dynamics and the incidence of each subsystem on the others. From the display of the complexity of the system and subsystems, the KBMFI is constructed, which allows an adequate representation of diffuse knowledge and the dynamics associated with the system. A fuzzy control heuristic is also designed that allows managing the KBMFI.

In the case of the planning of mining operations, the KBMFI and the associated heuristic allow the evaluation of the impact of the incidence of various factors such as reduction of the number of planned workers in a shift, faults in Load Haul and Dump LHD equipment, rock breakers, shafts, and trains, among others.

If someone is considering developing software from this proposal, it should be kept in mind that in the tool there should be an agent module that is informed (alert) of the acceptable critical values for each node, so that this node does not alter acceptable states (experts) of the nodes with which it interacts. In such case the agent must learn about the acceptable critical values, know and learn preventive measures; know and learn mitigation measures, and know and learn corrective measures.

Bibliography

[1] D. Alahakoon, S. K. Halgamuge, and B. Srinivasan, Dynamic self organizing maps with controlled growth for knowledge discovery, *IEEE Trans. Neural Networks*, Vol.11:601-614, 2000.

[2] F. Cordova, L. Canete, L. Quezada, F. Yanine, An Intelligent Supervising System for the Operation of an Undergound Mine, International Journal of Computers, Communications and Control, Vol. III: 259-269, 2008.

[3] M. Gupta, R.K. Ragade,Yager, Advances in Fuzzy Sets Theory and applications, North Holland, Amsterdam, 1979.

[4] B. Kosko, Fuzzy Ingineering. Prentice Hall, 1997.

[5] G. Martinez, Servente and Pasquini,Sistemas Inteligentes, NL Nueva Libreria, Argentina, 2003.

[6] T. McNeill, Fuzzy Logic a Practical Approach. Academic Press, 1997.

[7] H. Roman, Sobre Entropias Fuzzy, Tesis de doctorado, Universidad de Campinas, Brasil, 1989.

[8] E. Schnaider, A. Kandel, Applications of the Negation Operator in Fuzzy Production Rules, Fuzzy Sets and Systems, Vol. 34: 293-299, Noth Holland, 1990.

[9] W. Silder, J. Buckley, Fuzzy expert system and fuzzy reasoning, John Wiley and Sons Inc., New Jersey, 416, 2005.

[10] U. Tsoukalas, Fuzzy and Neural Approaches in Engineering. Wiley Interscience, 1997.

[11] S. Welstead, Neural Network and Fuzzy Logic Applications in C++. Wiley Interscience, 1994.

[12] L.A. Zadeh, The role of fuzzy logic in the management of uncertainly in Expert Systems, Aproximate Reasoning in Expert Systems, Elservier Science Pub., North Holland, 3-31, 1985.

[13] L.A. Zadeh et al (eds.), *From Natural Language to Soft Computing: New Paradigms in Artificial Intelligence*, Editing House of Romanian Academy, 2008.

[14] Zadeh, L.A., Outline of a new approach to the analysis of a complex systems and decision processed, IEEE Trans. Syst. Man Cybern., Vol. 3: 28-44, 1973.

[15] L.A. Zadeh, Fuzzy sets and fuzzy information: granulation theory, Beijing Normal University Press, Beijing, 1997.

[16] L.Zhong, W.A. Halang,G. Chen, Integration of Fuzzy Logic and Chaos Theory, Springer-Verlag, Berlin Heidelberg, 2006.

Int. J. of Computers, Communications & Control, ISSN 1841-9836, E-ISSN 1841-9844
Vol. V (2010), No. 5, pp. 675-683

Towards Open Agent Systems Through Dynamic Incorporation

C. Cubillos, M. Donoso, N. Rodríguez, F. Guidi-Polanco, D. Cabrera-Paniagua

Claudio Cubillos, Makarena Donoso
Nibaldo Rodríguez, Franco Guidi-Polanco
Pontificia Universidad Católica de Valparaíso
Av. Brasil 2241, Valparaíso, Chile
E-mail: {claudio.cubillos,nibaldo.rodriguez,fguidi}@ucv.cl
makarena.donoso@gmail.com

Daniel Cabrera-Paniagua
Universidad de Valparaíso
Valparaíso, Chile
E-mail:daniel.cabrera@uv.cl

Abstract: This work tackles the problem of providing a mechanism and infrastructure for allowing a given Multiagent System (MAS) to become open, allowing the incorporation of newly incoming agents to participate within the existing society. For this, a conceptual analysis of the so-called conciliation problem is presented, covering the diverse levels and issues involved in such a process. Our Dynamic Incorporation Architecture is presented, which implements an infrastructure for allowing the participation of external agents into a specific multiagent system by incorporating the appropriate behaviours upon arrival. Our multiagent architecture for dynamic incorporation covers three levels: semantics, communication and interaction and has been applyed in a book-trading e-market scenario.
Keywords: Multiagent System, Dynamic Incorporation Architecture, Open Agent Systems, PASSI.

1 Introduction

Software agents are defined as autonomous entities capable of flexible behavior denoted by reactiveness, pro-activeness and social ability [1]. Multiagent systems (MAS) consist of diverse agents that communicate and coordinate generating synergy to pursue a common goal. At present, Multiagent Systems (MAS) raises as a key paradigm for the development of next generation software systems which are required to be distributed, intelligent (autonomous, proactive), open and dynamic. While much work has been done by the research community in solving distribution and intelligence issues, little effort has been devoted to openness and dynamicity under MAS settings.

In the medium-to-long-term future we will see open multi-agent systems spanning multiple application domains, and involving heterogeneous participants developed by diverse design teams. Agents seeking to participate in these systems will be able to incorporate and learn the appropriate behavior for participation in the course of doing so, rather than having to prove adherence before entry (as happens today) [2].

However, up to now agent systems typically center on closed agent systems with ad-hoc designs and predefined communications protocols. In recent years, agent systems have evolved to the use of agreed protocols and languages thanks to a huge standardization effort (FIPA, OMG, W3C). Nowadays agent system openness is limited to the participation of any agent able to satisfy publicly-advertised standards. Moreover, typically communication protocols, languages

(ACLs) and domain knowledge model (ontology) are defined by the design team prior to any agent interactions. Therefore much work needs to be done for the above scenario to become true.

The current work presents a conceptual analysis of the so-called conciliation problem for then presenting the design of a multiagent architecture devoted to facilitate the dynamic participation of external agents into a specific multiagent system by incorporating the appropriate behavior upon arrival. A book-trade e-market has been used as study case to validate the multiagent implementation. This work gives continuity to our previous research in [4] [5].

2 Related Work

Relevant research has been developed around coalition formation; the process to form a group of agents and solve a problem via cooperation [4], some works on Dynamic Coalition Formation (DCF) [5] tackle the issue of dynamically building beneficial coalitions (coalition algorithms) among agents that can cope with environmental changes without restarting the negotiation process.

However, at present coalition formation for virtual organizations is limited, with such organizations being largely static. All of the existing work has been devoted to optimizing for a given agent, the decision of when to participate or not in a coalition (conformation/disband) but not in providing an infrastructure supporting such dynamic coalition conformations and their agents' heterogeneities.

The novelty of our work relies on 1) presenting a conceptual approach for conciliating agents' heterogeneities under an open MAS setting 2) present a solution based in dynamic behavior loading rather than in a mediated-architecture approach and 3) implementing a solution that solves heterogeneities at two levels: semantics and interaction (behavior) while leaving the implementation of the communication level (related services and conciliation)as further work.

3 The Three Conciliation Levels

For the Dynamic formation of agent systems to become true, a set of diverse issues need to be solved first. This concern: 1) The mechanism needed to provide the incorporation of a foreign agent to the society (agent system), 2) The ways to allow the agent to incorporate upon-arrival the society common knowledge (ontology) in order to interact, and 3) The means allowing the foreign agent to learn or incorporate upon-arrival the communication protocols and languages used by the society (AIPs [6], ACLs [7], etc) and the inherent business model.

Each of the above issues regards a different aspect of the communication infrastructure used by the agent society to interact and each one is covered by different areas of informatics and computer science. In general terms, the problem of conciliating the existing divergences on the communication capabilities of the entering agent and the existing society can be organized into three levels: firstly, the interaction level, involving the diverse agent interaction protocols (AIP's) to formalize the conversations among agents. Then the communication level, tackling the heterogeneinity at message level, that is, the protocol and language used on the messages. Finally, the semantic level, conciliating the possible divergences on the symbols used in the messages and the underlying knowledge models of the new agent and the MAS society. Each of these is further explained in the following.

3.1 Interaction Level

The highest level is the interaction one, tackling the conversations among pair of agents through agent interaction protocols (AIP's) that specify the underlying coordination / coopera-

tion mechanism. Examples can be the contract-net protocol (CNP) [11], the different auctions (e.g. English, Dutch, Vickrey, etc.), tuple-based negotiations, among others. These specific protocols are grounded into interaction diagrams that specify the roles of the participants and the expected messages to be sent and received under which conditions. In practical terms, the new agent will need to incorporate the behavior needed to perform a certain role within the MAS society. For example, turning back to the cooper e-market, the new agent will need to add or load the manager role for initiating a contract-net, which will allow him to make the call for proposals, evaluate and select the best proposal and award it. All these tasks will need to be incorporated at run-time by simply instantiating the corresponding behavior classes and adding it to the agent.

By having this solve, the agent will still need to understand the parameters required by each of the tasks it has loaded, moving us to the semantic level. Other alternative is to have the agent with the correct interaction protocol and role (e.g. manager of a contract-net) but have divergences on the message format or language used in the content, moving us to the Communication level. Both are described below.

3.2 Communication Level

Nowadays agent systems do use communication protocols of their choice that define the message structure or syntax and language used to express the content of the message. Examples of such can be FIPA ACL [8], KQML, an ad-hoc XML-based message envelope, or even the simple but effective concatenization of the different data in an specific order (e.g such as in low-level protocols). It can happen that the new agent does know the FIPA ACL Message format but not the Prolog language used in the content of the message or can do not know both. Now when considering the newly incoming agent, how can we enable the incorporation of those message formats and content language upon entry?

One scenario is when the new agent has already the adequate roles and behavior to interact (interaction level) but with a message format and content language that is different from the ones used within the agent society. In this case, a mediation service can be provided, by establishing the mapping of the different slots used on the two message formats. A more efficient approach is the use of a meta-format (meta-ontology for communication) used as message interchange format [9]. In this way each new format must provide its mapping to the meta-format and vice-versa instead of providing its mapping to all the other existing formats.

Other scenario is when the new agent does not have incorporated the interaction role nor the corresponding behaviors. In this case are directly adopted the format for messages and language used by the society when incorporating the behaviors, that is, when solving the problems at interaction level.

3.3 Semantic Level

Existing coordination strategies rely on standard interaction and communication protocols, assuming that all participants (agents) understand the shared domain knowledge usually modelled in terms of an ontology. This assumption is no longer valid in open systems in which an incoming agent (usually developed by another party and in another moment) needs to interact with a certain agent society to obtain some service (pursuing a specific goal) and does not know in advance such domain knowledge used to interact. Some questions to answer are: How can the agent request or get that ontology from the agent society and process it or understand it? or how can the agent society provide the incoming agent with all the required knowledge.

The basic problem here is how to conceal the concepts of the new agent with regard to the ones used by the society. Therefore the entering agent will need to be provided by a service

capable of aligning his knowledge models with the one of the MAS society, requiring ontology alignment and mapping techniques to solve it.

4 Dynamic Participation Architecture

This project used PASSI (Process for Agent Societies Specification and Implementation) as methodology of development, which uses UML as modelling language. For a detailed description please refer to [3]. Figure 1 Shows our conceived solution in terms of the conciliation levels described before.

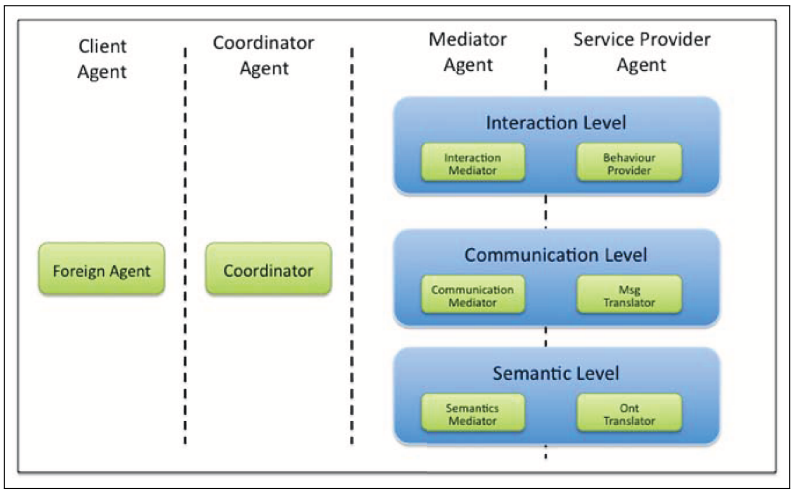


Figure 1: The 3-Tier architecture for conciliation

Firstly, the Foreign Agent (FA) is the one that wants to participate in a specific MAS to which it does not belong beforehand, thus requiring a conciliation process. A Coordinator manages the diverse steps in the agent incorporation from the initial request until the agent is ready to start interacting within the MAS society. Another type of agent is the Mediator, existing one for tackling each of the three levels of possible divergences (semantic, communication and interaction). In the first two cases, each one will have subscribed a set of Translator agents providing specific bridging services among a couple of ontologies, or a pair of message formats in correspondence with the level. In the case of the Interaction Mediator, it will be having associated a set of Behavior-Provider agents, each one containing the code for the roles in diverse interaction protocols (e.g. manager/bidder in Contract Net, auctioneer/auctionee in Auctions, etc.).

The Figure 2 shows when a new agent requests incorporation to the Book-trading MAS. For this, the FA agent sends a conciliation request to the Coordinator containing: the MAS domain name to which the FA wants to contact (Booktrading in our case); the identifier of the message protocol that the FA has; the ID of ontology that the FA knows; and the role name that the FA wants to perform within the Book-trading MAS. Such request is evaluated by the Coordinator checking whether if the FA needs conciliation services at any of the three considered levels; be provided with appropriate behaviors to participate (interaction), message-protocol translators (communication) and ontology translators (semantic). for this it compares the protocols and semantics used by the agent and by the domain MAS searching for heterogeneities.

The coordinator derivates to the Interaction Mediator the search for the requested role and related coded-behaviors. On its turn, the mediator contacts specific Behavior providers following the Contract-Net Protocol [11]. A similar process is carried out for the Communication Mediator and Semantic Mediator. In the first case, the Coordinator sends the ID of the message protocol

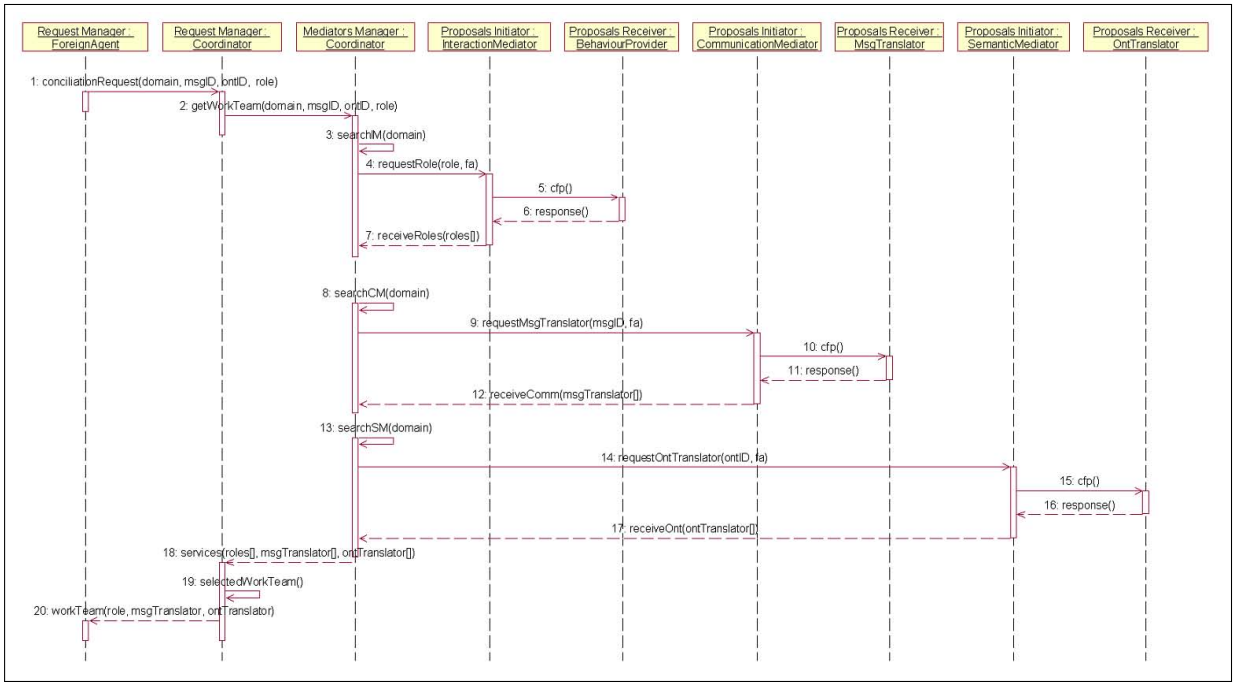


Figure 2: Scenario showing when a Foreign Agent requests a conciliation

used by the FA and the domain MAS, while in the second, sends the ontologies ID of the domain MAS and FA. In all three cases what is received is a list of possible providers and their proposals(bids), as the Coordinator is the one incharge of selecting the final providers, composing the Work Team and announcing it to the FA.

For selecting the appropriate members of the Work Team, the Coordinator should consider diverse aspects such as the cost, quality of service, reliability, etc. of each candidate according to its utility function. On its actual form, the Coordinator selects the providers based on the cost of each service. More variables within the objective function together with other selection schemas (e.g. foreign agent selection or a mixed approach) remains a matter of future work.

The conciliation process may result in the need of services with any combination of the three levels or even require no services in case of a perfect match. In this way, the Coordinator has the responsibility of structuring an appropriate Work Team for the FA. Usually, the Coordinator will select an OntologyTranslator (OT), a MessageTranslator (MT), and a BehaviorProvider (BP) agent.

It is important to mention that each role that the FA agent wants to perform needs a specific Work Team, therefore are tackled as different conciliation requests. However, nothing prevents a service provider of participating in diverse work teams, even for a same foreign agent. Together with requesting and selecting an appropriate work team for the Foreign Agent, another important process is the actual enforcement of those services. In practical terms, this means sending the behaviors from the provider to the FA, plus the possible translation services at communication and semantic level that could be needed.

The Figure 3 shows the services' enforcement, beginning with the Foreign Agent which decomposes the received working team. It gets the behaviors from the selected Behavior Provider while obtaining the corresponding codecs from the Message and Ontology Translators. Upon receiving the behaviors that compose the role, the FA takes the list of parameters required by the behaviors to work and translates it with the Ontology codec. This codec translates the concepts from the domain-MAs ontology to the FA ontology. In this way, the FA can map the values of

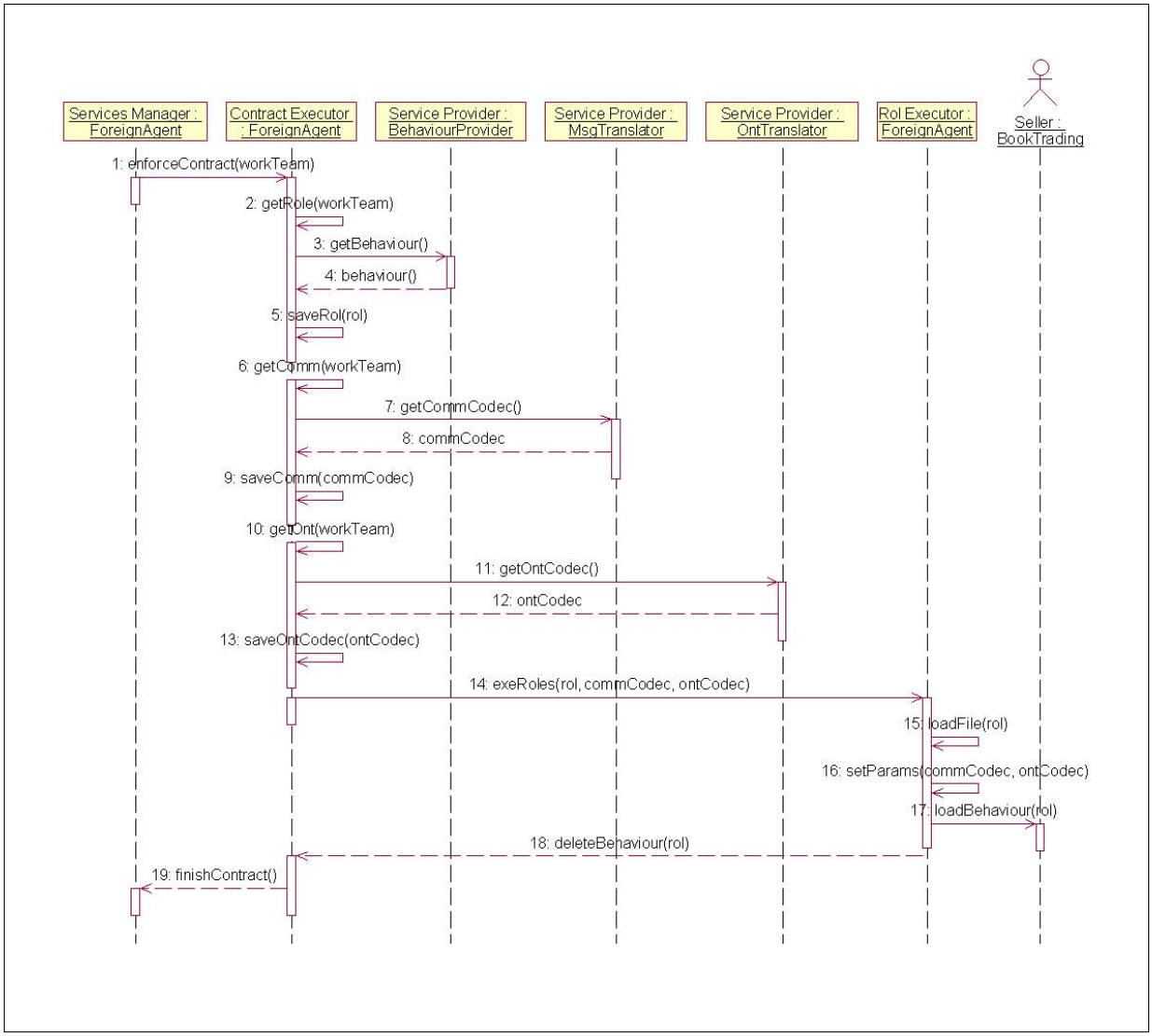


Figure 3: Scenario showing a Foreign Agent in the Contract enforcing process

the behavior parameters with its own attributes and provide the necessary arguments to them. On its turn, the communication codec is used to translate all the inbound/outbound messages among the message protocols used inside the behavior and by the domain MAS.

Regarding the Work Team composition, in a most general case we will be having several ways of composing types of BPs, CMs and STs to conciliate a same FA-to-target-MAS situation. However, due to the combinatorial character of the problem a simplification has been made considering only one possible combination of providers' types. Furthermore, the FA will usually need conciliation at two levels: the interaction and semantics, as usually the behaviors provided will already use the communication protocols employed by the target MAS.

5 Book-Trading Scenario

The study case is based on the well-known book trading MAS environment in which Seller and Buyer agents pursue their goals through the Contract-Net Protocol (CNP) [11]. The book-trading example included in the JADE development has been used as baseline. Buyer agents initiate

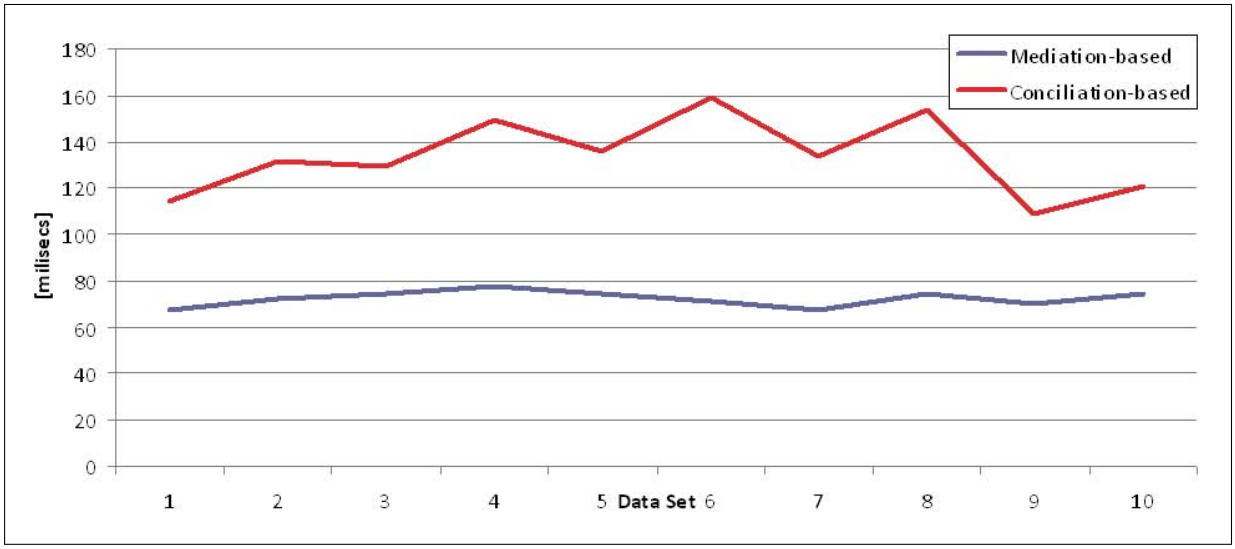


Figure 4: Performance comparison of the mediation-based and conciliation-based architectures

the interaction with a call-for-bids, adopting the Manager role of the CNP. The call includes the book title the buyer is looking for. On the other side, Seller agents assume as potential contractors, processing the calls and making bids for the requested title. In our implementation, the seller looks in its database for the title and sends a bid in case of having stock. The price within the bid is specific for each Seller. The buyer collects the answers and selects the cheapest one.

5.1 Experiments and Results

The objective of the study case was to prove the feasibility of our Dynamic Participation Architecture and evaluate the associated costs mainly in terms of performance. For this, the experiments focused in comparing a traditional book-trading market (as the one above) with a mediation-based approach and with our proposal. In the first case, Buyers and Sellers interact through a Mediator agent while in the last case, buyer agents correspond to Foreign Agents willing to dynamically participate in the book-trading MAS by adopting upon arrival the required behavior. In this particular case, the role requested for conciliation is manager of the CNP with its diverse tasks (e.g. call for proposal, bid selection, task awarding, etc.)

The test scenario has been generated through a random generation of book requests from a list of 30 titles. The number of sellers has been fixed to 10, each of which has one unit of each book in the list at a price that distributes uniformly U[50, 120]. A total of 10 sets of 100 requests has been generated. The simulation considers a main agent devoted to managing the creation of buyer agents, seller agents and the MAS for dynamic participation. The generation and arrival of buyer agents follows a Poisson distribution, hence the time between arrivals distributes Exponential, $\mathsf{E}(\lambda)$, with $\lambda = 2$ in terms of requests per second.

For more details on the architecture design and PASSI-UML diagrams please refer to [5]. Regarding the considered distributed environment, the simulations were carried out over PCs with Intel Pentium 4 of 2 GHz. with 256 MB Ram, connected through a 10/100 Mb. Router.

Figure 4 shows the service mean times for the 10 datasets, measuring the time required for the Buyer agents to buy their books within the book-trading e-market. For our conciliated approach, the graph shows the time spent for the contract-enforcing part only. The time spent in the first part of the process (obtaining a Working Team) was around 6 seconds. The mediated approach

does not consider the time spent in looking for a mediator as it already knows it. These two aspects were not considered in the comparison as a roaming agent will usually carry out many transactions with the same MAS before leaving, hence these initial processes happens only once. In [4] a comparison of a typical closed contract-net-based MAS and our proposal is presented.

6 Conclusions and Future Works

An agent-based software architecture for allowing the dynamic participation of Foreign Agents into an existing MAS has been described. Additionally, an implementation has been carried out under the book-trading domain giving an insight on the viability of the proposal. The use of a mechanism based on dynamic behavior loading (conciliating differences at communication and semantic levels) raises as a feasible approach for obtaining an open MAS system.

Further work considers: extending to other levels of conciliation (e.g. communication protocol stack), different ways of services' delivery (e.g. codec, mediation, tuple-based) and applying our solution to other application domains such as transportation, robotics and supply-chain.

7 Acknowledgement

This work has been partially funded by CONICYT through Fondecyt Project No. 11080284 and the Pontificia Universidad Católica de Valparaíso (www.pucv.cl), through Nucleus Project No. 037.115/2008 "Collaborative Systems".

Bibliography

[1] G. Weiss. *Multiagent Systems: A Modern Approach to Distributed Artificial Intelligence,* MIT Press, Massachusetts, USA. 1999.

[2] M. Luck, P. McBurney, C. Preist, 2003. *Agent Technology: Enabling Next Generation Computing. A Roadmap for Agent Based Computing,* AgentLink II.

[3] P. Burrafato and M. Cossentino. Designing a multiagent solution for a bookstore with the passi methodology. In Fourth International Bi-Conference Workshop on Agent Oriented Information Systems (AOIS-2002).

[4] C. Cubillos, M. Donoso, 2009. Towards Open Agent Systems: A Book-Trading Study Case. In Fourth International Conference on Computer Sciences and Convergence Information Technology (ICCIT '09), pp. 950–953.

[5] C. Cubillos, M. Donoso, D. Cabrera-Paniagua, 2009. Designing an Open Agent System for Book-Trading. Third International Symposium on Intelligent Information Technology Application (IITA 2009), pp. 578–581.

[6] H. Lee, C. Chen, Nov. 2006. Multi-Agent Coalition Formation for Long-Term Task or Mobile Network. Int. Conf. on Computational Intelligence for Modelling, Control and Automation, 2006 and Int. Conf. on Intelligent Agents, Web Technologies and Internet Commerce, pp. 52–57.

[7] M. Klusch, A. Gerber, 2002. Dynamic coalition formation among rational agents.*IEEE Intelligent Systems,* 17(3):42–47.

[8] Foundation for Intelligent Physical Agents (FIPA). 2002. FIPA Interaction Protocols (IPs) Specifications. Available at: http://www.fipa.org/repository/ips.php3.

[9] Y.Labrou, T. Finin, Y. Peng, 1999. Agent communication languages: The current landscape.*IEEE Intelligent Systems,* 14(2), 45–52.

[10] Foundation of Intelligent Physical Agents (FIPA). 2002. FIPA ACL Message Structure Specification. Doc. No. SC00061g, 03/12/2002. Available at: "http://www.fipa.org/specs/fipa00061/"

[11] M. Uschold, R. Jasper and P. Clark. Three Approaches for Knowledge Sharing: A Comparative analysis. In Proceedings of the 12th Knowledge Acquisition, Modelling and Management Workshop, KAW'99, Banff, Canada, October 1999.

[12] FIPA Contract Net Interaction Protocol Specification. Available at: www.fipa.org/specs/fipa00029/SC00029H.pdf

Int. J. of Computers, Communications & Control, ISSN 1841-9836, E-ISSN 1841-9844
Vol. V (2010), No. 5, pp. 684-692

Advanced Information Technology - Support of Improved Personalized Therapy of Speech Disorders

M. Danubianu, S.G. Pentiuc, I. Tobolcea, O.A. Schipor

Mirela Danubianu, Stefan Gheorghe Pentiuc, Ovidiu Andrei Schipor
"Ştefan cel Mare" University of Suceava
Romania, 720229 Suceava, 13 Universităţii
E-mail: {mdanub, pentiuc, schipor}@eed.usv.ro

Iolanda Tobolcea
"Alexandru Ioan Cuza" University of Iaşi
Romania, 700506 Iasi, 11 Bulevardul Carol I
E-mail: itobolcea@yahoo.com

Abstract: One of the key challenges of the Sustainable Development Strategy adopted by the European Council in 2006 is related to public health whose general objective envisages a good level of public health. One of the specific targets includes better treatments of diseases. It is true that there are affections which by their nature do not endanger the life of a person, however they may have a negative impact on her/his life standard. Various language or speech disorders are part of this category, but if they are discovered and treated in due time, they can be often corrected. The difficulty for researchers and therapists is to identify those children who have disorders that show a wide range of issues that cannot be solved spontaneously or which may lead to further significant deficiencies. Information technology in the latest years was used by specialists in order to assist and supervise speech disorder therapy. Consequently they have collected a considerable volume of data about the personal or familial anamnesis, regarding various disorders or regarding the process of personalized therapies. These data can be used in data mining processes that aim to discover interesting patterns which can help the design and adaptation of different therapies in order to obtain the best results in conditions of maximum efficiency. The aim of this paper is to present the Logo-DM system. This is a data mining system that can be associated with TERAPERS system in order to use the data from its database as a source for analysis and to provide new information based on an improved system of therapy. Through the use of appropriate techniques of data mining Logo-DM realizes predictions on the evolution and the final status of patients undergoing therapy and enriches the knowledge data of expert system embedded in TERAPERS.
Keywords: personalized therapy, data mining, classification, clustering, associations rules.

1 Introduction

Various forms of speech disorders affect an important percent of people. There are affections which, by their nature, do not endanger the life of a person, however may have a negative impact on her/his life standard. Discovered and treated in due time, they can be corrected, most often during childhood. The use of information technology in order to assist and supervise speech disorder ther-apy allows specialists to collect a considerable volume of data about the personal or familial anam-nesis, regarding various disorders or regarding the process of personalized therapy.

Even if these data can provide plenty of statistical information little useful knowledge can be obtained from it. In order to get such useful knowledge it is necessary to discover patterns in the data regarding the common characteristics of children with different types of diagnosis, about the connection between antecedents, personal and family behaviour and evolution of the child, or on the connection between the anamnesis and the response to different types of treatments or to different phases of the therapeutic process. These patterns are used to establish such a future strategy so as to maximize the benefits of the therapy and to minimize the costs.

What are the speech disorders? A speech disorder is a problem with fluency, voice, and/or how a person utters speech sounds. Classifying speech into normal and disorder is complex because the statistics points out that only 5% to 10% of the population has a completely normal manner of speaking, all others suffer from one disorder or another. The most common speech disorders are: stuttering, cluttering, voice disorders, dysartria and speech sound disorders. The speech disorder therapy should begin as soon as possible. Children enrolled in therapy early in their development (younger than 5 years) tend to have better outcomes than those who begin therapy later. During the therapy, speech therapists use a variety of strategies including: oral motor or feeding therapy, articu-lation therapy and language intervention activities [2]. During the language intervention activities the therapist will interact with a child by playing and talking. He may use pictures, books, objects, or ongoing events to stimulate language development. The therapist may also model correct pro-nunciation and use repetition exercises to build speech and language skills.

In the area of speech disorders there are some European projects developed as part of the EU Quality of Life and Management of Living Resources program, like: OLP (Ortho-Logo-Paedia) pro-ject [8], STAR - Speech Training, Assessment, and Remediation [12] [19], Speechviewer III developed by IBM [11] or ARTUR (Articulation Tutor) [17] [18]. Currently, the priorities at the international level focus on the development of information systems that can provide a personalised therapy. At the national level, little research has been conducted on the therapy of speech impairments [13]. TERAPERS project [1] [2], developed with the financial support granted by the National Agency for Scientific Research, contract ref. no. 56-CEEX-II03/27.07.2006 by the Research Center for Computer Science in the University "Stefan cel Mare" of Suceava, aims to assist and support the speech disorder therapists in their efforts to develop personalized programs for the therapy of dyslalia.

2 Data mining and its application in logopaedic area

Data mining is defined as the process of discovering non-obvious and potentially useful patterns in large data volumes. As exploration and analysis technique of large amounts of data in order to de-tect patterns or rules with a specific meaning, data mining may facilitate the discovery from appar-ently unrelated data, relationships that can anticipate future problems or might solve the studied problems.

Data mining represents one phase in the complex process of knowledge discovery in databases (KDD) [5]. According to CRISP-DM [15], the reference model for this process, KDD consists of a sequence of steps. These steps are presented in Figure 1.

Using appropriate methods, data mining can solve two broad categories of problems: prediction and description [10] [14]. The most used methods for prediction are classifications and regressions, and for description, clustering, deviation detection or association rules.

The specific logopaedic tasks performed by data mining fall into the following categories [3]:

- classification which places the people with different speech impairments in predefinited classes. Thus it is possible to track the size and structure of various groups. We can use

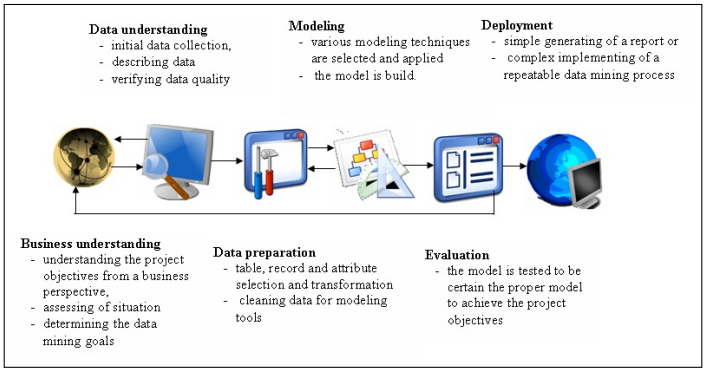


Figure 1: Crisp_DM process of Knowledge Discovery in Databases

classi-fication which is based on the information contained in many predictor variables, such as per-sonal or familial anamnesis data or related to lifestyle, to join the patients with different seg-ments.

- clustering which groups people with speech disorders on the basis of similarity of different features. It is an important task because it helps therapists understand their patients. Cluster-ing aims to finding subsets of a predetermined segment, with homogeneous behavior to-wards various methods of therapy that can be effectively targeted by a specific therapy but it is not based on the previous definition of groups.

- association rules aim to find out associations between different data which seem to have no semantic dependence. It may be a way to determine why a specific therapy program has been successful on a segment of patients with speech disorders and on the other was ineffective.

To conclude with we state that data mining can be a useful tool. Still, there is a limitation we have to consider. Data mining applications generate information by analyzing patterns of data ob-tained from the systems which assist and supervise the speech therapy. Such patterns can help pre-dict the evolution of the individuals that are currently in the process of therapy, or design a scheme of an appropriate therapy for them. However data mining technology can not provide information about impairments, people or behaviors that are not found in the databases that provide data for analysis.

3 Logo-DM System

3.1 Objectives

The idea of trying to improve the quality of logopaedic therapy by applying some data mining tech-niques started from TERAPERS project developed within the Research Center for Computer Sci-ence in the University "Stefan cel Mare" of Suceava. This project has proposed to develop a system which is able to assist speech therapists in their speech therapy of dislalya and to asses how the pa-tients respond to various personalized therapy programs. Starting in March 2008 the system is cur-rently used by the therapists from Regional Speech Therapy Center of Suceava.

At present, because of the limited time and the economical aspects involved, information regard-ing the therapy for each particular case is of interest [4]: what is the predicted final state for a child or what will be his/her state at the end of various stages of therapy, which the best

exercises are for each case and how they can focus their effort to effectively solve these exercises or how the family receptivity - which is an important factor in the success of the therapy - is associated with other as-pects of family and personal anamnesis. All this may be the subject of predictions obtained by ap-plying data mining techniques on data collected by using a computer based therapy system. It is also interesting, as part of the knowledge discovered by data mining algorithms, to be used to enrich the knowledge base of expert system embedded. To achieve this goal we propose the development of Logo-DM system.

Consequently its objectives are:

- analysis of data collected and their preprocessing in order to assure a proper quality for data mining algorithms
- feature selection for the elimination of those irrelevant or redundant
- the use of corresponding data mining methods and algorithms that can be applied in order to find models which can answer to problems raised in speech disorders therapy
- models evaluation and their validation on new cases
- to find new rules which can enrich the knowledge base of the expert system embedded in TERAPERS

3.2 System Architecture

Data mining aims at deriving knowledge from data. The architecture of a data mining system plays an important role in the efficiency with which data is mined. Considering the characteristic of the domain we have proposed for the system a two tier client server architecture. This architecture is presented in Figure 2.

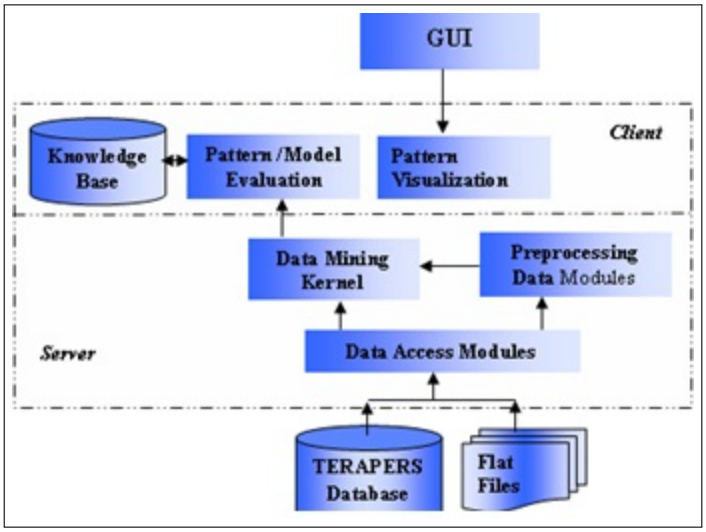


Figure 2: Logo-DM Architecture

On the client side there is the user interface (GUI) which allows the user to communicate with the system in order to select the task to perform, to select and submit the datasets on which data mining needs to be applied. Pattern evaluation and the post-processing step consisting in pattern visualiza-tion are performed also on the client. The knowledge base is the module where the background knowledge is stored.

The more difficult computational tasks of data mining operations are carried out on the server. Here, the data mining kernel contains modules able to perform classifications and association rule detection. Supplementary the pre-processing data module allows data to become suitable for apply-ing data mining algorithms.

3.3 Some aspects regarding the system implementation

It is well known that the best results of data mining algorithms are obtained by applying on data in data warehouses. But in this case the development of a data warehouse is not appropriate, so, it is used, as the primary source of data, a database that contains data collected from the different speech therapists' offices. In order to choose the right solution for the implementation of the system we have made an analysis of available data both its structure and content.

We have started from a scheme with over 60 tables and after deleting tables with irrelevant content for the intended purpose we have obtained, as underlying tables for the final data set, 27 tables as presented in Figure 3.

Content analysis can reveal interesting issues related to data quality or the need for transformation. We have made a first assessment of data quality through the following measures: completeness, conformity, accuracy, consistency and redundancy. The mechanisms provided by the used da-tabase management system have imposed a minimum, controlled redundancy and have assured data consistency. Values stored in fields correspond to reality, but unfortunately in some records useful data for analysis are missing. Therefore it is necessary to supplement data gaps, and where not pos-sible, the removal of the record for accurate results is suggested.

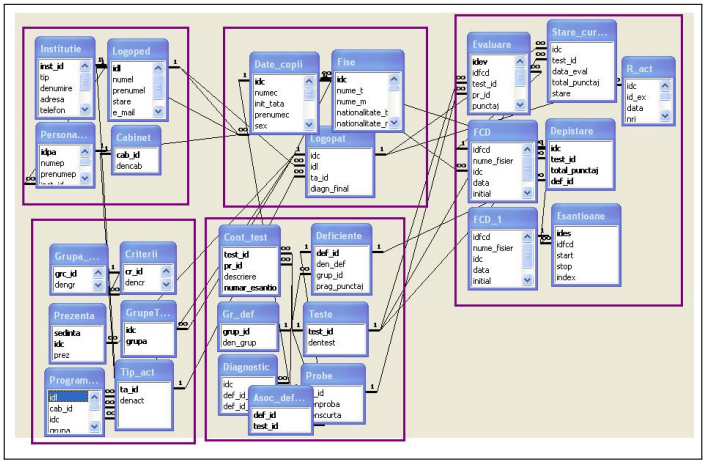


Figure 3: The useful part of database schema

Proper data for the analysis are subjected to the following types of transformation: transformations of the structure, and changes aimed value.

Structural transformations are dictated by the fact that there are fields in the database contain-ing data related to a complex of features to be addressed individually in the analysis. Values of transformations refer to the replacement of coded data by the rules, enabling, for example, the effective storage with descriptive values of characteristics allowing rapid interpretation of results.

An example of these transformations is the following. An issue addressed in the anamnesis form is related to the skills of the child. In Figure 4 we can see that there is a complex of skills of interest (verbal, perceptual, numeric, psycho-motor or special skills).

In the database, all these skills are in two distinct fields: one for general skills, which groups data regarding verbal, perceptual, numeric, psycho-motor and intelligence skills and one for

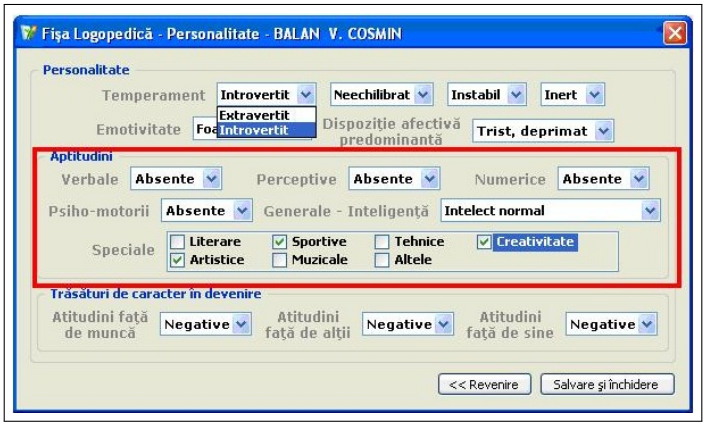


Figure 4: Sample of anamnesis data

special skills (Figure 5). The field called *'aptitudini'* is numeric and is represented in the table by a string of five bits, as shown in Figure 5. These bits, positioned from left to right, have the following meaning:

- the first bit - verbal skills (1- present, 0- absent)
- the second bit - perceptual skills (1- present, 0- absent)
- the third bit - numeric skills (1- present, 0- absent)
- the fourth bit - psycho-motor skills (1- present, 0- absent)
- the fifth bit - intelligence (1- normal intelligence, 0 - mental deficiency)

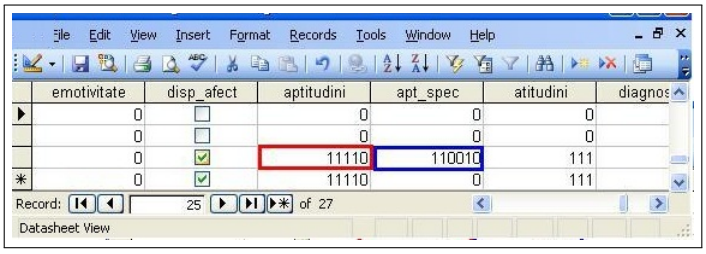


Figure 5: Data to be transformed

Since all these attributes may affect the analysis it is desirable that they can be addressed indi-vidually and explicitly in the final data set. For this purpose the original table structure is changed and values are converted to descriptive values as in Figure 6.

These changes have conducted to a modified form of the relational database used by Terapers. In the first phase, construction of target data sets for each of the methods to be applied in the system is through the application of relational expressions like those presented in (1).

$$\prod_{I_i} (T_1 \triangleright \triangleleft T_2 \triangleright \triangleleft \ldots \triangleright \triangleleft T_k) \tag{1}$$

where:

- Ii is a superset of the attributes regarding the useful characteristics for each method
- $T_1 \ldots T_k$ is the set of tables containing the attributes in the list of projection.

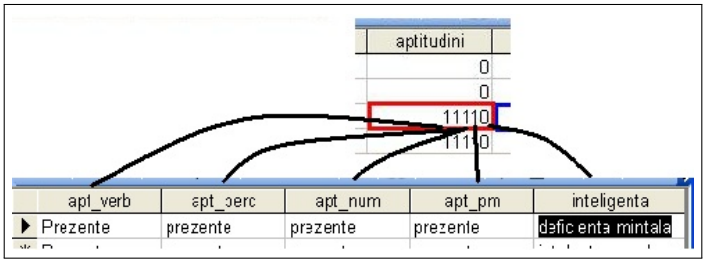


Figure 6: Transformed data

Each of these expressions was implemented in SQL, and has generated intermediate tables. For example, the target data set necessary to establish the profile of children with speech disorders, can be obtained by joining tables which contain: general data about children, family and personal an-amnesis, data on complex evaluation and diagnosis associated. The statement that performs that is presented in (2). The result is a table that contains 129 features.

```
create table caract_copii as
    select f.*, l.diagn_final
    from fise f, logopat l
    where f.idc = l.idc;
```
(2)

Data mining techniques were not designed to process large amounts of irrelevant features. Conse-quently before their application, a selection of the relevant features is required [6] [7]. The most im-portant objectives of feature selection are: to avoid over fitting and improve model performance. A variant of the mRMR method [9] for categorical values has been used for feature selection. It is based on mutual information criteria, formally defined, for two discrete random variables X and Y, as:

$$I(X;Y) = \sum_{y \in Y} \sum_{x \in X} p(x,y) \log \left(\frac{p(x,y)}{p_1(x) p_2(y)} \right) \tag{3}$$

where p(x,y) is joint probability distribution function of X and Y, and p1(x) and p2(y) are the marginal probability distribution functions of X and Y respectively.

For discrete random variable, the joint probability mass function is:

$$p(x,y) = p(X = x, Y = y) = p(Y = y | X = x) * p(X = x) = p(X = x | Y = y) * p(Y = y) \tag{4}$$

Since these are probabilities, we have

$$\sum_{x} \sum_{y} p(X = x, Y = y) = 1 \tag{5}$$

The marginal probability function, p (X = x) is:

$$p(X = x) = \sum_{y} p(X = x, Y = y) = \sum yp(X = x | Y = y) p(Y = y) \tag{6}$$

The criterion used is related to minimizing redundancy and maximizing relevance to the chosen characteristics. The result of tests performed on data prepared as described in the example mentioned above, revealed that, for classification, the minimum error is obtained if we deal with a number between 20 and 22 features selected. The target data set, obtained after these steps, is

subject to data mining algorithms. For an effective implementation of algorithms we have taken into account, and we tested, two possibilities: to use the Oracle Data Mining kernel (ODM) which offers the possibility to apply algorithms for classification, clustering and association rules and to use some open source implementations of relevant algorithms adapted and integrated into our own system.

We took into account the types of data included in the set and we used implementations in Oracle of Adaptive Bayes Network, Seeker Model and decision trees build with CART [16] and ID3/C4.5 for classification, for clustering the Oracle implementation of A-Clustering algorithm and for association rules Apriori algorithm. It should be noted that for the moment, the volume of data on which work is relatively low, because the system which is the main source of these data is operational for only several months.

4 Conclusions and Future Works

Considering the opportunity of data mining techniques application on data collected in the process of speech therapy, we have concluded that methods such as classification, clustering or as-sociation rules can provide useful information for a more efficient therapy. Consequently, we have designed and we are currently implementing a data mining system that aims to use data provided by TERAPERS system, developed by the Research Center for Computer Science in the University "Stefan cel Mare" of Suceava, in order to achieve an optimized personalized therapy of dyslalia. We have tested the modules for data pre-processing and on target data sets obtained from these modules we have applied more algorithms for detecting the most appropriate solutions for the data mining kernel. At present efforts are directed towards the implementation of evaluation patterns and visualization modules and towards building a user friendly interface.

Bibliography

[1] M. Danubianu, S.G. Pentiuc, O. Schipor, I. Ungureanu, M. Nestor, Distributed Intelligent System for Personalized Therapy of Speech Disorders, in Proc. of *The Third International Multi-Conference on Computing in the Global Information Technology ICCGI*, July 27- August 01, Athens, Greece, 2008.

[2] M. Danubianu, S.G. Pentiuc, O. Schipor, M. Nestor, I. Ungurean, D.M. Schipor, TERAPERS - Intelligent Solution for Personalized Therapy of Speech Disorders, *International Journal on Ad-vances in Life Science*, p.26-35, 2009.

[3] M. Danubianu, T. Socaciu, Does Data Mining Techniques Optimize the Personalized Therapy of Speech Disorders?, *Journal of Applied Computer Science and Mathematics*, p.15-19, 2009

[4] M. Danubianu, S.G. Pentiuc, T. Socaciu, Towards the Optimized Personalized Therapy of Speech Disorders by Data Mining Techniques, *The Fourth International Multi Conference on Computing in the Global Information Technology ICCGI 2009*, Vol: CD, 23-29 August, Cannes - La Bocca, France, 2009

[5] F.G. Filip, *Decizie asistata de calculator*, Ed. Tehnica, Bucuresti, 2005

[6] I. Guyon, A. Elisseeff, An introduction to variable and feature selection. *J. Mach Learn Res.*, 3, p.1157-1182, 2003

[7] H. Liu, H. Motoda, *Feature Selection for Knowledge Discovery and Data Mining.* Kluwer Aca-demic Publishers, Norwell, MA, 1998

[8] OLP (Ortho-Logo-Paedia) - Project for Speech Therapy (http://www.xanthi.ilsp.gr/olp);W.-K. Chen, Linear Networks and Systems (Book style).Belmont, CA: Wadsworth, p. 123-135, 1993

[9] H. Peng, F. Long, C. Ding, Feature Selection Based on Mutual Information: Criteria of Max-Dependency, Max-Relevance, and Min-Redundancy, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 27, No. 8, p. 1226-1238, 2005

[10] B. Reiz, L. Csató, Bayesian Network Classifier for Medical Data Analysis. *International Journal of Computers Communications & Control* Vol. 4, p: 65-72, 2009

[11] Speechviewer III - (http://www.synapseadaptive.com/edmark/prod/sv3)

[12] STAR Speech Training, Assessment, and Remediation (http://www.asel.udel.edu/speech)

[13] Tobolcea,I., *Interventii logo-terapeutice pentru corectarea formelor dislalice la copilul normal*, Editura Spanda, Iasi, 2002.

[14] P. Wessa, Quality Control of Statistical Learning Environments and Prediction of Learning Outcomes through Reproducible Computing, *International Journal of Computers Communications & Control* Vol. 4, p: 185-197, 2009

[15] R. Wirth, J. Hipp, CRISP-DM: Towards a standard process model for data mining. *In Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining*, pages 29-39, Manchester, UK, 2000

[16] www.salford-systems.com/
last visited October 2009

[17] www.speech.kth.se/multimodal/ARTUR/index.html
last visited August 2009

[18] O. Balter, O. Engwall, A.M. Oster, H. Kjellstrom, Wizard-of-Oz Test of ARTUR - a Computer-Based Speech Training System with Articulation Correction. *Proceedings of the Seventh International ACM SIGACCESS Conference on Computers and Accessibility*, Baltimore, October, 2005, pp.36-43.

[19] H.T. Bunnel, M.D. Yarrington, B.J. Polikoff, Articulation Training for Young Children, *Proceedings of 6th International Conference on Spoken Language Processing (ICSLP 2000)*, Beijing, China, October 16-20, 2000, vol.4, pp. 85-88.

Int. J. of Computers, Communications & Control, ISSN 1841-9836, E-ISSN 1841-9844
Vol. V (2010), No. 5, pp. 693-700

Meta-Rationality in Normal Form Games

D. Dumitrescu, R.I. Lung, T.-D. Mihoc

Dan Dumitru Dumitrescu, Rodica Ioana Lung, Tudor Dan Mihoc
"Babes Bolyai" University
Romania, Cluj-Napoca, St. Universitatii 5,
E-mail: {ddumitr, mihoct}@cs.ubbcluj.ro, rodica.lung@econ.ubbcluj.ro

Abstract: A new generative relation for Nash equilibrium is proposed. Different types of equilibria are considered in order to incorporate players different rationality types for finite non cooperative generalized games with perfect information. Proposed equilibria are characterized by use of several generative relations with respect to players rationality. An evolutionary technique for detecting approximations for equilibria is used. Numerical experiments show the potential of the method.
Keywords: non-cooperative games, evolutionary equilibrium detection, generative relations, Nash-Pareto, meta-strategy.

1 Introduction

The most common solutions proposed in Game Theory are the equilibrium concepts. Within the present day approaches each equilibrium concept is addressed separately, meaning that in a particular game players interact accordingly to a unique equilibrium concept. This restriction induces unrealistic results. For example, the concept of Nash equilibrium, alone, sometimes can lead to deceptive results so we need to cope with more complex situations.

In real life players can be more or less cooperative, more or less competitive and more or less rational, therefore agents guided by the different kind of equilibrium concepts should be allowed to interact.

We consider a generalized game where players are allowed to have different behaviours according to their rationality type. Players can have different behaviours/rationality types resulting in an adequate meta-strategy concept.

Game equilibria can be characterized using appropriate generative relations [4]. Thus Nash equilibrium is characterized by the ascendancy relation [6] and Pareto equilibrium by the Pareto domination. Combining the two relations may lead to different types of joined Nash–Pareto equilibria.

We introduce a new generative relation for Nash equilibrium and we use it to compose a new joined Nash-Pareto equilibria.

An evolutionary technique for detecting the two joined Nash–Pareto equilibria for generalized games is used.

2 Generalized games

In order to cope with different rationality types the concept of generalized game is defined [4].

Definition 1. A finite strategic *generalized game* is defined as a system by $\mathsf{G} = (\mathsf{N}, \mathsf{M}, \mathsf{U})$ where:

- $\mathsf{N} = \{1, ..., n\}$, represents the set of players, n is the number of players;

- for each player $i \in N$, S_i represents the set of actions available to him, $S_i = \{s_{i_1}, s_{i_2}, ..., s_{i_{m_i}}\}$; $S = S_1 \times S_2 \times ... \times S_n$ is the set of all possible situations of the game;
- for each player $i \in N$, M_i represents the set of available meta-strategies, a meta-strategy is a system $(s_i|r_i)$ where $s_i \in S_i$ and r_i is the i^{th} player rationality type;
- $M = M_1 \times M_2 \times ... \times M_N$ is the set of all possible situations of the generalized game and $(s_1|r_1, s_2|r_2, ..., s_n|r_n) \in M$ is a meta-strategy profile.
- for each player $i \in N$, $u_i : S \to \mathbf{R}$ represents the payoff function.

$$U = \{u_1, ..., u_n\}.$$

Remark 2. In a generalized game the set of all possible meta-strategies represents the meta-strategy search space.

The rationality type of a player usually represents the player bias towards a certain equilibrium.

3 Generative relations for generalized games

Three generative relations are considered in this section. Two of them correspond to Pareto and Nash equilibria. The third induces a new type of joined Nash–Pareto equilibrium.

3.1 n_P–strict Pareto domination

We consider the n_P–strict Pareto domination in order to be able to combine several concepts of Nash and Pareto domination.

In a finite strategic generalized game consider the set of players Pareto biased

$$I_P = \{j \in \{1, ..., n\} | r_j = \text{Pareto}\}$$

and $n_P = \mathtt{card} I_P$.

Let us consider two meta strategy profiles x and y from M.

Definition 3. The meta strategy profile x n_P–strict Pareto dominates the meta strategy profile y if the payoff of each Pareto biased player from I_P using meta strategy x is strictly greater than the payoff associated to the meta strategy y, i.e.

$$u_i(x) > u_i(y),\ \forall i \in I_P.$$

Remark 4. The set of non dominated meta strategies with respect to the n_P–strict Pareto domination relation when $n_P = n$ is a subset of the Pareto front.

3.2 Nash - ascendancy

Similar to Pareto equilibrium a particular relation between strategy profiles can be used in order to describe Nash rationality. This relation is called Nash-ascendancy (NA).

A strategy is called Nash equilibrium [5] if each player has no incentive to unilaterally deviate i.e. it can not improve the payoff by modifying its strategy while the others do not modify theirs.

We denote by (s_{i_j}, s^*_{-i}) the strategy profile obtained from s^* by replacing the strategy of player i with s_{i_j} i.e.

$$(s_{i_j}, s^*_{-i}) = (s^*_1, s^*_2, ..., s^*_{i-1}, s_{i_j}, s^*_{i+1}, ..., s^*_n).$$

Definition 5. The strategy profile x Nash-ascends the strategy profile y, and we write $x <_{NA} y$ if there are less players i that can increase their payoffs by switching their strategy from x_i to y_i then vice versa.

In [6] is introduced an operator

$$k : S \times S \to \mathbf{N},$$

$$k(y, x) = \mathtt{card}\{i \in \{1, ..., n\} | u_i(x_i, y_{-i}) \geq u_i(y), x_i \neq y_i\}.$$

$k(y, x)$ denotes the number of players which benefit by switching from y to x.

Proposition 6. *The strategy x Nash-ascends y (x is NA-preferred to y), and we write $x <_{NA} y$, if the inequality*

$$k(x, y) < k(y, x),$$

holds.

According to [6] the set of all strategies from S non-dominated by respect of Nash ascendancy relation equals the set of Nash equilibria.

This result proves that the Nash ascendancy is the generative relation for the Nash equilibrium.

3.3 Differential generative relation of Nash equilibrium (DGN)

A new generative relation for Nash equilibrium is proposed. This relation relies on the payoff difference between perturbed and non perturbed strategies.

We introduce the measure

$$m(y, x) = \sum_{i \in N} (u_i(x_i, y_{-i}) - u_i(y))$$

Definition 7. The strategy x dominates y, and we write $x <_{DGN} y$, if the inequality

$$m(x, y) < m(y, x),$$

holds.

3.4 Joint Nash–Pareto domination

Let us consider two meta-strategies $x = (x_1|r_1, x_2|r_2, ..., x_n|r_n)$ and $y = (y_1|r_1, y_2|r_2, ..., y_n|r_n)$.

Let us denote by I_N the set of Nash biased players (N-players) and by I_P the set of Pareto biased players (P-players). Therefore we have $I_N = \{i \in \{1, ..., n\} | r_i = \text{Nash}\}$.

We consider the operators k_P and k_N defined as:
$k_P(x, y) = \mathtt{card}\{j \in I_P | u_j(x) > u_j(y), x \neq y\}$ and respectively $k_N(x, y) = \mathtt{card}\{i \in I_N | u_i(y_i, x_{-i}) \geq u_i(x), x_i \neq y_i\}$.

Remark 8. $k_P(x, y)$ measures the *relative efficiency* of the meta strategies x and y with respect to Pareto rationality and $k_N(x, y)$ measures the *relative efficiency* of the meta strategies x and y with respect to Nash rationality.

Definition 9. The meta strategy x N–P dominates the meta strategy y if and only if the following statements hold

1. $k_P(x, y) = n_P$
2. $k_N(x, y) < k_N(y, x)$

In what follows we consider that efficiency relation induces a new type of equilibrium called *joined Nash-Pareto equilibrium.*

3.5 Joint Differential Nash Pareto domination

A new domination relation with respect to Nash-Pareto equilibrium is introduced by using differential generative relation of Nash equilibrium.

Definition 10. The meta strategy x DGN–P dominates the meta strategy y if and only if the following statements hold

1. $k_P(x, y) = n_P$
2. $m(x, y) < m(y, x)$

4 Detecting joint N–P equilibria in generalized games

Consider a three player non-cooperative game. Let r_i be the rationality type of player i. If $r_1 = r_2 = r_3 =$ Nash then all players are Nash biased and the corresponding solution concept is the Nash equilibrium. If $r_1 = r_2 = r_3 =$ Pareto then all players are Pareto biased and the corresponding equilibria are described by the set of strictly non dominated strategies (Pareto front).

We also intend to explore the joint cases where one of the players is Nash biased and others are Pareto and the one where one is Pareto and the others are Nash biased.

In order to detect the joined Nash–Pareto equilibria of the generalized game an evolutionary approach is used. For a certain equilibrium the corresponding generative relation allows the comparison of two meta-strategies. This comparison may guide the search towards the game equilibrium.

Let us consider an initial population of meta strategies for the generalized three player game. Each member of the population has the form

$$x = (s_1|r_1, s_2|r_2, s_3|r_3).$$

Non domination (with respect to a generative relation) is considered for fitness assignment purposes. Evolutionary Multiobjective Optimization Algorithms [3] are efficient tools for evolving strategies based on a non domination relation.

The state of the art NSGA2 [2] has been considered to illustrate how generative relations can be used for evolutionary detection of proposed equilibria.

A population of 100 strategies has been evolved using a rank based fitness assignment technique. In all experiments the process converges in less than 30 generations.

5 Numerical experiments

In order to illustrate the proposed concepts the oligopoly Cournot model is considered (see for instance [4]).

Let q_1, q_2 and q_3 denote the quantities of an homogeneous product - produced by three companies respectively. The market clearing price is $P(Q) = a - Q$, where $Q = q_1 + q_2 + q_3$, is the aggregate quantity on the market. Hence we have

$$P(Q) = \begin{cases} a - Q, & \text{for } Q < a, \\ 0, & \text{for } Q \geq a. \end{cases}$$

Let us assume that the total cost for the company i of producing quantity q_i is $C_i(q_i) = c_i q_i$. Therefore, there are no fixed costs and the marginal cost c_i is constant, $c_i < a$. Suppose that

the companies choose their quantities simultaneously. The payoff for the company i is its profit, which can be expressed as:

$$\begin{aligned} \pi_i(q_i, q_j) &= q_i P(Q) - C_i(q_i) \\ &= q_i \left[a - (q_i + q_j) - c_i\right]. \end{aligned}$$

Several experiments have been performed for this game by using RED technique [4].

The symmetric Cournot model with parameters $a = 24$ and $c_1 = c_2 = c_3 = 9$ is considered.

According to the data from the Table 1 in less than 30 generations the algorithm converges to the Nash equilibrium point (14.00, 14.00, 14.00) for each relation. We observe that the differential Nash domination provides more accurate results than the Nash ascendency. We must consider however the particular nature of this Cournot game. For other types of games a normalisation of the deviations must be done in order to sum them.

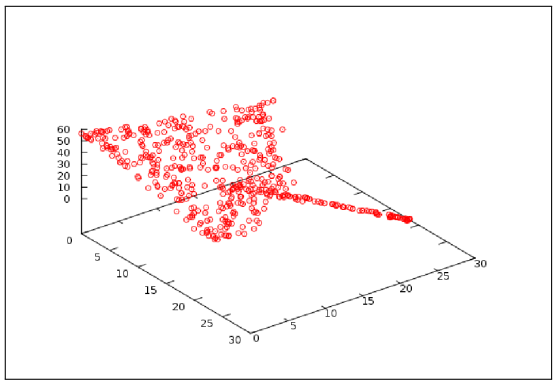


Figure 1: The payoffs for the Nash-Nash-Pareto front detected in less than 30 iterations for the symmetric Cournot game with the Nash–Pareto generative relation

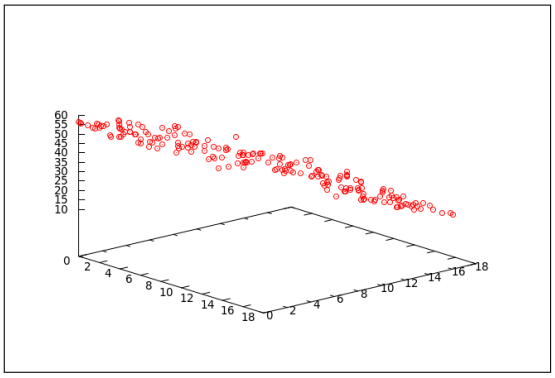


Figure 2: The payoffs for the Nash-Nash-Pareto front detected in less than 30 iterations for the symmetric Cournot game with the differential Nash–Pareto generative relation

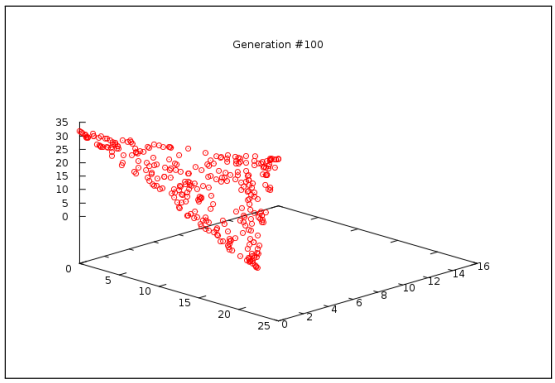


Figure 3: The payoffs for the Nash-Pareto-Pareto front detected in less than 30 iterations for the symmetric Cournot game with the Nash–Pareto generative relation.

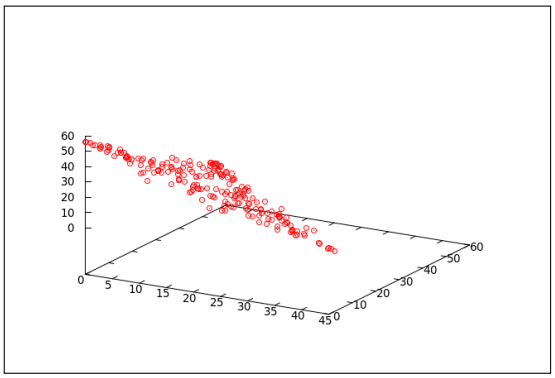


Figure 4: The payoffs for the Nash-Pareto-Pareto front detected in less than 30 iterations for the symmetric Cournot game with the differential Nash–Pareto generative relation.

The resulting front in the Nash-Nash-Pareto case spreads from the standard Nash equilibrium corresponding to the two player–Cournot game (25.00, 25.00) to the Nash equilibrium corresponding to the three player–Cournot game, and from there to the edges of Pareto front for

Table 1: Average payoff and standard deviation of the final populations in 30 runs with 100 meta-strategies after 30 generations for the symmetric Cournot model where all three players are Nash biased using Nash ascendency and differential Nash generative relations.

N-N-N	Average payoff			St. dev.			Maximum payoff			Minimum payoff		
player	p1	p2	p3	p1	p2	p3	p1	p2	p3	p1	p2	p3
	Nash ascendency relation											
Average	14.05	14.06	14.05	0.03	0.04	0.04	14.85	15.57	15.00	12.25	12.49	12.45
St. Dev.	0.02	0.02	0.02	0.08	0.09	0.08	1.39	2.80	1.83	3.25	3.00	3.05
	Differential Nash relation											
Average	14.06	14.06	14.06	0.00	0.00	0.00	14.06	14.06	14.06	14.06	14.06	14.06
St. Dev.	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Table 2: Average payoff and standard deviation of the final populations in 30 runs with 100 meta-strategies after 30 generations for the symmetric Cournot model where two player are Nash biased and one is Pareto for both joint Nash–Pareto and joint Differential Nash–Pareto generative relations.

N-N-P	Average payoff			St. dev.			Maximum payoff			Minimum payoff		
player	p1	p2	p3	p1	p2	p3	p1	p2	p3	p1	p2	p3
	Joint Nash–Pareto relation											
Average	10.99	11.01	29.80	52.81	53.02	182.28	25.92	25.71	56.24	0.00	0.00	0.49
St. Dev.	0.36	0.33	0.78	1.75	2.33	17.62	0.92	0.88	0.00	0.00	0.00	1.67
	Joint Differential Nash–Pareto relation											
Average	8.50	8.48	35.53	22.76	22.67	165.84	18.47	18.92	56.24	0.00	0.00	7.60
St. Dev.	0.35	0.31	0.85	0.25	0.25	0.54	2.88	2.93	0.00	0.00	0.00	5.60

the Nash–Pareto equilibria (see Figure 1). For differential Nash-Pareto (see Figure 2) the front spreads from vicinity of the Nash equilibrium for Cournot game to the edge of the Pareto front corresponding to the Pareto player. The numerical results are presented in Table 2.

As we can see in the Figure 3 in the Nash-Pareto-Pareto case for Nash–Pareto generative relation the result is similar to the Pareto front. In the same case for differential Nash–Pareto generative relation (Figure 4) the Pareto front is deformed in the Nash player's corresponding edge. The numerical results are presented in Table 3.

6 Conclusions and future work

A new generative relation for Nash equilibrium based on differences between perturbations is introduced. Generative relations between meta strategies induce corresponding solutions concepts named Joined Nash–Pareto equilibrium, respectively joint differential Nash–Pareto equilibrium.

An evolutionary technique for detecting approximations of the generalized equilibria is used. The ideas are exemplified for Cournot games with three players and two types of rationality.Results indicate the potential of the proposed technique.

Future work will address generalized games having other rationality types and other methods

Table 3: Average payoff and standard deviation of the final populations in 30 runs with 100 meta-strategies after 30 generations for the symmetric Cournot model where one player is Nash biased and the other two Pareto for both joint Nash–Pareto and joint Differential Nash–Pareto generative relations.

N-P-P	Average payoff			St. dev.			Maximum payoff			Minimum payoff		
player	p1	p2	p3	p1	p2	p3	p1	p2	p3	p1	p2	p3
	Joint Nash–Pareto relation											
Average	17.74	18.52	18.44	242.42	247.83	247.97	56.23	56.24	56.24	0.00	0.00	0.00
St. Dev.	0.40	0.36	0.42	7.36	6.98	6.82	0.04	0.00	0.00	0.00	0.00	0.00
	Joint Differential Nash–Pareto relation											
Average	15.13	19.83	19.76	154.37	250.62	248.44	48.90	56.24	56.24	0.00	0.00	0.00
St. Dev.	0.86	0.77	0.56	0.78	0.29	0.32	3.02	0.00	0.01	0.00	0.00	0.00

of combining them.

7 Acknowledgements

This research is supported partially by the CNCSIS Grant ID508 *"New Computational paradigms for dynamic complex problems"* funded by the MEC and from the SECTORAL OPERATIONAL PROGRAMME HUMAN RESOURCES DEVELOPMENT, Contract POSDRU 6/1.5/S/3 *"Doctoral studies: through science towards society"*, Babeş - Bolyai University, Cluj - Napoca, România.

Bibliography

[1] Bade, S., Haeringer, G., Renou, L.: *More strategies, more Nash equilibria*, Working Paper 2004-15, School of Economics University of Adelaide University, 2004.

[2] Deb, K., Agrawal, S., Pratab, A., Meyarivan, T.: *A Fast Elitist Non-Dominated Sorting Genetic Algorithm for Multi-Objective Optimization: NSGA-II*, Marc Schoenauer, Kalyanmoy Deb, Günter Rudolph, Xin Yao, Evelyne Lutton, Juan Julian Merelo, and Hans-Paul Schwefel, editors, Proceedings of the Parallel Problem Solving from Nature VI Conference, Paris, France, 2000. Springer, Lecture Notes in Computer Science, 1917, 849-858.

[3] Deb, K.: *Multi-objective optimization using evolutionary algorithms*, Wiley, 2001.

[4] Dumitrescu, D., Lung, R.I., Mihoc, T.D.: *Evolutionary Equilibria Detection in Non-cooperative Games*, Book Series: LNCS, Publisher Springer Berlin / Heidelberg, Volume 5484 / 2009, Book: Applications of Evolutionary Computing, 2009, 253-262.

[5] Lung, R. I., Muresan, A. S., and Filip, D. A.: *Solving multi-objective optimization problems by means of natural computing with application in finance*, In Aplimat 2006 (Bratislava, February 2006), pp. 445-452.

[6] Lung, R., I., Dumitrescu, D.: *Computing Nash Equilibria by Means of Evolutionary Computation*, Int. J. of Computers, Communications & Control, 2008, 364-368

[7] Maskin, E. : *The theory of implementation in Nash equilibrium:A survey*, in: L. Hurwicz, D. Schmeidler and H. Sonnenschein, eds., Social Goals and Social Organization (Cambridge University Press), 1985,173-204

[8] McKelvey, R., D., McLennan, A.: *Computation of equilibria in finite games*, In H. M. Amman, D. A. Kendrick, and J. Rust, editors, Handbook of Computational Economics, Elsevier,1996.

[9] Nash.,J.,F.: *Non-cooperative games*, Annals of Mathematics, 54:286-295, 1951.

[10] Osborne, M. J., Rubinstein, A.: *A Course in Game Theory*, MIT Press, Cambridge, MA, 1994

Int. J. of Computers, Communications & Control, ISSN 1841-9836, E-ISSN 1841-9844
Vol. V (2010), No. 5, pp. 701-709

Stable Factorization of Strictly Hurwitz Polynomials

Ö. Eğecioğlu, B. S. Yarman

Ömer Eğecioğlu
Department of Computer Science
University of California, Santa Barbara
CA 93106, USA
E-mail: omer@cs.ucsb.edu

B. Siddik Yarman
Department of Electric and Electronics Engineering
College of Engineering, Istanbul University
34320 Avcilar, Istanbul, Turkey
E-mail: sbyarman@gmail.com

Abstract: We propose a stable factorization procedure to generate a strictly Hurwitz polynomial from a given strictly positive even polynomial. This problem typically arises in applications involving real frequency techniques. The proposed method does not require any root finding algorithm. Rather, the factorization process is directly carried out to find the solution of a set of quadratic equations in multiple variables employing Newton's method. The selection of the starting point for the iterations is not arbitrary, and involves interrelations among the coefficients of the set of solution polynomials differing only in the signs of their roots. It is hoped that this factorization technique will provide a motivation to perform the factorization of two-variable positive function to generate scattering Hurwitz polynomials in two variables for which root finding methods are not applicable.
Keywords: Routh-Hurwitz stability, Hurwitz polynomial, stable factorization, Newton's method.

1 Introduction

In many microwave communication system design, modeling and simulation problems, description of lossless two ports in one or two kinds of elements is essential [5]. In the design of microwave matching networks, amplifiers or in modeling passive one port devices such as antennas, lossless two ports are either described in terms of driving point immitance or reflectance functions [6,7]. The methods known as Real Frequency Techniques (RFT) are excellent tools for design and modeling [5,8]. Once the independent descriptive parameters are selected, numerical implementations of real frequency techniques demands the construction of strictly Hurwitz polynomials. For example, in the simplified real frequency technique (SRFT), the numerator polynomial $h(p) = h_0 + h_1 p + \cdots + h_n p^n$ of the driving point input reflectance $S_{11}(p) = \frac{h(p)}{g(p)}$ completely specifies the scattering parameters of the lumped element reciprocal lossless two port as follows:

$$S_{12} = S_{21} = \frac{f(p)}{g(p)} \quad \text{and} \quad S_{22} = \frac{f(p)}{f(-p)} \frac{h(-p)}{g(p)} \tag{1}$$

provided that the monic-polynomial $f(p)$ which is constructed on the transmission zeros of the system under consideration, is pre-selected. In this representation, the denominator polynomial $g(p) = g_0 + g_1 p + \cdots + g_n p^n$ is generated as a strictly Hurwitz polynomial from the equation

$$G(p^2) = g(p)g(-p) = h(p)h(-p) + f(p)f(-p) = G_0 + G_1 p^2 + \cdots + G_n p^{2n} \tag{2}$$

which is obtained by means of the lossless condition. Once $f(p)$ is selected, (2) is specified in terms of the real coefficients $\{h_0, h_1, \ldots, h_n\}$ of $h(p)$. For many practical problems, it may be sufficient to choose $f(p)$ as $f(p) = p^k, k \leq n$. In this case, (2) results in a set of quadratic equations such that

$$\begin{aligned}
G_0 &= g_0^2 = h_0^2 \\
G_1 &= -g_1^2 + 2g_0g_2 = -h_1^2 + 2h_0h_2 \\
&\vdots \\
G_i &= (-1)^i g_i^2 + 2(g_{2i}g_0 + \sum_{j=2}^{i}(-1)^j g_{j-1}g_{2j-i+1}) = (-1)^i h_i^2 + 2(h_{2i}h_0 + \sum_{j=2}^{i}(-1)^j h_{j-1}h_{2j-i+1}) \\
&\vdots \\
G_k &= G_{(i=k)} + (-1)^k \\
&\vdots \\
G_n &= (-1)^n g_n^2 = (-1)^n h_n^2
\end{aligned} \tag{3}$$

It should be mentioned that the general form of $f(p)f(-p)$ may be described as

$$F(p^2) = f(p)f(-p) = F_0 + F_1p^2 + \cdots + F_np^{2n}. \tag{4}$$

Then it is straightforward to revise (4.3) with the help of (4.4). At this point it is the crucial issue to generate $g(p)$ as a strictly Hurwitz polynomial either employing (2) or (4.3). If one employs (2), it is sufficient to find the roots of $G(p^2)$ and then, construct $g(p)$ on the left halfplane roots of $G(p^2)$, yielding $g(p) = g_0 + g_1p + \cdots + g_np^n$. This has been the common practice of the SRFT. However, if the problem under consideration demands the construction of lossless two-ports with two kinds of elements, then there is no way to carry out the computation by means of root finding techniques. In this case, one has to rewrite (2) in two variables as

$$G(p, \lambda) = g(p, \lambda)g(-p, -\lambda)$$

and revise (4.3) accordingly. Eventually one needs to solve (4.3) to generate $g(p, \lambda)$ as a "two variable scattering Hurwitz polynomial" [1,2]. In this representation the complex variable $p = \sigma + j\omega$ is associated with first kind of elements and the complex variable $\lambda = \Sigma + j\Omega$ is associated with the second kind of elements of the lossless two-port. Actually, this way of posing the problem may be understood as the factorization of the two variable polynomial $G(p, \lambda)$ as $g(p, \lambda)g(-p, -\lambda)$, which in turn yields the scattering Hurwitz polynomial $g(p, \lambda)$. Based on the knowledge of the authors, there is no explicit solution for the factorization of two variable polynomials in the current literature. However, for the single variable case, root finding techniques provide excellent results as described within SRFT. Therefore, in this paper, to provide an insight to the general factorization problem, an attempt will be made to come up with a numerical procedure to solve (4.3) which is specified in single variable, with the hope that the numerical procedure presented in this paper may be extended to cover the two variable factorization case.

2 Mathematical problem statement

Let $G(z^2) = G_0 + G_1z^2 + G_2z^4 + \cdots + G_nz^{2n}$ be a real polynomial with $G_0 > 0$. Consider a factorization of G of the form

$$G(z^2) = g(z)g(-z) \tag{5}$$

for a real polynomial $g(z) = g_0 + g_1 z + g_2 z^2 + \cdots + g_n z^n$ as required in (4.3). Call (5) a *stable factorization* of G, if the polynomial g is stable: that is, the real parts of the zeros of g are strictly negative. We also refer to a stable polynomial as *strictly Hurwitz*. From physical considerations that give rise to the problem, $G_0, G_1, \ldots, G_n$ are such that G admits a stable factorization. Our aim is to determine the coefficients of $g(z)$ as a function of $G_0, G_1, \ldots, G_n$.

2.1 On root finding

This one dimensional problem is theoretically solvable quite easily by root finding: Since G is a real polynomial, it can be factored as

$$G(z^2) = c(z^2 - \alpha_1)(z^2 - \alpha_2)\cdots(z^2 - \alpha_n)$$

with $c > 0$ and the α_i complex. For $i = 1, 2, \ldots, n$, let $\beta_i = \pm\sqrt{\alpha_i}$, where the sign is picked so that β_i has a negative real part. Then $g(z) = \sqrt{c}(z - \beta_1)(z - \beta_2)\cdots(z - \beta_n)$, and the g_i can be computed from this product. However, we wish to avoid this approach as the real motivation behind the treatment of the one variable case is the factorization problem in two variables to generate scattering Hurwitz polynomials, for which root finding techniques do not apply.

2.2 Basic elements of Routh-Hurwitz stability

The conditions for a real polynomial

$$g(z) = g_0 + g_1 z + g_2 z^2 + \cdots + g_n z^n \tag{6}$$

with $g_0 > 0$ to be strictly Hurwitz are given in terms of the positivity of the Hurwitz determinants

$$\Delta_i = \det \begin{bmatrix} g_1 & g_3 & g_5 & \cdots & g_{2i-1} \\ g_0 & g_2 & g_4 & \cdots & g_{2i-2} \\ 0 & g_1 & g_3 & \cdots & g_{2i-3} \\ 0 & g_0 & g_2 & \cdots & g_{2i-4} \\ \cdot & \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdot & \cdots & g_i \end{bmatrix} .$$

The indices in each row increase by two and the indices in each column decrease by one. The term g_j is taken to be zero if $j < 0$ or $j > n$. Note that $\Delta_1 = g_1$.

Theorem 1. (Routh-Hurwitz stability) A necessary and sufficient condition that the polynomial (6) is strictly Hurwitz is that $\Delta_1, \Delta_2, \ldots, \Delta_n$ be all positive [3].

Since $\Delta_n = g_n \Delta_{n-1}$, the condition that Δ_{n-1} and Δ_n be positive is equivalent to the requirement that Δ_{n-1} and g_n be positive. Furthermore, a necessary condition for (6) to be strictly Hurwitz is that all coefficients g_0 through g_n be positive.

3 The main quadratic system

Comparing coefficients in (5), we derive a quadratic system of $n+1$ equations in the variables $g_0, g_1, \ldots, g_n$:

$$G_k = \sum_{i+j=2k} (-1)^i g_i g_j, \qquad (k = 0, 1, \ldots n) \tag{7}$$

This is the system we are aiming to solve in the factorization problem. The additional constraint is that the polynomial (6) is stable. When $n = 5$,

$$\begin{aligned} G_0 &= g_0^2 \\ G_1 &= -g_1^2 + 2g_0g_2 \\ G_2 &= g_2^2 + 2g_0g_4 - 2g_1g_3 \\ G_3 &= -g_3^2 - 2g_1g_5 + 2g_2g_4 \\ G_4 &= g_4^2 - 2g_3g_5 \\ G_5 &= -g_5^2 \end{aligned} \tag{8}$$

So in this case the stable factorization problem is to find a solution $(g_0, g_1, g_2, g_3, g_4, g_5)$ of the quadratic system (8) in which each $g_i > 0$, and in addition the constraints

$$\Delta_2 = \det\begin{bmatrix} g_1 & g_3 \\ g_0 & g_2 \end{bmatrix} > 0, \quad \Delta_3 = \det\begin{bmatrix} g_1 & g_3 & 0 \\ g_0 & g_2 & g_4 \\ 0 & g_1 & g_3 \end{bmatrix} > 0, \quad \Delta_4 = \det\begin{bmatrix} g_1 & g_3 & g_5 & 0 \\ g_0 & g_2 & g_4 & 0 \\ 0 & g_1 & g_3 & g_5 \\ 0 & g_0 & g_2 & g_4 \end{bmatrix} > 0$$

are satisfied. In the general case $G_0, G_1, \ldots, G_n$ with $G_0 > 0$ are given as input. The output the solution required is real $g_0, g_1, \ldots, g_n$, with $g_0 > 0$ such that $g_0, g_1, \ldots, g_n$ is a solution of the associated quadratic system (7) of $n + 1$ equations and $g(z) = g_0 + g_1z + \cdots + g_nz^n$ is strictly Hurwitz. We assume that the G_k are given so that the system has a solution of the required type.

3.1 Newton's method

We consider the vector valued function $f : \mathbb{R}^{n+1} \to \mathbb{R}^{n+1}$ which has as its set of real zeros the solutions to the quadratic system (7). For $n = 5$, $f : \mathbb{R}^6 \to \mathbb{R}^6$ is $f = (f_0, f_1, \ldots, f_5)^t$ with

$$\begin{aligned} f_0 &= x_0^2 - G_0 \\ f_1 &= -x_1^2 + 2x_0x_2 - G_1 \\ f_2 &= x_2^2 + 2x_0x_4 - 2x_1x_3 - G_2 \\ f_3 &= -x_3^2 - 2x_1x_5 + 2x_2x_4 - G_3 \\ f_4 &= x_4^2 - 2x_3x_5 - G_4 \\ f_5 &= -x_5^2 - G_5 \end{aligned}$$

We compute the Jacobian matrix as

$$J_f = 2\begin{bmatrix} x_0 & 0 & 0 & 0 & 0 & 0 \\ x_2 & -x_1 & x_0 & 0 & 0 & 0 \\ x_4 & -x_3 & x_2 & -x_1 & x_0 & 0 \\ 0 & -x_5 & x_4 & -x_3 & x_2 & -x_1 \\ 0 & 0 & 0 & -x_5 & x_4 & -x_3 \\ 0 & 0 & 0 & 0 & 0 & -x_5 \end{bmatrix}.$$

We calculate by elementary operations

$$\det(J_f) = 2^6(-x_0x_5)\det\begin{bmatrix} -x_1 & x_0 & 0 & 0 \\ -x_3 & x_2 & -x_1 & x_0 \\ -x_5 & x_4 & -x_3 & x_2 \\ 0 & 0 & -x_5 & x_4 \end{bmatrix} = 2^6(-x_0x_5)\det\begin{bmatrix} x_1 & x_3 & x_5 & 0 \\ x_0 & x_2 & x_4 & 0 \\ 0 & x_1 & x_3 & x_5 \\ 0 & x_0 & x_2 & x_4 \end{bmatrix}$$

and $\det(J_f) = 2^6(-x_0)\Delta_5$. For general n we have a similar identity relating the Jacobian of f and Δ_n as

$$\det(J_f) = 2^{n+1}(-1)^n x_0 \Delta_n. \tag{9}$$

Thus the Jacobian J_f does not vanish at $(g_0, g_1, \ldots, g_n)$ at if $(g_0, g_1, \ldots, g_n)$ corresponds to a stable $g(x)$. In other words, starting from an initial point that is close enough to the stable solution, the Jacobian of f does not vanish. Starting with an initial vector $X_0 = (x_0, x_1, \ldots, x_n)^t$ we compute the iterates by Newton's method as

$$X_{n+1} = X_n - J_f^{-1}(X_n) f(X_n)$$

until successive iterates are within a given tolerance. The invertibility of J_f at the point X_n is guaranteed for X_n close to a stable solution $(g_0, g_1, \ldots, g_n)$. However a real solution $g(z)$ of the quadratic system found by Newton's method is not necessarily strictly Hurwitz. The polynomial we want is obtained from $g(z)$ by flipping the sign of some of its roots and making each one have negative real part, even though we do not have access to the roots themselves.

Example 1. Suppose $G(x^2) = G_0 + G_1x^2 + G_2x^4 + G_3x^6 + G_4x^8$ with $G_0 = 9.244$, $G_1 = 72.286$, $G_2 = 217.183$, $G_3 = 296.638$, $G_4 = 155.673$. $G(x^2) = g(x)g(-x)$ where

$$g(x) = g_0 + g_1x + g_2x^2 + g_3x^3 + g_4x^4$$

with $g_0 = 3.040$, $g_1 = 2.289$, $g_2 = 12.749$, $g_3 = 4.637$, $g_4 = 12.476$ is strictly Hurwitz. Starting with the initial random vector of coefficients $(1.933, 2.008, 0.181, 0.870, 2.582)$, and tolerance 0.01, Newton's method converges to the polynomial

$$3.040 + 0.004x + 11,887x^2 - 0.003x^3 + 12.477x^4$$

of the quadratic system in 14 iterations. This polynomial is not stable. Its roots are $-0.0928 \pm 0.6956j$ and $0.0929 \pm 0.6972j$.

3.2 An auxiliary problem

The necessity of being able to "flip" the sign of certain roots of a given real polynomial as indicated above results in the following auxiliary problem:

> *Given a real polynomial* $g(x) = g_0 + g_1x + \cdots + g_nx^n$ *of degree* n *with* $g_0 > 0$, *construct the real polynomial* $h(x) = h_0 + h_1x + \cdots + h_nx^n$ *with* $h_0 = g_0$, *such that the roots of* h *are* $\pm$ *roots of* g *and* h *is strictly Hurwitz.*

If we could generate the polynomials h whose roots differ from the roots of g only in their sign, then we could test each polynomial generated by the Routh-Hurwitz criteria to see if it is stable. But there cannot be an analytic way involving radicals to do this: Consider a generic fifth degree polynomial $g(x) = g_0 + g_1x + \cdots + g_5x^5$ with $g_0 > 0$. Let r be a real root of g, and let $h(x) = h_0 + h_1x + \cdots + h_5x^5$ be the polynomial with $h_0 = g_0$, which has identical roots as $g(x)$, except for its fifth root it has $-r$ instead of r. Since g_4 and h_4 are the negative of the sums of the roots of $g(x)$ and $h(x)$ respectively, we have $r = \frac{1}{2}(h_4 - g_4)$. We can also calculate $g(x)/(x - r)$ by synthetic division and compute the roots of this quartic by radicals. Thus if there were a way of computing the coefficients of $h(x)$ from those of $g(x)$ by means of radicals, then this would allow us to express the roots of a general fifth degree polynomial by radicals.

4 Algorithmic approaches to finding $g(x)$

There are two essentially distinct approaches to find the strictly Hurwitz polynomial $g(x)$ given the input data $G_0, G_1, \ldots, G_n$ with $G_0 > 0$. Both have a random component.

A1 : Generate an initial vector $X_0 = (x_0, x_1, \ldots, x_n)^t$ and run Newton's method starting with X_0. Let the converged polynomial be $h(x)$. If $h(x)$ passes the Routh-Hurwitz criteria, then it is the strictly Hurwitz polynomial desired and we are done. If not, generate another X_0 and continue.

A2 : Generate an initial vector $X_0 = (x_0, x_1, \ldots, x_n)^t$ and run Newton's method starting with X_0. Let the converged polynomial be $h(x)$. If $h(x)$ passes the Routh-Hurwitz criteria, then it is the strictly Hurwitz polynomial desired and we are done. If not, use the coefficients of $h(x)$ to generate another X_0 and continue.

A1 is simple to implement. On the other hand the number of executions of the Newton method is fewer for A2, which essentially goes from a computed $h(x)$ to another polynomial whose roots are negatives of some of the roots of $h(x)$. We shall indicate a number of methods for A2.

Example 2. In [8], the data given for a monopole antenna is modeled using the linear interpolation technique proposed for positive real functions. We used this problem for the experimental evaluation of A1 and A2. For this model $G(x^2) = G_0+G_1x^2+G_2x^4+G_3x^6+G_4x^8$ with $G_0 = 9.244$, $G_1 = 72.286$, $G_2 = 217.183$, $G_3 = 296.638$, $G_4 = 155.673$. Employing A1, the strictly Hurwitz polynomial

$$g(x) = g_0 + g_1x + g_2x^2 + g_3x^3 + g_4x^4$$

with $g_0 = 3.040$, $g_1 = 2.289$, $g_2 = 12.749$, $g_3 = 4.637$, $g_4 = 12.476$ was found. The initial vectors $X_0 = (x_0, x_1, x_2, x_3, x_4)^t$ were generated by picking x_i independently and uniformly in the range $0 < x_i < \sqrt{G_0}$. The average number of different starting points required for the Newton method for convergence to $g(x)$ with a tolerance of 0.001 is about 8 with a standard deviation of 5.

Next we consider two properties of the family of polynomials which are solutions to (4.3) and differ only in the signs of their roots.

4.1 Selection of starting points

For A2, we use the following idea: Suppose we have two real solutions $g(x)$ and $h(x)$ to the quadratic system of equations (7). Then $G(x^2) = g(x)g(-x) = h(x)h(-x)$ where $g(x)$ and $h(x)$ have the same roots up to signs. Define $F(x) = h(x)/g(x)$. Since $h(x)/g(x) = g(-x)/h(-x)$, $F(x)$ satisfies the functional equation

$$F(x)F(-x) = 1\ . \tag{10}$$

Put $F(x) = c_0 + c_1x + c_2x^2 + \cdots$ with $c_0 = 1$. >From (10), we have

$$\begin{aligned}
1 &= c_0^2 \\
0 &= -c_1^2 + 2c_2 \\
0 &= c_2^2 + 2c_0c_4 - 2c_1c_3 \\
0 &= -c_3^2 + 2c_0c_6 - 2c_1c_5 + 2c_2c_4 \\
0 &= c_4^2 + 2c_0c_8 - 2c_1c_7 + 2c_2c_6 - 2c_3c_5 \\
\vdots &= \qquad\qquad \vdots
\end{aligned}$$

The general form of the k-th equation for $k \geq 1$ is

$$0 = \sum_{i+j=2k} (-1)^i c_i c_j.$$

In this infinite system, each of $c_2, c_4, c_6, \ldots$ can be expressed in terms of the coefficients $c_1, c_3, c_5, \ldots$. In fact, we can represent c_{2k} as a polynomial in $c_1, c_3, \ldots c_{2k-1}$. >From the second equation, $c_2 = \frac{1}{2}c_1^2$. Using this with the third equation we get

$$c_4 = -\tfrac{1}{2}c_2^2 + c_1c_3 = -\tfrac{1}{8}c_1^4 + c_1c_3 \quad \text{and} \quad c_6 = \tfrac{1}{2}c_3^2 + c_1c_5 - \tfrac{1}{2}c_1^3c_3 + \tfrac{1}{16}c_1^6 \ .$$

In the general case we can write

$$c_{2k} = \tfrac{1}{2}(-1)^{k+1}c_k^2 + \sum_{i=1}^{k-1}(-1)^{i+1}c_i c_{2k-i} \ . \tag{11}$$

In (11), we repeatedly substitute the expressions obtained for the earlier coefficients with even indices, we arrive at the expression of c_{2k} in terms of $c_1, c_3, c_5, \ldots$. Therefore

$$F(x) = 1 + c_1x + \tfrac{1}{2}c_1^2x^2 + c_3x^3 + (c_1c_3 - \tfrac{1}{8}c_1^4)x^4 + c_5x^5 + (\tfrac{1}{2}c_3^2 + c_1c_5 - \tfrac{1}{2}c_1^3c_3 + \tfrac{1}{16}c_1^6)x^6 + c_7x^7 + \cdots$$

Thus $h(x) = g(x)\left(1 + c_1x + \frac{1}{2}c_1^2x^2 + c_3x^3 + (c_1c_3 - \frac{1}{8}c_1^4)x^4 + c_5x^5 + \cdots\right)$ for some real numbers $c_1, c_3, c_5, \ldots$. We can use the form of the coefficients of $F(x)$ to pick a new starting point if the solution $g(x)$ we obtain from the Newton's method fails to be stable.

For algorithm A2, we generate a new initial point $h(x)$ for Newton's method from the current computed solution $g(x) = g_0 + g_1x + \cdots + g_nx^n$ with $g_0 > 0$ by setting $h_k = \sum_{i=0}^{k} g_i c_{k-i}$ for $k = 0, 1, \ldots, n$ with $c_0 = 1$ and $0 = \sum_{i=0}^{k} g_i c_{k-i}$ for $k > n$. We then express the even indexed c_i in terms of the odd index ones. After this stage, the h_k's involve only $c_1, c_3, \ldots, c_{n-1}$ (or up to c_n if n is odd.) We pick random values for these c_i's satisfying these constraints.

Example 3. For the data in Example 2, we considered the experimental evaluation of A2. The initial vector $X_0 = (x_0, x_1, x_2, x_3, x_4)^t$ was generated by picking x_i independently and uniformly in the range $0 < x_i < \sqrt{G_0}$. Following that the algorithm jumps to the next initial vector using the ideas presented above. The average number of iterations to converge to the strictly Hurwitz polynomial within a tolerance of 0.001 was 2, with a standard deviation of 1.

4.2 A linear algebraic property

Given a real polynomial $g(x) = g_0 + g_1x + \cdots + g_nx^n$ of degree n, we briefly consider the problem of constructing a new polynomial $h(x) = h_0 + h_1x + \cdots + h_nx^n$ whose roots depend on the roots of g, without actually finding the roots themselves. Without loss of generality, $g_n = 1$.

Suppose the roots of g are $\beta_1, \ldots, \beta_n$ and the required roots of h are $p(\beta_1), \ldots, p(\beta_n)$ for some polynomial p. Consider the companion matrix of g defined by

$$C = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ \vdots & & \ddots & \vdots \\ 0 & & & 1 \\ -g_0 & -g_1 & \cdots & -g_{n-1} \end{bmatrix}$$

The characteristic polynomial of C is $\det(xI - C) = g(x)$. Then h can be expressed in terms of only the coefficients of g as $\det(xI - p(C)) = h(x)$ without calculating the zeros $\beta_1, \ldots, \beta_n$. For example for $g(x) = g_0 + g_1 x + x^2$ with zeros β_1, β_2,

$$C = \begin{bmatrix} 0 & 1 \\ -g_0 & -g_1 \end{bmatrix}$$

and the characteristic polynomial of $C^2 - 3C$ has zeros $\beta_1^2 - 3\beta_1, \beta_2^2 - 3\beta_2$. We compute

$$C^2 - 3C = \begin{bmatrix} 0 & -2 \\ 3g_0 + g_0^2 & 3g_1 + g_1^2 \end{bmatrix}$$

and therefore

$$h(x) = \det(\begin{bmatrix} x & 2 \\ -3g_0 - g_0^2 & x - 3g_1 - g_1^2 \end{bmatrix}) = 2g_0(3 + g_0) - g_1(3 + g_1)x + x^2 \ .$$

The reason for this is that C is similar to an upper triangular matrix with $\beta_1, \ldots, \beta_n$ on the diagonal [4],

$$BCB^{-1} = \begin{bmatrix} \beta_1 & & & \\ 0 & \beta_2 & & * \\ \vdots & & \ddots & \\ 0 & \ldots & 0 & \beta_n \end{bmatrix}, \text{ so that } Bp(C)B^{-1} = \begin{bmatrix} p(\beta_1) & & & \\ 0 & p(\beta_2) & & * \\ \vdots & & \ddots & \\ 0 & \ldots & 0 & p(\beta_n) \end{bmatrix}.$$

Functions other than polynomials can be used for p (e.g. mixtures of exponential and certain rational functions). However for this approach to work, each β_i must be transformed by the same function p. We only want to change the sign of one of the β_i at a time.

Let $I(i)$ be the matrix that is obtained from the identity matrix by changing the ith 1 to a -1. We would like to construct the matrix $s_i(C)$ such that

$$Bs_i(C)B^{-1} = I(i)BCB^{-1}. \tag{12}$$

Then the characteristic polynomial of $s_i(C)$ has zeros $\beta_1, \ldots, \beta_{i-1}, -\beta_i, \beta_{i+1}, \ldots, \beta_n$. >From (12), $s_i(C) = BI(i)B^{-1}C$. If $b_1, \ldots, b_n$ are the column vectors of B and $c_1, \ldots, c_n$ are the row vectors of B^{-1}, then $s_i(C)$ and C are related by

$$s_i(C) = (I - 2b_i c_i)C \ . \tag{13}$$

We do not need the matrix B exactly (this may involve finding eigenvalues, which are not permitted in this approach). The characteristic polynomial of the perturbed matrix in (13) will be used as a starting point for the next iteration of Newton's method, so b_i and c_i and the outer product $b_i c_i$ can be approximate.

5 Conclusions and further work

We have proposed a stable factorization procedure generate strictly Hurwitz polynomial from a given strictly positive even polynomial. The factorization process is carried out directly to find the solution of a set of quadratic equations in many variables employing Newton's method.

It is hoped that the method presented in this paper generalizes to two-variable polynomials. This would make possible the generation of scattering Hurwitz polynomials, which are the two-dimensional analogues of strict Hurwitz polynomials.

Bibliography

[1] Aksen A., "Design of Lossless Two-ports with Mixed, Lumped and Distributed Elements for Broadband Matching," PhD. Dissert., Lehrstuhl Für Nachrichtentechnik, Ruhr Universitaet Bochum, 1994.

[2] Fettweis A., "On the Scattering Matrix and the Transfer Scattering Matrix of Multi Dimensional Lossless Two-ports," Int. J. of Communication, vol. 36, pp. 374-381.

[3] Henrici P., *Applied and Computational Complex Analysis*, Vol. II, Wiley, New York, 1977.

[4] Lang S., *Linear Algebra*, Addison-Wesley, Reading, MA, 1966, p. 183.

[5] Yarman B. S., "Broadband Networks," Wiley Encyclopedia of Electrical and Electronics Engineering, Vol. II, pp. 589-604, 1999.

[6] Yarman B. S. and Aksen A., "A Reflectance-based Computer Aided Modeling Tool for High Speed/High Frequency Communication Systems," Proc. IEEE-ISCAS 2001; 4, pp. 270-273.

[7] Yarman B. S., Aksen A. and Kilinc A., "Immitance Data Modeling via Linear Interpolation Techniques," Proc. IEEE-ISCAS 2002; 3, pp. 527-530.

[8] Yarman B. S., Kilinc A. and Aksen A., "Immitance Data Modeling via Linear Interpolation Techniques: a Classical Circuit Theory Approach," Int. J. of Circuit Theory and Applications, 2004; 32, pp. 537-563.

[9] Yarman B. S. and Carlin H. J., "A Simplified Real Frequency Technique Applied to Broadband Multi-stage Amplifiers," IEEE Trans. MTT; 30, pp. 2216-2222, 1982.

Int. J. of Computers, Communications & Control, ISSN 1841-9836, E-ISSN 1841-9844
Vol. V (2010), No. 5, pp. 710-718

Bounded Rationality Through the Filter of the Lisbon Objectives

R. Fabian, M.J. Manolescu, L. Galea, G. Bologa

Ralf Fabian
"Lucian Blaga" Univeristy of Sibiu
Romania, 550012 Sibiu, 5-7 Dr. Ion Ratiu st.
E-mail: ralf.fabian@ulbsibiu.ro

Misu-Jan Manolescu, Loredana Galea, Gabriela Bologa
Agora University, Oradea.
Cercetare Dezvoltare Agora
E-mail: rectorat@univagora.ro, loredana.galea@univagora.ro, gabi_fiat@yahoo.com

Abstract: Information and Communication Technologies (ICT) have created best conditions for grows of knowledge societies. An emerging global information society serves to building global knowledge societies as source for further development. Conventional paradigms of sciences starts to be more blemish and prone to redefinition of there foundations, understood as scientific knowledge. The perspective of knowledge and ideals of rationality are both heavily influenced by a new contemporary scientific thinking, through tools, inherent of autonomy and uncertainty. A new understanding of the world in terms of open dynamic heterogeneous uncertain systems is needed. Among the conclusions: classical rational reasoning is mainly aiming at effectiveness, not at uncertain knowledge processing, because of its temporality (mainly its ineffectiveness in dealing with future events); a bounded-rationality approach enables both, better economic models and better modelling, being based on trends in economic modelling as well as on agent-oriented software engineering.
Keywords: knowledge society paradigms, bounded rationality, "just in time" paradigm, uncertain knowledge processing.

1 Introduction

The approach proposed in this paper, trials to a better reaction to uncertain and rapidly changing environments as result of ICT implications in the "grows of economy". As far target the paper has three specific aims: (a) Showing that both bounded rationality and agents are, at the same time, unavoidable restrictions, and valuable means when developing applications. (b) Underlining the inadequacy of conventional formal methods with algorithm-based computational techniques applied in AI, in general but firstly for treating uncertainty in applications. (c) Outlining, on this groundwork, an agent-oriented approach to combine synergistically bounded rationality and agents in modelling.

In a society that strives to be progressively technically, technology is becoming a tool for social interaction bridging the strands between online and offline activities, respectively, digital and social behaviour. Studies demonstrate clearly that the Information and Communication Technologies (ICT) experienced in people's everyday life sets a milestone for an active participation in the Knowledge Society (KS) development [9].

The "*Lisbon Strategy*", developed by the European Council for the time period 2000-2010, aims to make the EU "the most dynamic and competitive knowledge-based economy in the world capable of sustainable economic growth with more and better jobs and greater social cohesion,

and respect for the environment by 2010" [20] [9] by considering an economic, a social and an environmental pillar.

Information Society policies address core objectives of the Lisbon Strategy: drive productivity growth, create an open and competitive digital economy, and stimulate innovation to tackle changes of globalization and demographic change.

ICT affects economic performance by driving innovation through investment in ICT; improving business processes and reduce companies administration costs; increasing efficiency in public administration; enhance productivity grow and create new markets.

For the period beyond 2010 the EU develops a new strategy to make a recovery from financial crises and speed up towards a greener, more sustainable and innovative economy. In a formal "Consultation on The Future "EU 2020" Strategy" [9] [20] the EU Commission considers that the key drivers of EU 2020 should be focused on the following thematic: Creating value by basing growth on knowledge; Empowering people in inclusive societies; and creating a competitive, connected and greener economy.

Conventional paradigms of sciences starts to be more blemish and prone to redefinition of there foundations understood as scientific knowledge. Accepted aspects are questioned and relocate the place of man in the world. The perspective of knowledge and ideals of rationality are both heavily influenced by a new contemporary scientific thinking, through tools, inherent of autonomy and uncertainty.

Innovative theoretical solutions, such as the approach of complexity, change old fashion models of rationality. [15] [16] The new rationality prioritizes the view of complexity as an essential characteristic of the surrounding reality (social and non-social). The classical ideal of rationality is centred on the supremacy of reason, which is to be overcome by the approach of complexity. A new understanding of the world in terms of dynamic systems is needed, where interaction between the elements of the system and the environment rules.

For the society, education systems are institutional spaces for generation and transmission of complex knowledge. A premise for a global knowledge based society is an adequate amount of highly qualified scientists. The Universities have been aware in the last decades to expend access to higher education. Beside the need fore more graduates (this objective has reached a significant progress since 2000, increasing the number of Mathematics, Science and Technology graduates, by more then 25%. [9]), an even more important issue is to assure that after graduating, lifelong learning takes place in work and leisure time.

The insertions of new ICT and artificial intelligence in the human society, together with the promotion through the Lisbon objectives, have reached, in the last century, every area of human activities, from education and economy to political systems. If we endeavour to undertake any kind of professional research and development (R&D), we have to analyze the applicability and changes brought us through them.

Premises for the research described here, as well as for other research surveyed in this paper: a) In the last decade most non-trivial IT applications are meant for, and hosted by, open, dynamic, and uncertain environments. b) Ever more services have to be provided in line with the "just in time" (JIT) paradigm; c) developing applications for JIT services implies both bounded rationality (fact of life) and artificial intelligence (powerful IT instrument) as cardinal design-space dimensions. [3] [22] [8] The basic idea is that the dynamics of every facet of our society (first of all, economy) are so intense that no field of study could afford a slower development pace.

Present-day IT environments, except for some irrelevant applications, move fast from limited, homogeneous, changing slowly, deterministic (even if partly approximated or even unknown) towards (OHDUE). That means: "open and heterogeneous (the resources involved are unalike and their availability is not warranted), dynamic (the pace of exogenous and endogenous changes

is high) and uncertain (both information and its processing rules are revisable, fuzzy, uncertain and intrinsically non-deterministic" [22] [8] - as most environmental and almost all human stimuli generators).

This paper is organized as follows: After summarizing the history of the undertaking and revising basic concepts, (Section 2), its rationale and approach are outlined in (Section 3) together with considerations highlighting the disadvantages of conventional IT, because of inadequate interaction in dynamic environments. Decision making in from the perspective of bounded rationality is suggested. Conclusions and future work (Section 6) close the paper.

2 History and basic concepts

Conventional IT undertakings relaying on formal methods of computation used for modelling economic systems [10], showed the limitations of the conceptual bases. The rather rigid character and incapability to adapt over time to a rapid changing environmental context, has been addressed afterwards by introducing "total fuzzy grammars" [12].

Trying to manage non deterministic intelligent systems, from economic modelling [12] [10] to natural language processing for ChatBots [11], with intrinsic deterministic tools, proved the be inadequate of this kind endower. Tools unable to deal with a certain degree of uncertainty are not worth to be considered in any kind of modelling real life scenarios.

The approach from [22] addresses bounded-rationality and enables better economic modelling, being based on trends in economic modelling as well as on agent-oriented software engineering.

Uncertainty was analyzed with regard to approximation and undecidability [22] [3] and emphasis that they are two incommensurable kinds of uncertainty: approximation is deterministic and atemporal and undecidability is nondeterministic and has basic temporal dimension. Thus, approximation aims at optimization and efficiency and doesn't fit as instrument for uncertainty. On the other hand, undecidability has to deal with future events in any decision making process and can not be treated exclusively through approximation or other conventional prediction methods (probabilistic, stochastic, etc.)

Research directions on *non-algorithmic* approaches begun in 2007 and have the role to support the approach in this paper. Since the conclusions from [21] [4] are general, they can be easily adapted to real world applications by further investigations and development.

Information and knowledge. The essence of communication in KS is the message to be transmitted. To clarify the meaning used in this paper, of three overlapping and often misleadingly used concepts, regarding message transmission, we consider answering the following questions: What is (will be) passing through modern electronic pipelines? Data, Information or Knowledge?

Data alone caries no meaning and as such represents the lowest level of abstraction. It is used as carrier for collections of numbers, characters, images, bits and bytes etc. Data represents "qualitative or quantitative attributes of a variable or set of variables" (http://en.wikipedia.org/wiki/Data). For data to become information they have to be interpreted in such way that it gets associated with a meaning. Information bears a diversity of meanings from everyday usage to technical settings and is depending on the context of use. Hence, we can see it as second level of abstract while the highest one will be knowledge.

According to Oxford English Dictionary, knowledge is defined as: a) expertise, and skills acquired by a person through experience or education; the theoretical or practical understanding of a subject, (2) what is known in a particular field or in total; facts and information; c) awareness or familiarity gained by experience of a fact or situation. Though, the defining of knowledge is subject for several ongoing philosophical debates.

Putting al three terms together we can say that information is the result of processing data such that it enlarges the knowledge of a person. Therefore, trough the pipelines are passing a mix of all three, decisive being the nature of sender and receiver.

Information and knowledge society. Information Society refers to a society where information and ICT are explicitly ruling. A common vision on the Information Society, "where everyone can create, access, utilize and share information and knowledge, enabling individuals, communities and peoples to achieve their full potential in promoting their sustainable development and improving their quality of life...", was settled by the World Summit on the Information Society (WSIS, Declaration of Principals, Geneva phase 2003).

The idea of Information Society is based on technological breakthroughs while the Knowledge society encompasses much broader social, ethical and political dimensions.

Historically, the Information Society concept was overtaken rapidly by the idea of the Knowledge Society. The main issues of settling for the idea of Knowledge Society were that the Information Society was to restricted for policy and knowledge is more usable than information.

"Knowledge is today recognized as the object of huge economic, political and cultural stakes, to the point of justifiably qualifying the societies currently emerging." [25] A KS is "a society that creates shares and uses knowledge for the prosperity and well-being of its people".

New dynamics have emerge since the "Third Industrial Revolution" (that of new technologies) and constantly evolved from education, science and technique to cultural aspects. Knowledge is central to these changes. The need for an accepted diversity means that knowledge societies will have to be societies of shared knowledge. ICT, especially by the emergence of the internet to a public network, carries new opportunities for a universal access to knowledge. Consequently, it is unacceptable that ICT should lead to a fatalistic technological determinism.

The changes in sciences and implicitly in everyday life produced a revolution in the conception of man by the means of how to regard and produce knowledge and sciences. The long term impact of ICT to knowledge production is described by Manuel Castells as "application of such knowledge to knowledge generation and information processing/communication devices, in a cumulative feedback loop between innovation and the uses of innovation." [18]

3 Rationale and Approach

The acknowledgement of the systemic nature of the world implies a vision of the complex reality together with there diverse interrelationships. It is necessary to adopt methods which allow working from a holistic perspective in order to address adequately complex problems.

Now, more then ever, in the Knowledge Society, any activity is affected in one way ore another by ICT tools. Scientists are able to create state of the art technologies designed for collaboration and transdisciplinarity.

The paper proposes an approach for achieving and promoting the Lisbon objectives through the ICT perspective by narrowing the gape between the technologic perspective and the anthropocentric one. In this endower we consider three paradigms as main pillars for sustainable development in new societies:

- Open Heterogeneous Dynamic Uncertain Environment (OHDUE) - as natural real world environment;
- Bounded Rationality (BR) - as kind of rationality in decision making;
- Just in time paradigm (JIT) - as limit of perception.

We concentrate now on ICT paradigms impulsed by the Lisbon objectives to absorb the speed of change in/for the KS, to sustain the pillars considered above. Since the required change of mentality is profound and urgent, the approach must avoid "solutions in search of a problem". Hence, total pragmatism to begin with: a) To validate the approach - at least in ovo - a

relevant, cardinal, and for the most part ill-defined real-world problem is focused on: decision making in economic modelling. b) The solution must reduce complexity: b1) cognitive; that means obvious functionality, no sophisticated concepts or instruments (agents, temporal logics, explicit uncertain information processing, computability theory and computational complexity theory, Bayesian methods, certainty factors, etc.); b2) structural; that means simple mechanisms, immediate applicability in current designs; conventional software engineering. [8]

IT paradigms shift from "client-server" to "computing as interaction". However, since the prevalence of older paradigms are still active and the issue is central to this paper, a historical perspective is helpful: For over 40 years, determinism and bivalent logic were the pillars of Computer Science; likewise, algorithms were the backbone of computer programs, complying with their etymon: pro-gramma = what is written in advance. When early real-time applications (firstly, operating systems) required less autistic programs, algorithms tried to adapt accepting "unsolicited input", to fit the incipient non-determinism due to user free will. Bivalence not only survived, but also, strongly backed by hardware, grew in importance. In the early 70's, the role of bivalent logic transcended the borders of narrow data processing, penetrating "Computer-Aided x", where x stays for almost any intellectual activity (including decision-making). [22] [5]

Uncertainty as epistemic concept. The relationship between uncertainty and non-determinism is intricate but irrelevant in decision-making context, because uncertainty can appear in deterministic problems too (playing chess is a manifest example: certainty about the best move is given up to speed - better said, to inexorable time restrictions). Thus, research can avoid non-determinism and focus only on uncertainty, its species and degrees. Due to its vital role in any kind of decision-making, a subsection in [7] is dedicated to this topic.

Uncertain knowledge processing. Since "in the knowledge-based process planning system, the attribute values are usually miscellaneous, heterogeneous and uncertain" and in "manufacturing process planning, experts often make decisions based on different decision thresholds under uncertainty [...], a novel approach to integrating fuzzy clustering is proposed" in [28] able to "discover association rules more effectively and practically in process planning with such thresholds".

Due to the characteristics of this century, when speed and efficiency are vital in all fields of interest, decision making plays an important role in almost every domain of human activity. In such tasks, alternatives are considered and there outcomes are measured until a certain threshold is reached, at which time the solution is found. Human do not use explicit probabilistic/stochastic thinking when they make time constrained decisions [22] [3] [8].

Undecidability. Definitions on the Web: a) "property of knowledge representation language that the truth of some of the true statements within that language cannot be established by any algorithm" (www.dbmi.columbia. edu/homepages/wandong/KR/krglossary.html); b) "Undecidable has more than one meaning in mathematical logic: b1) A decision problem is called (recursively) undecidable if no algorithm can decide it, such as for Turing's halting problem. b2) "Undecidable" is sometimes used as a synonym of "independent", where a formula in mathematical logic is independent of a logical theory if neither that formula nor its negation can be proved within the theory." (en.wikipedia.org/wiki/Undecidability). Obviously such definitions are too "full of mathematics" to be thoroughly comprehended in common decision-making. On the contrary, the definition "Within a system, a statement is said to be undecidable when it cannot be shown to be either true or false." (ddi.cs.uni-potsdam.de/Lehre/TuringLectures/Math Notions.htm is relevant for this paper. Indeed, even when the fact that accurate numeric data are hard to get is accepted, the emphasis is on approximated, predicted, evaluated by rule of thumb, or even on intrinsically fuzzy data, rather than on missing ones (lacking sensor or market information, delayed previous decisions, server crash etc.). However, decisions are difficult because a relevant event not happened yet, not because a result is imprecise.

Approximation. A web definition of approximation states "inexact representation of some-

thing that is still close enough to be useful. Although approximation is most often applied to numbers, it is also frequently applied to such things as mathematical functions, shapes, and physical laws." (en.wikipedia.org/wiki/Approximation). These definition is relevant as regards: a) both the commonsensical and the scientific (mainly, mathematical) meanings of approximation; b) its role as degree of uncertainty (as "measure" of: imprecision, difference between a reported value and a real one, possible error or range of error, etc.); c) its major function as optimization tool. Indeed, approximation, seen as a "don't care"-like uncertainty, can speed up remarkably data processing in key IT subdomains (e.g., image processing, form and motion recognition, surface design, web classification, networks). [3] [22] But on the other hand: approximation is inherently unsuitable for "don't know" - like uncertainty, not even in deterministic contexts. [8]

Bounded Rationality. The very principle of bounded rationality had been developed a half-century ago by the Chicago School and endorsed brilliantly by Simon [24]. Rubinstein describes "models in which procedural aspects of decision making are explicitly included" in [23]. Kahneman received a Nobel Prize for seeing bounded rationality as a means to improve economic modelling [16] linking it to psychological processes or to communication faults that could explain (or not) ill-applied statistical thinking in decision-making. Gigerenzer proposed alternatives for decision making, based on simple heuristics [15] that "lead to better decisions than the theoretically optimal procedure" (en.wikipedia.org/wiki/Bounded rationality); for instance, priority heuristics [18]). A more recent advance by Edward Tsang presents the Rubinstein approach in a computational point of view, referred as CIDER (Computational Intelligence Determines Effective Rationality) theory. This way of interpreting BR enables to reason about economic systems when the full rationality assumption is relaxed. [26]

The problem is that none of the aspects mentioned above are taken into account sufficiently, neither in economic modelling, nor in decision-making. Moreover, ignoring the fact that bounded rationality is "a form of behaviour associated with uncertainty where individuals do not examine every possible option open to them" (www.pestmanagement.co.uk/lib/glossary/glossary_b.shtml), the mathematical tools still recommended for modelling economic processes (taking place in OHDUE) are ill-applied when they try to deal with undecidability and complexity.

Bounded rationality is defined as concept that decision makers (irrespective of their level of intelligence) have to work under three unavoidable constraints: (a) only limited, often unreliable, information is available regarding possible alternatives and their consequences; (b) human mind has only limited capacity to evaluate and process the information that is available, and (c) only a limited amount of time is available to make a decision.

Therefore even individuals who intend to make rational choices are bound to make satisfying (rather than maximizing or optimizing) choices in complex situations. Consequently, considering bounded rationality in modelling, would have two convergent, major beneficial effects: reducing attempts in the counterproductive direction (i.e., treating undecidability) and promoting efforts in the valuable course (i.e., offering "just in time" answers). Thus, the paradox of approximation complexity in modelling would fade away.

Inaccurate complexity management. Due the growing information complexity encountered in every research fields, machine learning techniques gain increasing popularity. Software tools (like Weka and RapidMiner) offer state of the art feature selection techniques and a large number of similarity functions allowing clustering and instance based learning techniques, respectively to deal with height dimensional domains. When complexity is overwhelming (requiring unaffordable resources), instead of simplifying the model, subsymbolic paradigms - mainly artificial neural networks (ANN) or evolutionary algorithms (EA) - are used, ignoring their general weaknesses (no explanations, weak convergence, poor reliability because of inadequate exploitation/exploration ratio) or specific ones: need for training example sets (for ANN) or endogamic limitation (for EA). For instance, what responsible human will make critical (e.g., macroeconomic) decisions

without any justification?

Modern Artificial Intelligence. In an historical early phase of intelligent system development, when expert systems based on the Newell-Simon hypothesis ("A physical symbol system has the necessary and sufficient means for general intelligent action") where "on vogue", they start to disappoint in all nuances of the world. A prompt reaction upon the limits of the symbolic paradigm came timely by replacing "GOFA" (Good Old-Fashioned AI) with "BIC" (Biologically Inspired Computing). Bringing all paradigms together, "modern artificial interlace" means now agents [17] [29] [1]. By a formal standard [14] the agent is now acknowledged as process.

The user impact of agent technology through ICTs [6] [5] breaks up from reversing the three questions: "What for?" - the aim; "What?" - the architecture; "How?" - the structure. Redressing the balance from an end user viewpoint to a user centered one, the three questions become a different interpretation: "What for?" - to get easy and fast help; "What?" - new applications domains become affordable; "How?" - by imposing user needs and not ICT ones. "Technological determinism" for "technology pushed" users is not distinctive for agents, it is just a matter of impact measurement.

4 Conclusions and future work

In the last century we witnessed an explosive growth in the number and diversity of networked devices and portals as main communication channels bringing a height degree of mobility, heterogeneity, and interactions among devices connected to global networks. Knowledge is the engine for sustainable growth. In a fast-changing world, what makes the difference is education and research, innovation and creativity. Building on its strengths in technology and knowledge, Europe should tap fully the potential of the digital economy.

Economic models cannot anymore avoid the consequences of bounded rationality. It has to accept undecidability as concept and to resort to approximation as means. To be relevant, the endower should always address real world problems and not artificially models born and living in abstracted laboratory simulated conditions.

Software agents will be seamlessly integrated in our everyday life and therefore they have to constantly sense and react to the environment. Great opportunities for agent oriented software architecture are opened by the dynamic and heterogeneous environmental character.

Negative impacts of ICTs are partially rooted in the uncertainty coming from the inability to assimilate the complexity, diversity, magnitude and space of the ICT world. Positive impacts increase the possibility of communication by significantly widening the filed of existing applications and facilitates the emergence of new ones. In this context, the negative or positive impact is wrongly attributed to the technology itself, and rather to there applications.

The gape between the user expectations and the technological offer is deepened due insufficient innovation and lake in use of new agent oriented potential.

The future research will focus on agent oriented software engineering as powerful just in time solution in an environment striving toward a knowledge society, where traditional rationality has to be replaced (ore at least heavily supported) by bounded rational approaches, in order to be able dealing with the increasing complexity of real life scenarios.

Bibliography

[1] gentLink III. Agent based computing. AgentLink Roadmap: Overview and Consultation Report. University of Southampton, sept. 2005. http://www.agentlink.org/roadmap/al3rm.pdf

[2] Alain Ambrosi, Valérie Peugeot and Daniel Pimienta, World Matters: Multicultural perspectives on information societies, Sally Burch - The Information Society / the Knowledge Society, C & F Éditions, 2005, ISBN 2-915825-03-3, http://vecam.org/article517.html

[3] Bă rbat Boldur E., Fabian Ralf, Brumar Cristina, Muntean Roxana, Bounded Rationality and approximation in modern artificial intelligence, Proceedings of The International Workshop - New Approaches, Algorithm and Advanced Computational Techniques in Approximation Theory and its Applications, Sibiu, Romania, September, 2007, pg. 64 67, Lucian Blaga University Press, ISBN 978 973 739 949 2

[4] Bărbat, B.E. From e-Learning to e-Nursing, Applying Non-Algorithmic Paths. Medinf 2007 Workshop: E-Learning aspects in medicine and nursing. 29th International Conference of the Romanian Medical Informatics Society, Sibiu, 2007.

[5] Bărbat, B.E., A. Moiceanu. I, Agent. The good, the bad and the unexpected: The user and the future of information and communication technologies (B. Sapio et al, Eds.), Conf. Proc., Brussels, COST Action 298 Participation in the Broadband Society, CD-ROM ISBN: 5-901907-17-5, 2007

[6] Bărbat, B.E., S.C. Negulescu, A.E. Lascu, E.M. Popa. Computer-Aided Semiosis. Threads, Trends, Threats. Proc. of the 11th WSEAS International Conference on COMPUTERS (ICCOMP '07) (N.E. Mastorakis et al, Eds.), 269-274, Agios Nikolaos, Crete, 2007.

[7] Bărbat, B.E., S.C. Negulescu, S. Pleşca. Emergence as Leverage and Non-Algorithmic Approaches in Agent-Oriented Software. Studies in Informatics and Control Journal, 16, 4, 2007.

[8] Dziţac Ioan, Bărbat Boldur E., Artificial Intelligence + Distributed Systems = Agents, Int. J. of Computers, Communications & Control, IV, 1, 17-26, (http:// www.britannica.com/bps/additionalcontent/18/36182542/Artificial-Intelligence–Distributed-Systems–Agents), 2009.

[9] European Commission, http://ec.europa.eu (visited December 2009)

[10] Fabian Ralf, Development of Techniques for Process Modeling With Abstract State Machines. Proceedings of the International conf. The Impact of European Integration on the National Economy, Faculty of Economics and Business Administration, Babe?-Bolyai University, Cluj-Napoca, 244-252, 2005.

[11] Fabian Ralf, Marcu Alexandru-Nicolae, Natural language processing implementation on Romanian ChatBot, Proceedings of the 9th WSEAS International Conference on SIMULATION, MODELLING AND OPTIMIZATION, Budapest Tech, Hungary, 2009, pg. 440-445, WSEAS Press, ISSN: 1790-2769, ISBN: 978-960-474-113-7

[12] Fabian Ralf, V. Crăciunean, Popa E. M.. Intelligent system modelling with total fuzzy grammars. Proc. of the 8th WSEAS International Conference, Mathematical Methods and Computational Techniques in Electrical Engineering (MMACTEE), 82-87, Bucharest, 2006.

[13] Facing The Challenge. The Lisbon strategy for growth and employment. Report from the High Level Group chaired by Wim Kok, November 2004, (visited December2009) http://ec.europa.eu/growthandjobs/pdf/kok_report_en.pdf

[14] FIPA - Foundation for Intelligent Physical Agents - Specifications, 2005, http://www.fipa.org/specifications/index.html (visited December 2009)

[15] Gigerenzer, G., R. Selten. Bounded Rationality. MIT Press, Cambridge, 2002.

[16] Kahneman, D. Maps of Bounded Rationality: Psychology for Behavioral Economics. Lecture (when receiving Nobel Prize; revised version). Stockholm, Nobel Foundation, 2002.

[17] Luck, M., P. McBurney, C. Priest. A Manifesto for Agent Technology: Towards Next Generation Computing. Autonomous Agents and Multi-Agent Systems, 9, 203-252, Kluwer Academic Publishers, 2004.

[18] Manuel Castells, The Information Age: Economy, Societyand Culture, Vol. 1, The Rise of the Network Society. Malden, Mass./Oxford, Blackwell, 2000.

[19] Maria Joăo Rodrigues, Europe, Globalization and the Lisbon Agenda, Edward Elgar Publishing Limited, 2009, ISBN 978-1-84844-199-6.

[20] Official site of the Lisbon Strategy http://portal.cor.europa.eu/lisbon/ (visited December 2009)

[21] Pah, I., A. Moiceanu, I. Moisil, B.E. Bărbat. Self-Referencing Agents for Inductive Non-Algorithmic e-Learning. Proc. of the 11th WSEAS International Conference on COMPUTERS (ICCOMP '07) (N.E. Mastorakis et al, Eds.), 86-91, Agios Nikolaos, Crete, 2007.

[22] Popa Emil. M., Fabian Ralf, Brumar Cristina, A Bounded Rational Review of Approximation and Undecidability in Economic Modelling, Proceedings of the 12th WSEAS International Conference on Computers - New Aspects of Computers, Heraklion, Greece, July 2008, pg. 303 308, Published by WSEAS Press, ISSN 1790 5109, ISBN 978 960 6766 85 5

[23] Rubinstein, A. Modeling Bounded Rationality. MIT Press, Cambridge, 1998.

[24] Simon, H.A. Models of Bounded Rationality. MIT Press, Cambridge, 1997.

[25] Towards knowledge societies, UNESCO World Report, UNESCO Publishing, 2005, ISBN 92-3-204000-X, http://unesdoc.unesco.org/images/0014/001418/141843e.pdf

[26] Tsang P. K. Edward, Computational Intelligence Determines Effective Rationality, International Journal of Automation and Computing, 2008.

[27] Windt, K., F. Böse, T. Philipp. Criteria and Application of Autonomous Cooperating Logistic Processes - Proc. of the 3rd Int. Conference on Manufacturing Research - Advances in Manufacturing Technology and Management, (J.X. Gao et al, Eds.), Cranfield, 2005.

[28] Zadeh L.A., et al. (Eds.). From Natural Language to Soft Computing: New Paradigms in Artificial Intelligence, Romanian Academy Publishing House, 2008.

[29] Zambonelli, F., A. Omicini. Challenges and Research Directions in Agent-Oriented Software Engineering. Autonomous Agents and Multi-Agent Systems, 9, 253-283, Kluwer Academic Publishers, 2004.

Int. J. of Computers, Communications & Control, ISSN 1841-9836, E-ISSN 1841-9844
Vol. V (2010), No. 5, pp. 719-726

A Homogeneous Algorithm for Motion Estimation and Compensation by Using Cellular Neural Networks

C. Grava, A. Gacsádi, I. Buciu

Cristian Grava, Alexandru Gacsádi, Ioan Buciu
University of Oradea
Faculty of Electrical Engineering and Information Technology
Oradea, Romania
E-mail: {cgrava,agacsadi,ibuciu}@uoradea.ro

Abstract: In this paper we present an original implementation of a homogeneous algorithm for motion estimation and compensation in image sequences, by using Cellular Neural Networks (CNN). The CNN has been proven their efficiency in real-time image processing, because they can be implemented on a CNN chip or they can be emulated on Field Programmable Gate Array (FPGA). The motion information is obtained by using a CNN implementation of the well-known Horn & Schunck method. This information is further used in a CNN implementation of a motion-compensation method. Through our algorithm we obtain a homogeneous implementation for real-time applications in artificial vision or medical imaging. The algorithm is illustrated on some classical sequences and the results confirm the validity of our algorithm.
Keywords: cellular neural networks, motion estimation, Horn & Schunck method.

1 Introduction

The motion estimation and compensation algorithms were developed for different applications as artificial vision, video information compression, medical imaging, digital and high-definition television, video-telephony, virtual-reality and multimedia techniques. Motion estimation allows one to reduce the temporal redundancy in a sequence of images in order to reduce the transmission rate and has been widely used in television signal coding (e.g. motion-compensated (MC) prediction, MC interpolation) and videoconference services [1]. To avoid this limitation in this paper we propose a fully parallel solution in order to realize the motion estimation and compensation, using CNN [2], as a competing alternative to classical computational techniques. The advantage of the algorithms that can be implemented on CNNs is that these kinds of neural networks already exists in hardware version [3], [4] and thus we can obtain real-time applications. In our case, because the motion estimation and compensation methods have generally a great computational cost, we develop homogeneous algorithms (that is the estimation part and compensation part are implemented in the CNN environment) for real-time applications that can be after that applied in artificial vision or medical imaging. After a small introduction, in the second part of this paper we present an overview of motion estimation and compensation methods, followed by a section that presents the CNN implementation of the Horn and Schunck motion estimation method and a section that presents the CNN implementation of the motion compensation. In the section dedicated to experimental results we present results that confirm the validity of our algorithm and we finalize our paper with a section of conclusions giving also some perspectives to our work.

2 Overview on motion estimation and compensation methods

The most used motion estimation methods are [5]:

- differential methods (or gradient methods); in this case the motion being estimated based on the spatial and temporal gradients of images [6], [7], [1];
- block-based methods (or correlative methods). These methods could be classified in phase-correlation methods and block-matching methods. In the case of phase-correlation methods, the motion is estimated based on the Fourier phase-difference between two blocks from two successive images. These methods are less used in practice because of the high noise-sensitivity. In the case of block-matching methods the location of the block (in the following or previous images) that best matches the reference block in the current image is searched, based on a certain matching or difference criteria. Both methods are usually applied in the case of a translation movement, but could be also adapted for other spatial models of the movement [8].

The principle of almost all motion estimation methods is that the brightness intensity of each pixel is constant along the motion trajectory or is modifying in a predictable way. This hypothesis of brightness intensity preservation of each point (x,y,t) along the motion trajectory can be expressed through the equation of *Displaced Frame Difference* (DFD), between the t and $t-1 = t-\Delta t$ instants [9]:

$$DFD(x,y) = \Phi(x-dx, y-dy, t-dt) - \Phi(x,y,t), \tag{1}$$

where $\Phi(x,y,t)$ denote the brightness distribution of the image at the moment t and $\mathbf{d} = [dx, dy]^T$ is the displacement vector between the moments t and $t' = t-\Delta t$ (dx and dy being the displacement vectors on x and y direction, respectively).

Differential motion estimation methods are based on spatial and temporal gradients of a sequence of images. If the brightness intensity of a pixel is not varying in time, then $d\Phi/dt = \Phi^t = 0$ [1]. The first order Taylor development of this last relation, has as result the "equation of movement constraint" EMC (or "optical flow equation" OFE) that links the spatial and temporal gradients of brightness intensity [9]:

$$\frac{\partial\Phi}{\partial x}\frac{\partial x}{\partial t} + \frac{\partial\Phi}{\partial y}\frac{\partial y}{\partial t} + \frac{\partial\Phi}{\partial t} = 0. \tag{2}$$

We can rewrite:

$$\Phi^x \cdot v_x + \Phi^y \cdot v_y + \Phi^t = 0, \tag{3}$$

where Φ^x and Φ^y are the spatial gradients, Φ^t is the temporal gradient of the brightness intensity and $v_x = dx/dt$, $v_y = dy/dt$ are the velocities on x and y directions [1].

As it can be observed in eq. 3, the optical flow equation has two unknowns (v_x, v_y), hence the system is under-determined leading to an ill-posed issue. In order to obtain both movement components (v_x, v_y), it has to be introduced a second constraint, to obtain a fully determined system (two equations with two unknowns). One of the possible constraints is offered by the well-known Horn & Schunck motion estimation method [1], which assumes that all the neighbor pixels have similar movement (we say that the velocity field is uniform or smooth). It results that it has to be minimized an energy:

$$E^2 = E^2_{flow} + \gamma E^2_{uniformity} \tag{4}$$

The first term correspond to the difference related to the projection of the velocity vectors on the spatial gradient (as in the EMC) and the second term correspond to the difference related

to a smooth field, g being the weighting term between the two terms. The uniformity constrain is expressed by the equation:

$$E^2_{uniformity} = \left(\frac{\partial v_x}{\partial x}\right)^2 + \left(\frac{\partial v_x}{\partial y}\right)^2 + \left(\frac{\partial v_y}{\partial x}\right)^2 + \left(\frac{\partial v_y}{\partial y}\right)^2 \quad (5)$$

$$E^2_{flow} = \left(\Phi^x \cdot v_x + \Phi^y \cdot v_y + \Phi^t\right)^2 \quad (6)$$

The solution is obtained after a Gauss-Seidel minimization [1]. Equation 6 will be minimized when the error (that means the difference between two successive values of (v_x, v_y), will be considered as being the minimal or when the maximum chosen number of iterations will be reached. This method is not limited to translations as block-matching method and the computations are shorter, but the movement amplitude has to be small (less than three pixels) because of the considerations regarding Taylor development. The solution to avoid the constraint regarding small amplitude of movement is to use the multi-resolution technique [1]. Using two consecutive frames ($\Phi_i(x, y, t_i)$ and $\Phi_f(x, y, t_f)$) of a sequence (Fig. 1), after the application of a motion estimation algorithm, as Horn and Schunck method, for each pixel it results an estimation of its movement in the both two directions (x,y) of the system of co-ordinates that is attached to image plane.

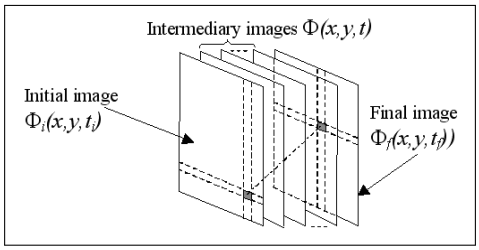


Figure 1: The illustration of motion compensation.

The purpose of motion compensation is that based on estimated motion information and starting from a reference image (the initial image Φ_i, in Fig. 1) to obtain an estimation of the real comparison image (the final image Φ_f) that was used in the process of motion estimation [10].

3 The CNN implementation of the Horn & Schunck motion estimation method

Regarding CNN gray-scale image processing generally, variational computing based template design is possible if the design constrains are respected [11], [12]. In order to analytically determine the template, cost functions or energies are used in the designing step. An important designing aspect is represented by the way to associate each energy function with one of the A, B, C or D template, as well as the way to chose the layers number of the CNN. Taking into account the characteristics of the existing CNN chip it is recommended to use only mono-layer CNN and only A and B templates. In the cost functions some weights can be introduced in order to maintain the state values in the linear zone of the state-output transfer characteristic. The motion estimation results using the two images, $\Phi_1(x, y, t)$ and $\Phi_2(x, y, t + \Delta t)$, a two-layer CNN structure and the *Hosch.tem* (see Fig. 2). After the cost function minimization [13], for the *Hosch.tem* it results:

- polarization images $Z_1 = \Phi^x\Phi^t$ and $Z_2 = \Phi^y\Phi^t$,

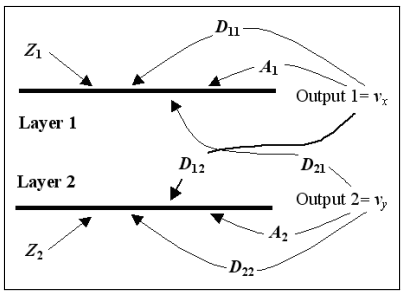


Figure 2: Two-layer CNN structure for the proposed Horn & Schunck motion estimation method.

- nonlinear templates $\boldsymbol{A}_1$ and $\boldsymbol{A}_2$:

0	a	0
a	1-4a	a
0	a	0

where the parameter a also includes the γ parameter from the equation (6), obviously weighted with the constants that results at the energy minimization. Nonlinear $\mathsf{D}-\mathsf{type}$ template D is:

0	0	0
0	d_{kl}	0
0	0	0

where each element d_{kl} is expressed as follows: $d_{11}(v_x) = (\Phi^x)^2 \cdot v_x$, $d_{21}(v_y) = \Phi^x\Phi^y \cdot v_y$, $d_{22}(v_y) = (\Phi^y)^2 \cdot v_y$, and $d_{12}(v_x) = \Phi^x\Phi^y \cdot v_x$, respectively.

In Fig. 3 the images with the estimated motion in the case of the "taxi" real sequence are presented. The first two images represent two images of the well-known sequence in motion estimation reference, "Hamburg-taxi", and in the last two images the motion v_x and v_y (or displacement) images are presented. The two last images represent the displacement in the two spatial directions. The combination between these two images could be also represented as a single image with vectors corresponding to the displacement of each pixel of the reference image, as we will present in the section presenting other experimental results. For a better visualization, the motion images, v_x and v_y, are not calibrated in the CNN domain. If we want to use these images in image compensation we have to calibrate them in the CNN domain $[-1, +1]$.

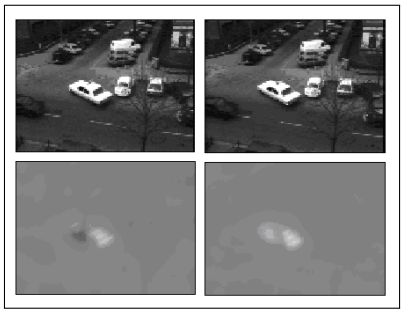


Figure 3: Two images of the Hamburg-taxi real sequence, used in our experiments.

4 The CNN implementation of motion compensation

In order to develop a motion compensation algorithm that can be directly implemented on a CNN chip, we have to decompose such an algorithm as elementary steps that can be then implemented on the existing hardware. As a result of the motion estimation process, a pixel could be stationary or can change its position in one of the eight elementary directions: N, N-E, E, S-E, S, S-W, W, N-W. After the application of a motion estimation technique, the pixels of the intermediary image frame $\Phi(x, y, t)$, that corresponds to any given moment $t \in (t_i, t_f)$, could be classified in the following categories (see Fig. 1), where t is the time-position of the intermediary image, between the initial image and the final image:

- Pixels of "*a*" type that has an identical position in the two consecutive images. These pixels does not change, at a given moment $t \in (t_i, t_f)$, neither their positions nor their values;
- Pixels of "*b*" type, that will moves as a result of the fact that the corresponding pixels in the two images that contains the motion information has a value greater than a current elementary value (that could be view as a quantum or a threshold). The value of these pixels is not changing, but inserting intermediary images between the initial and final image, Φ_i and Φ_f, their positions are changing successively with one elementary value (one quantum) corresponding to the spatial discretization. The maximum number of intermediary images that could be inserted equals the maximum number of elementary values (quantum) that could be identified in the images that contains motion (or displacement) information;
- Pixels of "*c*" type are those pixels that will change their values because will be covered by the pixels that will arrive in that position, overlapping the initial pixel:

$$c(t) = \mathrm{shift}(b(t)) \tag{7}$$

- Pixels of "*d*" type, with unknown values, that are the result of the displacement of "*b*" type pixels, that liberates a location but there is no pixel arriving in that location. In each step of the movement, the value of these pixels could be determined through spatial CNN spline-cubic interpolation [10]:

$$d(t) = \lfloor d(t-1) \cdot \overline{c}(t) + b(t) \cdot \overline{c}(t) \rfloor \cdot b(1) \tag{8}$$

- Pixels of "*e*" type that at the current time during the processing will have the same value as in the initial image Φ_i, due to the movement of the pixels (arrivals and departures of the pixels). The values of these pixels will be restored from the initial image:

$$e(t) = c(t-1) \cdot b(t) \cdot \overline{c}(t) \cdot \overline{d}(t) \tag{9}$$

Each intermediary step has as result an associate image and to create this intermediary image it has to be done the following operations:

- determine the "*c*" type pixels, that is the displacement with one pixel in the direction resulting from the motion information;
- interpolation, in order to determine the values of unknown pixels, that is the "*d*" type pixels.

For each intermediary image, the value of a pixel results after the determination of the type of that pixel. The state of a pixel could change during the processing. The initial and final image,

Φ_i and Φ_f, and the images that contain the motion information have the same dimensions. In this paper, all images are converted to standard CNN gray-scale images, taking values between -1 to +1 (see Fig. 5). The convention is that pixels with negative value are coding a displacement to the left (Fig. 5 a) or to up, respectively (Fig. 5 b) and the pixels having positive values are coding the displacements to the right (Fig. 5 a) or to down, respectively (Fig. 5 b). The values of the pixels coding the motion are multiples of the minimum detectable value of the motion. In order to detect the pixels that will change their position and to move the pixels, the *treshold.tem* and *shift.tem* templates family are used [14]. The determination of the value of pixels of "*d*" type could be made for each intermediate image or only to the final image. In order to avoid the modification of the pixels of "*a*" type or in order to restore the pixels of "*e*" type, some mask-images are created during the processing, using the equations (7), (8) and (9).

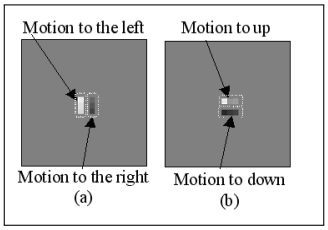


Figure 4: Conventions in the images containing the motion information.

5 Experimental results

In this section some experimental results obtained by using the "CadetWin" (CNN Application Development Environment and Toolkit under Windows [14]) are presented. The images containing the motion estimation, v_x and v_y, are calibrated in the CNN domain $[-1, +1]$. The processing time depends on the number of interpolations and on the number of the motion estimation quantum, that results after spatial discretization, that is on the total number of images inserted between the initial and final image, Φ_i and Φ_f. Due to parallel processing, this total processing time is independent by the dimensions of the original images or by the number of moving pixels. In order to illustrate the implemented method in the case of real images and in the case of a complex movement, starting from an initial image of a well-known "tennis-table" sequence, we have simulated a complex motion, using the Free Form Deformation principle, resulting (see Fig. 5) three real images of a sequence (Φ_1, Φ_2, Φ_3). In Figure 5 we also represented the "Motion Estimation Field" obtained after applying our CNN implementation of the Horn and Schunck motion estimation method. Starting from the first image of the real sequence (Φ_1) and this "Motion Estimation Field" we can obtain the "Motion Compensated image" ($\widehat{\Phi}_2$), that represent an estimation of the real image (Φ_2).

As it can be observed, the biggest errors between the real image and the motion compensated image ($\Phi_2 - \widehat{\Phi}_2$) are located in the region with a high gradient of intensity that usually corresponds to the regions with different motion. Another cause oh these errors could also be the discrete nature of the image spatial support and the interpolations that are necessary.

6 Conclusions

Generally, in the case of serial implementation of a motion compensation algorithm, the processing time depends on the image dimensions. In the case of our CNN motion estimation and compensation algorithm, that uses only 3×3 linear templates, the algorithm can be directly

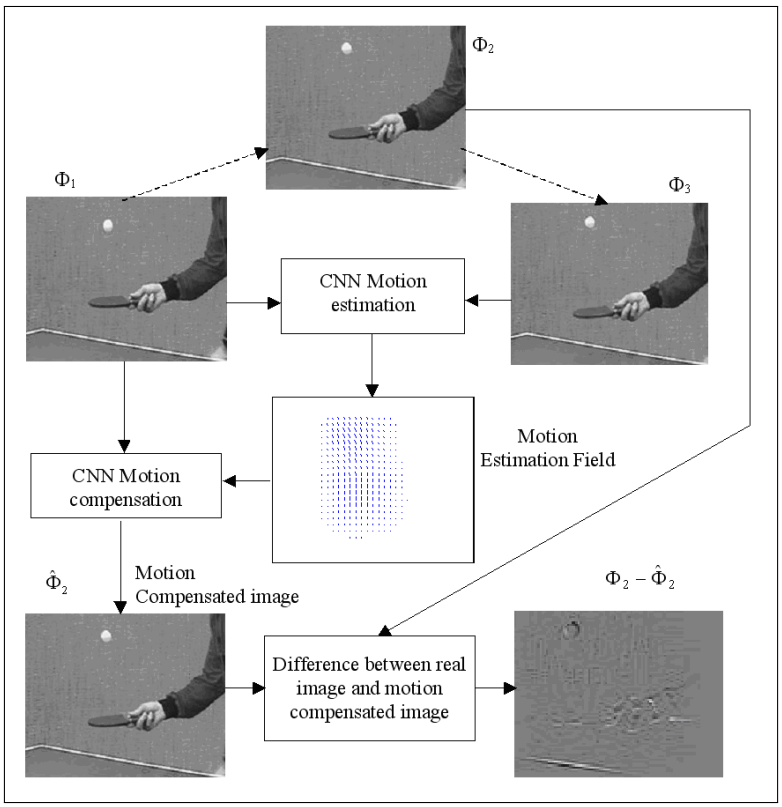


Figure 5: Motion estimation and compensation using CNN: principle and results.

implemented on the CNN-Universal Chip [4] and thus the image processing become completely parallel. The advantage of using the CNN hardware platform is that the total processing time doesn't depend on image dimensions, being dependent only on the number of displacement steps that has to be performed and thus we can obtain real-time applications with applications in artificial vision and medical imaging. Taking into account that our CNN motion estimation and compensation algorithm is based on the cost functions minimization (usually partially differential equations) resulting nonlinear templates, our attention is focused on the FPGA implementation of our algorithm, on a digital emulator of the CNN [11], [12].

7 Acknowledgement

This work was partially supported by a grant from the Romanian National University Research Council, PNCDI Program ID-668/2008.

Bibliography

[1] Horn B.K.P. and Schunck B.G, Determining Optical Flow, *Artificial Intelligence*, Vol.17, pp. 185-203, 1981.

[2] Chua L.O. and Yang L, Fuzzy Control Rules in Convex Optimization, *IEEE Transactions on Circuits and Systems*, Vol.35, pp.1257-1290, 1998.

[3] Cembrano G.L., Rodríguez-Vázquez A., Espejo-Meana S., and Domínguez-Castro R., ACE16k: A 128×128 Focal Plane Analog Processor with Digital I/O, *Int. J. Neural Syst.,* Vol.17, Issue 6, pp. 427-434, 2003.

[4] Roska T. and Chua L.O., The CNN universal machine: an analogic array computer, *IEEE Transactions on Circuits and Systems,* Vol.40, pp. 167-173, 1993.

[5] Konrad J., Motion detection and estimation, *Image Processing Handbook, Networking and Multimedia,* pp. 207-227, 2000.

[6] Bruhn A., Weickert J., Feddern C., Kohlberger T., Schnorr C., Real-time optic flow computation with variational methods, *Computer Analysis of Images and Patterns,* pp. 222-229, 2003.

[7] Brox T., Bruhn A., Papenberg N., Weickert J., High accuracy optical flow estimation based on a theory for warping, *ECCV,* pp. 25-36, 2004.

[8] Wei W., Hou Z.-X., Guo Y.-C., A displacement search algorithm for deformable block matching motion estimation, *Proc. of IEEE International Symposium on Communications and Information Technology,* pp. 457-460, 2005.

[9] Barron J.L., Fleet D.J., Beauchemin S., Performance of Optical Flow Techniques, *International Journal of Computer Vision,* Vol. 12, Issue 1, pp. 43-77, 1994.

[10] Grava C., Gacsádi A., Gordan C., Maghiar T., Bondor K, Motion Compensation using Cellular Neural Networks, *Proc. of the European Conference on Circuit Theory and Design (ECCTD),* Vol. I, pp. I-397-I-400, Krakow, Poland, 2003.

[11] Kincses Z., Nagy Z., Szolgay P, Implementation of nonlinear template runner emulated digital CNN-UM on FPGA, *Proc. of the 10th International Workshop on Cellular Neural Networks and Their Applications,* pp. 186-190, Istanbul, Turkey, 2006.

[12] Nagy Z., Vörösházi Zs., Szolgay P., Emulated Digital CNN-UM Solution of Partial Diferential, *Int. Journal of Circuit Theory and Applications,* Vol. 34, Issue 4, pp. 445-470, 2006.

[13] Gacsádi A., Grava C., Tiponut V., Szolgay P., A CNN implementation of the Horn & Schunck motion estimation method, *Proc. of the 10th International Workshop on Cellular Neural Networks and Their Applications,* pp. 381-385, Istanbul, Turkey, 2006.

[14] *** CadetWin, CNN application development environment and toolkit under Windows. Version 3.0, Analogical and Neural Computing Laboratory, Hungarian Academy of Sciences, Budapest, 1999.

Int. J. of Computers, Communications & Control, ISSN 1841-9836, E-ISSN 1841-9844
Vol. V (2010), No. 5, pp. 727-734

Towards Low Delay Sub-Stream Scheduling

W. Guofu, D. Qiang, W. Jiqing, B. Dongsong, D. Wenhua

Wu Guofu, Dou Qiang, Wu Jiqing,
Ban Dongsong,Wenhua Dou
National University of Defense Technology
School of Computer Science
Changsha, Hunan, P.R.China
E-mail: {gfwu,qdou,jqwu,dsban,whdou}@nudt.edu.cn

Abstract: Peer-to-Peer streaming is an effectual and promising way to distribute media content. In a mesh-based system, pull method is the conventional scheduling way. But pull method often suffers from long transmission delay. In this paper, we present a novel sub-stream-oriented low delay scheduling strategy under the push-pull hybrid framework. First the sub-stream scheduling problem is transformed into the matching problem of the weighted bipartite graph. Then we present a minimum delay, maximum matching algorithm. Not only the maximum matching is maintained, but also the transmission delay of each sub-stream is as low as possible. Simulation result shows that our method can greatly reduce the transmission delay.
Keywords: P2P streaming, scheduling, sub-stream, weighted bipartite graph, matching

1 Introduction

The emerging peer-to-peer (P2P) systems have appeared to be the most promising driving force for video streaming over the Internet. By distributing the workload to a large number of low-cost computing hosts such as PCs and workstations, one can eliminate the need for a costly centralized server and at the same time improve the system's scalability. There are already commercial products emerging, e.g., *PPLive* [1], *CoolStreaming* [3].

Any P2P streaming system consists of two distinct but related components: (i) an Overlay Con- struction mechanism, and (ii) a Content Scheduling mechanism. To improve the performance of P2P streaming systems, many studies focus on the overlay construction. However, the content scheduling mechanism also has greatly impaction on the performance. Carefully designed scheduling mechanism could have a better tradeoff among maximum streaming rate, minimum transmitting delay and control overhead. The overlay construction falls into two categories: tree(s) and mesh. In the mesh overlay, the pull method which is very similar to that of Bit-Torrent protocol [2]is widely used. Recent result [5]shows that the protocol of unstructured mesh overlay outperforms the traditional multi-tree approaches much in many aspects.

Several media content scheduling mechanisms for mesh-based system have been proposed. *CoolStreaming/DONet* [3] and *Chainsaw* [4] proposes pull-based scheduling framework. JM Li [6] designs a scheduling algorithm to manipulate the order of data blocks to improve the transmission efficiency in P2P streaming system. M Zhang [7] defines priorities for different blocks according to their rarity property and their emergency property, and tries to maximize the average priority sum of each node. Based on the traffic from each neighbor, a node in *GridMedia* [8] subscribes the pushing packets from its neighbors at the end of each time interval. *Pulsar* [9] combines push-based operations along a structured overlay with flexibility of pull operations. *LStreaming* [10] uses sub-streams in the push-pull streaming system.

This paper presents an effectual content scheduling strategy which can largely reduce the content transmission delay meanwhile remaining the main advantage of the pure pull method. First, the original stream is divided into k sub-streams, and each sub-stream has the same rate. A node requests one sub-stream instead of one content block from its neighbors. Then the sub-stream scheduling problem is transformed into the matching problem of weighted bipartite graph. The well known Hungarian Algorithm which solves the maximum matching problem is ameliorated. Not only maximum matching is reserved by the new improved algorithm, but also the transmission delay of each sub-stream is as low as possible. The main difference between our scheme and *LStreaming* is the choice of available neighbors.

The following paper is organized as follows: Section 2 describes our motivation. In Section 3, we present our algorithm for the sub-stream scheduling problem. Simulation result shows In Section 4. Finally, we conclude our work in Section 5.

2 Motivation

2.1 Delay analysis of the pull-based scheduling strategy

The main drawback of pull-based scheduling strategy is that content blocks suffer from long delay. Now we give a quantitative analysis showing why this strategy resulting in long transmission delay. In such system, the media content is divided into equal size blocks, each of which has a unique sequence number. Every node (including the source node) periodically broadcasts all of its neighbors a bit vector called Buffer Map (BM) which represents the availability of useful blocks in its buffer pool. According to the announcement, each node decides from which neighbor to ask for which blocks, and periodically sends requests to its neighbors for the desired blocks. When a neighbor receives the request, it puts the desired blocks into its output queue, waiting to send out. For the efficiency of information exchange, BM and requests will only be sent periodically so that dozens of packets can be mapped into a single packet. We denote the interval between two buffer map packets or two request packets as T.

Figure 1 shows a typical process that a content block goes from one node to a neighbor. At the time t_1, a fresh block arrives in node A. Because the BM packets will only be sent at the beginning of each time slot, the useful information will wait until the time t_2, when a new time slot begins. After a period of time t_{d1}, at the time t_3, the BM packet reaches at node B. Time t_{d1} contains two components: propagation latency and transmit latency. The propagation latency lies on the length of links between node A and node B, while the transmit latency lies on the length of packet and the available bandwidth. The width of the shadow represents the transmit latency. Because the length of BM and request packets is small, we ignore the latency. At the time t_3, node B would make a design whether or not to request the fresh block from node A. As the same reason, node B sends the request packet at the time t_4. At the time t_5, node A receives the request packet, and then it puts the fresh block to the output queue. After a queue delay t_w, the fresh block is sent out. Because the length of content block is large, we can't ignore the transmit latency of the packet. At the time t_7, the fresh block is received by node B. Now we can compute the one hop delay, which is the interval between the time t_1 and t_7. Apparently, t_p and t_r are uniform random variable on the length T of time slot, and they are independent. Their average values are both $T/2$. We suppose the packet with largest waiting time will be sent first, the blocks before the time t_1 reached at node A at the same time slot will be sent before the fresh block, so the queue delay t_w is $T - t_p$. Thus, the average latency $t_{1-delay}$ for one packet in one hop can be computed as

$$t_{1-delay} = E[t_p + t_{d1} + t_r + t_{d2} + t_w + t_{d3}] = \frac{3T}{2} + 3\overline{t_d} + \frac{l}{u} \tag{1}$$

Where $\overline{t_d}$ is the average end-to-end delay, and l is the length of content block, and u is the average upload bandwidth of nodes.

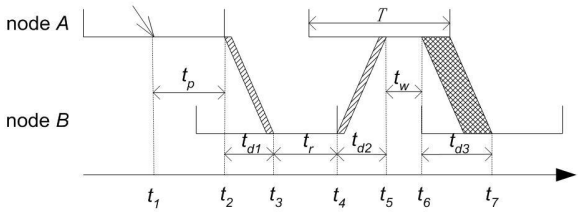


Figure 1: one hop latency

2.2 Using sub-stream to reduce the one hop latency $t_{1-delay}$

We can reduce the latency by adopting a shorter cycle T', but this will bring in more information packets. Meanwhile the three end-to-end delays can't be avoidable. A naive solution is that when node A receives the fresh content block, it directly sends it to node B. In this case, the one hop latency is only $\overline{t_d} + \frac{l}{u}$. But node B may receive repeated blocks from different neighbors under blind packets scheduling. An efficient way is that node A sends specifically sub-stream(s) to node B, while other neighbors send another sub-stream(s) to node B. Different sub-streams are not intersected. Node B decides to subscribe which sub-stream from which neighbor, and requests for sub-stream(s) instead of content block(s). Theoretically, node B requests for sub-stream(s) only once, and then just waits to receive media content. Figure 2 shows the creation

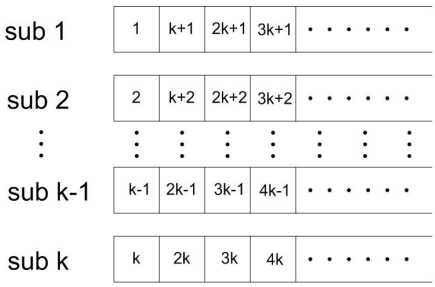


Figure 2: content blocks in different sub-streams

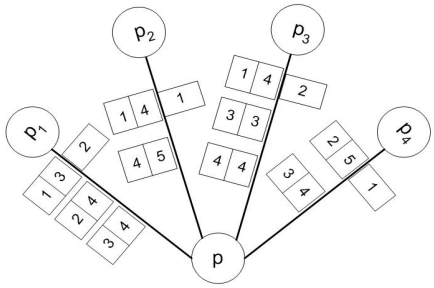


Figure 3: the information peerp collects

of sub-streams. The original media content is divided into blocks with equal size, and each block has a unique sequence number, counting from number 1. Suppose we divide the original stream into k sub-streams, then what content blocks each sub-stream has is expressly described in figure 2. Which sub-stream a content block with sequence number m belonging to is determined by the following function $f(m)$.

$$f(m) = \begin{cases} m \mod k & \text{if } x \mod k \neq 0 \\ k & \text{else} \end{cases} \tag{2}$$

3 Sub-stream Scheduling

In our framework, participating peers adopt gossip protocol to self-organize into unstructured mesh overlay. In this section, we study the problem of sub-stream scheduling, that is to say a

node subscribe which sub-stream from which neighbor. We suppose in the initial period, the node uses pull-based method to get the content blocks. After that, push-based method works.

3.1 Exchanging information

Peers periodically broadcast sub-stream information and available upload bandwidth to neighbors. The period can be long, because other cases may trigger a node to broadcast the information or send the information to chosen neighbors. When a node joins the system, all of its neighbors send information packets to it actively. The information includes the sub-streams the neighbor can provide $\{s_{ij}\}$, latency of each sub-stream from the source to the neighbor $\{h_{ij}\}$ and the available upload bandwidth of the neighbor $\{c_i\}$. s_{ij} =1 presents $neighbor_i$ can provide $sub-stream_j$, else $s_{ij} = 0$. h_{ij} presents the number of hops of $sub-stream_j$ from the source to $neighbor_i$. If $s_{ij} = 0$ then $h_{ij} = +\infty$. c_i is not the real bandwidth, but it presents the multiples of one sub-stream rate. For example, $neighbor_i$'s available upload bandwidth can support transmitting two sub-streams to local node, then $c_i = 2$. After a short period, the local node gets information from all of its neighbors. Suppose the original stream is divided into 4 sub-streams. Figure 3 shows the information that $peer_p$ collects from all of its neighbors. The information on the left side of every connection presents the sub-streams neighbors can provide and their latency starting from the stream source node, while the number on the right side presents how many sub-streams the neighbor's available upload bandwidth can support. For example, in figure 3, neighbor p_1 can provide sub-stream 1, 2, 3 respectively, and the corresponding latency is 3, 4, 4 hops respectively. Meanwhile the available bandwidth between p_1 and p can support 2 sub-streams.

3.2 Constructing weighted bipartite graph

More than one neighbors can provide the same sub-stream, so we should make a decision which neighbor is more suitable as a provider for one sub-stream. Here we transform the sub-stream scheduling problem into the matching problem of weighted bipartite graph. First, we construct the weighted bipartite graph, according to the information collecting from neighbors. The weighted bipartite graph can be expressed by a quadruple group $G = (X, Y, E, W)$, where X presents the set of sub-stream providers, Y presents the set of sub-streams, E presents the set of connections between X and Y, and W presents set of weight on each connection. There is at most one connection between any two elements in X and Y respectively. So if a neighbor's upload bandwidth can support several sub-streams, we should characterize this situation in our bipartite graph. Here, we allow same elements to coexist in set X to deal with this situation. If $neighbor_i$'s available upload bandwidth can support c sub-streams, then we copy c same elements of $neighbor_i$ into set X. When all neighbors are put into set X, we complete the construction of set X. Next we build the connections between X and Y. If any element $x \in X$ can provide sub-stream $y \in Y$, then a connection element $e = (x, y)$ is added into connection set E. The corresponding weight of connection e is the hops of sub-stream y from stream source to provider x. We use the example in figure 3 to illustrate the construction of weighted bipartite graph. Neighbor p_1's available upload bandwidth can support 2 sub-streams, so there are 2 p_1 elements in the set X, as shown in figure 4. Neighbor p_1 can provide sub-stream 1, 2, 3 respectively, and the corresponding latency is 3, 4, 4 hops respectively. Accordingly, in the weighted bipartite graph, both of the two provider p_1 have connections with sub-stream 1, 2, 3 respectively, and the weight on each connection is 3, 4, 4 respectively. The whole weighted bipartite graph transformed from figure 3 is shown in figure 4. The number on each edge presents the weight of each edge.

3.3 Minimum delay, maximum matching algorithm (MDMMA)

After the weighted bipartite graph has been constructed, we try to find the best matching. Here we mainly consider two targets: first, more sub-streams should be transmitted in parallel, because this can make use of peers' upload bandwidth as much as possible, speeding up the transmission of the media content; second, hops of each sub-stream from source to the node should be as little as possible, because this can lead to lower delay. Existent matching algorithms of bipartite graph only care one of the two targets. In this paper, we enhance the Hungarian Algorithm which is the well known algorithm to solve the maximum matching of bipartite graph. The ameliorated Hungarian algorithm first insures that the matching is the maximum matching, and then it tries to find the lowest latency matching. The algorithm in detail is described in table 1.

The main changes occur in step 1.2 and step 1.4. In step 1.2, when there are several candidate "uncheck" providers, we choose the provider whose connected edge has the minimum weight. In step 1.4, to find the augment chain, we start from the vertex whose connected edge has the minimum weight, making sure that the edge with the lowest weight is put into the matching. Although the running time of the algorithm is $O(n^3)$, n is usually is less than 30, so the peer can find the appropriate neighbors quickly.

Table 1 Min Delay Max Matching Algorithm

Min Delay Max Matching

input $G = (X, Y, E, W)$

output M

0 set $M = \phi$, and set all vertexes in G with no label.

1.1 let $S = \{x/x \in X$, and exist $y \in Y$, $(x, y) \in M\}$. if $X/S = \phi$, then finish;
else label every elements $x \in X/S$ with "-1" and "uncheck".

1.2 if all $x \in X$ is checked, finish;
else choose the "uncheck" x_i whose connected edge has the minimum weight.

1.3 if all $y \in \{y/y \in Y$, and $(x_i, y) \in E\}$ have been labeled,
then label x_i with "checked", goto step 1.2.

1.4 let $P = \{y/y \in Y$, and $(x_i, y) \in E$, and y is not labeled$\}$, label all $y \in P$ with "i".
let $Q = \{y/y \in Y$, and exist $x \in X, (x, y) \in M\}$.
if $P/Q \neq \phi$, then choose $y_j \in P/Q$ whose connected edge has the minimum weight, goto step 2;
else for every $y_j \in P$, label $x_p((x_p, y_j) \in M)$ with "j" and "unchecke", label x_i with "checked", goto step 1.2.

2 find the augment chain C starting from y_j until found $x(\in S)$ with label "-1", $M = M \oplus C$, cancel labels of all vertex in G, goto step 1.1.

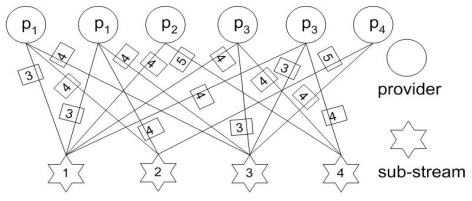


Figure 4: weighted bipartite graph

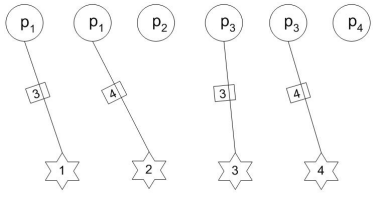


Figure 5: min delay, max matching

The ameliorated Hungarian algorithm is heuristic. It can insure that the matching is the maximum matching, but it can't insure that the matching is minimum delay matching theoretically. Applying the Min Delay Max Matching Algorithm to the weighted bipartite graph in figure 4, we get the following matching as shown in figure 5. According to the result, neighbor p_1 provide sub-stream 1, 2 to peer p, and neighbor p_3 provides sub-stream 3, 4 to peer p

respectively.

3.4 Broadcasting new sub-stream information

When a node decides to subscribe which sub-stream from which neighbor, it sends sub-stream request to the appointed neighbor. If the neighbor rejects the request, it cuts the matching connections in the weighted bipartite graph, and finds another provider with lowest latency. It requests again until it subscribe the sub-stream successfully or there is no useful connection. Then it broadcasts the subscribed sub-streams to its neighbors. The hop of each sub-stream in the information packets is adding the accepted hops by 1. When a neighbor accepts the request, it sends out the content blocks belonging to the desired sub-stream in the push window. If there is sub-stream(s) which can't find proper provider, the node should find more new neighbors.

3.5 Other discussion

Monitor neighbor's available upload bandwidth. We measure the interval of consecutive packets of the same sub-stream monitor neighbor's upload bandwidth. If the interval remains changeless (or change slightly), we suppose the bandwidth of the neighbor is affluent. If the interval changes greatly, we suppose the bandwidth of the neighbor is deficient. The rational reason behind this conclusion is that: if the bandwidth is affluent, there will no queue in the neighbors, little burst will happen, so the interval is changeless; if the bandwidth is deficient, burst will often happen, which leading to changing intervals.

Request missing blocks from neighbors with affluent bandwidth. Packets in transmission may be lost. When the missing content blocks enter into the pull widow, we request the block explicitly from the neighbor with affluent available upload bandwidth.

Sub-stream re-scheduling. Due to network dynamics, neighbors' upload capacity may fluctuate. When a neighbor is not competent as a sub-stream provider, a new provider should be chosen. Also we choose the appropriate provider from the neighbors with affluent. The neighbor who can provide sub-stream with fewer hops has higher priority.

4 Performance Evaluation

4.1 Simulation settings

We use the random model of GT-ITM [11] to generate the underlying topology with 5000 routers. Each link transmitting delay is set with a uniform random variable within [10ms, 100ms]. We randomly choose 2000 routers and connect one peer with one router. A peer randomly selects 10 to 15 other nodes as its neighbors to construct the mesh overlay. We set the playback rate of the original stream is 512kbps, and divide the stream into 32 sub-streams, each sub-stream with rate of 16kbps. The size of each content block is 1k bytes. The upload capacity of video source is 5 Mbps. All peers are assumed to be DSL users with three type of available upload bandwidth of 1 Mbps, 512 kbps and 256 kbps. These three types of peer represent 15%, 60% and 35% of the total peers. We suppose peer's download capacity is infinite. We run the simulation in the environment of Matlab7.0.

4.2 Simulation result

Here we compare our algorithm MDMMA with the pure pull-based method. We mainly care about the transmission delay of the stream. Figure 6 shows the cumulative distribution function of the content block transmission delay of the two methods in steady state. We can see that our

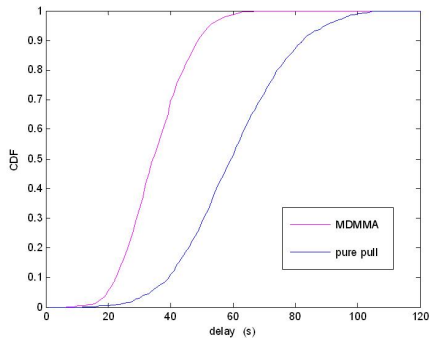


Figure 6: content block transmission delay CDF

algorithm MDMMA reduce the transmission delay greatly. For example, when 90% of the nodes get the content blocks, it only costs less than 50 seconds in our method, while in the push-based method, it costs more than 80 seconds. We can conclude that our proposed method is much better than pure pull-based method in the delay performance.

5 Conclusion

P2P streaming system consists of two components: (i) Overlay Construction Mechanism, and (ii) Content Scheduling Mechanism. Studies show that mesh-based system is better than tree-based system especially in high churn rate of node. In this paper, we present new scheduling mechanism to improve the performance of mesh-based P2P streaming systems. Our contribution includes two folds: first, we transform the sub-stream scheduling problem into the matching problem of the weighted bipartite graph; second, we present a minimum delay, maximum matching algorithm. Not only the maximum matching is maintained, but also the transmission delay of each sub-stream is as low as possible. Simulation shows that our method can reduce the transmission delay greatly.

Bibliography

[1] *PPlive*, http://www.pplive.com/.

[2] *Bittorrent*, http://bitconjuer.com/.

[3] X.Zhang, J.Liu, and et al. "Coolstreaming/donet: A data-driven overlay network for efficent media streaming". *In Proc. of INFOCOM 2005*, US, pp.2102-2111, Mar.2005.

[4] V.Pai, K.Kumar, and et al. "Chainsaw: Eliminating trees from overlay multicast". *Peer-to-Peer System ˇô,* pp.127-140, Nov.2005.

[5] N.Magharei, R.Rejaie, and Y.Guo. "Mesh or multiple-tree: A comparative study of p2p live streaming services". *In Proc. of INFOCOM 2007*, USA, pp.1424-1432, May.2007.

[6] JM.Li, C.K.Yeo, and B.S.Lee. "Peer-to-peer streaming scheduling to improve real-time latency". *In Proc. of Multimedia and Expo*, China, pp.36-39, Jul.2007.

[7] M.Zhang, Y.Q.Xiong, and et al. "Optimizing the throughput of data-driven peer-to-peer streaming". *IEEE Transactions on Parallel and Distributed systems*, Vol.20, No.1,pp.97-110, May.2008

[8] M.Zhang, J.G.Luo, and et al. "A peer-to-peer network for live media streaming - using a push-pull approach". *In Proc. of the 13th annual ACM internatioan conference on Multimedia*, Singapore, pp.287-290, 2005.

[9] T.Locher, R.Meier, and et al. "Push-to-pull peer-to-peer live streaming". *In Proc. of DISC 07*, Germany, pp.388-402, 2007.

[10] Z.J.Li, Y.Yu, and et al. "Towards low redundancy push-pull P2P live streaming". *In Proc. of of ACM Sigcomm 2008 Demo*, USA, Aug. 2008.

[11] K.C.Ellen, W.Zegura and S.Bhattacharjee. "How to model an internetwork". *In Proc. of Infocom 1996*, USA, pp.594-602, 1996.

Int. J. of Computers, Communications & Control, ISSN 1841-9836, E-ISSN 1841-9844
Vol. V (2010), No. 5, pp. 735-743

Full-Text Search Engine using MySQL

C. Gyorodi, R. Gyorodi, G. Pecherle, G. M. Cornea

Cornelia Gyorodi, Robert Gyorodi, George Pecherle, George Mihai Cornea
Department of Computer Science
Faculty of Electrical Engineering and Information Technology
University of Oradea, Str. Universitatii 1, 410087, Oradea, Romania
E-mail: cgyorodi@uoradea.ro, rgyorodi@rdsor.ro, gpecherle@uoradea.ro, generalmip@yahoo.com

Abstract: In this article we will try to explain how we can create a search engine using the powerful MySQL full-text search. The ever increasing demands of the web requires cheap and elaborate search options. One of the most important issues for a search engine is to have the capacity to order its results set as relevance and provide the user with suggestions in the case of a spelling mistake or a small result set. In order to fulfill this request we thought about using the powerful MySQL full-text search. This option is suitable for small to medium scale websites. In order to provide sound like capabilities, a second table containing a bag of words from the main table together with the corresponding metaphone is created. When a suggestion is needed, this table is interrogated for the metaphone of the searched word and the result set is computed resulting a suggestion.
Keywords: full-text, search, MySQL, index, search engine, ranking, metaphone, Levenstein

1 Introduction

Before the advent of the search engine, users had to search manually through dozens or hundreds of articles to find the ones that were right for them. Nowadays, in our more user-centered world, we expect the results to come to the user, not the other way around. The search engine gets the computer to do the work for the user.

Full-text search is widely used for various services of the Internet. A high-speed and a more efficient full-text search technology are necessary because of the amount of increasing handled document and corresponding document data every day [6].

According to the MySQL manual, full-text is a "natural language search"; the words are indexed and appear to represent the row, using the columns you specified. As an example, if all your rows contain "MySQL" then "MySQL" won't match much. It's not terribly unique, and it would return too many results. However, if "MySQL" were present in only 5% of the rows, it would return those rows because it doesn't appear too often to be known as a keyword that's very used.

MySQL full-text search doesn't have too many user-tunable parameters. If you want more control over your search you will have to download the sources and compile them yourself after you've made the changes you wanted. Anyway, MySQL full-text is tuned for best effectiveness. Modifying the default behaviour in most cases can actually decrease effectiveness [1].

Our implementation comes to bring a plus of functionality to the basic search capabilities that MySQL offers by returning better quality results to the user. For small data sets to be searched, its performance can be easily compared with the one of the more advanced dedicated full-text search engines.

2 Full-Text Search

In text retrieval, full-text search refers to a technique for searching a computer-stored document or database. In a full-text search, the search engine examines all of the words in every stored document as it tries to match search words supplied by the user.

When dealing with a small number of documents it is possible for the full-text search engine to directly scan the contents of the documents with each query, a strategy called serial scanning. This is what some rudimentary tools, such as grep, do when searching.

However, when the number of documents to search is potentially large or the quantity of search queries to perform is substantial, the problem of full-text search is often divided into two tasks: indexing and searching. The indexing stage will scan the text of all the documents and build a list of search terms, often called an index, but more correctly named a concordance. In the search stage, when performing a specific query, only the index is referenced rather than the text of the original documents. Some indexers also employ language-specific stemming on the words being indexed, so for example any of the words "drives", "drove", or "driven" will be recorded in the index under a single concept word "drive" [2].

Before passing on to the search, we would like to talk a little about how MySQL full-text indexes work. A MySQL full-text query returns rows according to relevance. But what is relevance? It is a floating-point number based on formulas. Researchers have shown that these formulas produce results that real users want.

For every word that isn't too short or isn't a too common one, MySQL calculates a number to determine its relevance inside the text. To be noticed that MySQL will not increase weight if two keywords are close to each other. Local weight, global weight, and query weight are the only things that matter. MySQL can work with stemming but as we've seen from different tests it isn't working very well, usually bombarding the user with tons of irrelevant results.

There are three formulas we have to mention for full-text index [3].
local weight = (log(dtf)+1)/sumdtf * U / (1+0.0115*U)
global weight = log((N-nf)/nf)
query weight = local weight * global weight * qf

where:
dtf - How many times the term appears in the row
sumdtf -The sum of "(log(dtf)+1)" for all terms in the same row
U - How many unique terms are in the row
N - How many rows are in the table
nf - How many rows contain the term
qf - How many times the term appears in the query

Notice that local weight depends on a straight multiplier, the term-within-row frequency times the unique frequency.

But most simply: If a term appears many times in a row, the weight goes up.

Why does local weight depend on how many times the term is in the row? Think of the document you are reading now. I inevitably mention "MySQL" and "full-text" several times. That's typical: if words appear several times, they are likely to be relevant.

Notice that global weight depends on an inverse multiplier, the count of rows MINUS the count of rows that the term appears in. Put most simply: If a term appears in many rows, the weight goes down.

To illustrate this better we've chosen to give a practical example. For this example, we've used a smaller database in order to be able to calculate the relevancies. To see what is in a full-

text index, we can use the myisam_ftdump program. It comes with the standard distribution. The name of owner table is demo and it is located in dmp database. The table consists in 2 columns, the second one being the one with full-text index.

In order to see the local weights of the different words we have to use the command dump index. This command will return the data offsets and the word weights (Fig. 1).

```
C:\wamp\bin\mysql\mysql5.1.30\bin>myisam_ftdump c:\wamp\bin\mysql\mysql5.1.30\data\dmp\demo 1 -d
```

88	0.9560229 create	c0	0.9560229 mysql
0	0.8421108 database	0	0.8421108 open
88	0.9560229 engine	0	0.8421108 popular
44	0.9456265 feature	44	0.9456265 search
0	0.8421108 free	88	0.9560229 search
44	0.9456265 fultext	c0	0.9560229 search
c0	0.9560229 fultext	c0	0.9560229 slow
44	0.9456265 grate	0	0.8421108 source
44	0.9456265 indexing	0	0.8421108 world
0	1.4258175 mysql *		
88	0.9560229 mysql		

Figure 1: Dump index

In order to see the global weights of the words we have to use the command calculate per-word stats. This command will return for us a word count and the global weight (Fig. 2).

```
C:\wamp\bin\mysql\mysql5.1.30\bin>myisam_ftdump c:\wamp\bin\mysql\mysql5.1.30\data\dmp\demo 1 –c
```

1	1.0986123 create	3	-1.0986123 mysql **
1	1.0986123 database	1	1.0986123 open
1	1.0986123 engine	1	1.0986123 popular
1	1.0986123 feature	3	-1.0986123 search
1	1.0986123 free	1	1.0986123 slow
2	0.0000000 fultext	1	1.0986123 source
1	1.0986123 grate	1	1.0986123 world
1	1.0986123 indexing		

Figure 2: Calculate per-word stats

For the first formula: (log(dtf)+1)/sumdtf * U/(1+0.0115*U);
where Dtf is how many times the term appears in the row
"MySQL" appears 2 times in row 0
so log(dtf()+1) = 0.6931472 + 1 = 1.6931472
sumdtf - The sum of "(log(dtf)+1)" for all terms in the same row

"MySQL" appears 2 times in row 0, so add log(2)+1
"world" appears 1 times in row 0, so add log(1)+1
"popular" appears 1 times in row 0, so add log(1)+1
"open" appears 1 times in row 0, so add log(1)+1
"source" appears 1 times in row 0, so add log(1)+1
"database" appears 1 times in row 0, so add log(1)+1
"free " appears 1 times in row 0, so add log(1)+1

so sumdtf = log(2)+1 + (log(1)+1)*6 = 7.6931472

U - How many unique terms are in the row there are 7 unique terms in row 0
so U/(1+0.115*U) = 7/(1+0.0115*7) = 6.478482

local weight = 1.6931472 / 7.6931472 * 6.478482= 1.4258175 (check Figure 2 for the first occurrence of the term “MySQL” *)

For the second formula: log((N-nf)/nf);
where:
N - How many rows are in the table; there are 4 rows in the ’demo’ table
nf - How many rows contain the term; the term "special" occurs in 3 rows

log((4-3)/3)= -1.0986123 (check Fig. 3 for the term “MySQL” **)
Note that because this term appears in more than 50% of the rows it has a negative global weight. In an actual search, this term will practically be ignored, only the adjacent terms being representative.

3 Implementation

The primary table is the table where we keep all the data that has to be searched and ranked. Also, from this table we will create later the “bag of words” together with their metaphones in order to provide correction suggestions to the user. The primary table will consist of 6 columns, the first one being the row id and the others all being meant for full-text search. The structure of the table is:

ID - unique identifier for every row
URL - the URL where the text was taken from
TITLE - the title of the page the text was taken from
CONTENT - everything on the page that is not part of the formatting and functionality (only text, no HTML or other scripts, including the dealers and strong text)
HEADERS - H1,... contents from the page. They will have higher relevancy than STRONG and CONTENT but lower than URL and TITLE
STRONG - words that appear between B or STRONG tags. They will have higher relevancy than ordinary text from CONTENT but lower than all the others.

During the construction of the table in the indexing process we decompose every page to obtain the required fields. First we will obtain the URL and the TITLE fields. After we have those two we can get rid off the head part of the web page and move on to the body. Before parsing out the entire HTML and other script tags from the text, we will have to take out the header and the strong keywords that will have a higher relevancy in our future search. After we’ve done that we can parse all the remaining script tags and send the resulting data into the MySQL database.

An improvement to this table would be a new column that contains all the keywords linking to this page. This way you can extend the article relevancy behind the actual page. Those linking keywords can be considered the ones between the anchor tag or the most relevant keywords from that page (calculated by their density after we’ve previously removed the stop words) or a combination of both. However, this approach requires many more pages to be scanned and many of them could come from outside our website.

After the first indexing is complete we can create the full-text index on the table. From now on we can make full-text searches on that table. Any new insert into the database will

be attached on the fly internally by MySQL so that we don't have to worry about this [9]. The SQL syntax for doing this is: ALTER TABLE 'content' ADD FULLTEXT ('url','title', 'content','headers','strong')

Searching the database is not very hard. Actually we will only use a query to apply a search on the database and let MySQL do the rest. It is a lot harder to determine the best values for the weights of the results. Depending on those constants, we will have different search results as we will show in the following examples.

In order to query the database we will use the following query. The query is virtually composed of three parts. The first part is meant to determine the relevancies of every column in part, the second one is meant for determining the matching rows and the last one is for ordering. You can see the query in Fig. 3.

```
SELECT * ,
MATCH (`url`) AGAINST ('search term') AS relUrl,
MATCH (`title` ) AGAINST ('search term') AS relTitle,
MATCH (`content` ) AGAINST ('search term') AS relContent,
MATCH (`headers` ) AGAINST ('search term') AS relHeaders,
MATCH (`strong` ) AGAINST ('search term') AS relStrong
FROM `content`
WHERE MATCH (`content` , `url` , `title` , `headers` , `strong`) AGAINST ('search term')
ORDER BY relUrl *U + relTitle *T + relContent *C+ relHeaders *H+ relStrong *S DESC
```

Figure 3: Search query

relUrl, relTitle, relContent, relHeaders and relStrong are the different relevancies returned by MySQL after the preliminary search of every different column in part. The parameters U, T, C, H and S represent the importance of those relevancies in part.

The query will execute like this: first, different relevancies will be returned from the search on the individual columns. Those relevancies will be later used to calculate the order of the results. After we have got the relevancies we have to move on to the retrieval of all the rows that contain the terms that we searched. In order to do this we have to select all the matches inside the text fields ('content', 'url', 'title', 'headers', 'strong'). Once we have those results we will have to reorder them by relevance. In order to do this we will determine the ORDER BY argument using the following formula.

$$\text{relevance} = \text{relUrl} *U + \text{relTitle} *T + \text{relContent} *C + \text{relHeaders} *H + \text{relStrong} *S$$

Figure 4: Relevance formula

The relevance for every column should be carefully selected in order to achieve the best result order. The URL relevance is not so important in our opinion. We have added it because all the big search engines have it. We will treat it as it would have the same relevance as the page title. The next column on the relevance scale, after the URL and the Title would be the Headers relevance.

Let's think about a certainty search. We want to search for an article about "Mysql Fulltext". The script will first determine the pages that contain the keywords inside the Title and URL of the page because those are the most suitable for us. The next in importance is the Headers column.

The last two of the columns are the Content column and the Strong column. The Strong column contains keywords that the writer of that page thought they are of a higher importance, with strong relevance to the subject.

The order of the relevancies is clear. The plain content relevance is the lower one, followed by the keywords relevance (Strong), Headers, Title which is about the same to the URL relevance. The problems appear when we have to determine the best magnitude for those relevance parameters.

If the difference of the relevance parameters is too small, we will not have effective search results ordering. But if we fall into the other side, we can provide excessive ordering based only on the hierarchy we've determined for the parameters and less based on the relevancies of the results.

A schematisation of the proposed indexing and search algorithm is provided below:
A) Indexing
1. The web page is processed by reading its contents.
2. The internal links from the page are determined.
3. The unique internal links and full page contents are inserted in a 'pages' table.
4. The next link to follow is taken from the 'pages' table and the new corresponding page is processed by going to step 1, until there are no more unprocessed pages.
5. The 'pages' table is processed to determine the page title, headings and strong text and the results are placed in a new 'contents' table.

B) Full-Text Search
1. Query the 'contents' table against a search term.
2. Determine the relevancies of every column in part.
3. Determine the matching rows.
4. Order the results by relevance, calculated using the formula in Fig. 4. The parameters U, T, C, H and S are chosen by us and they are an indicator of the importance of each of the relevancies (URL, Title, Content, Headers and Strong Text).

4 Testing

To test the algorithm, we have set up a database of 320 MB (including data and index) and a total number of 11,670 records. The database was built by indexing a subset of the wikipedia.com website, using the indexing algorithm described in the previous section.

For testing purposes, we have chosen the following values for the U, T, C, H and S parameters. The user can set his own values. A higher value means a higher importance of that element:
U=1.14, T=1.14, C=1, H=1.3, S=1.2

The testing process was done by asking different users to give ratings to each of the search results. We have performed a total of 143 tests. The ratings given by the users proved that our algorithm is more relevant than the default MySQL method by 19.47% .

Here are some examples of tests done with specific keywords:
Keyword: "programming"
Results: Our algorithm (the left panel) returned relevant results in positions 1 and 2 (an article titled "Computer Programming" and another one "Application programming interface"). This was because our algorithm assigned a higher importance to the page title and URL (1.14) than the page contents (1). The right panel did not return many relevant results (a result which was close to our expectations was in position 2, but it was not very specific because it was about a specific programming language and not about programming languages in general).

Keyword: "universal serial bus"
Results: Our algorithm (the left panel) returned a relevant result in position 1 (an article titled

"Universal Serial Bus"). The right panel did not return any relevant results (instead it returned pages that contain "universal" only and not the whole search term).

5 Search Suggestions

In many cases, there appears the necessity to correct the spelling mistakes made by the users when they don't know exactly how to spell something or are just making a typing mistake. To do this, we will have to give them the closest correct form that best matches the initial search.

Metaphone is a phonetic algorithm, an algorithm published in 1990 for indexing words by their English pronunciation. The algorithm produces variable length keys as its output, as opposed to Soundex's fixed-length keys. Similar sounding words share the same keys.

Metaphone was developed by Lawrence Philips as a response to deficiencies in the Soundex algorithm. It is more accurate than Soundex because it uses a larger set of rules for English pronunciation. Metaphone is available as a built-in operator in a number of systems, including later versions of PHP. The original author later produced a new version of the algorithm, which he named Double Metaphone, that produces more accurate results than the original algorithm. However I will use the first version of the script because it returns accurate enough keys with a much better performance than the second one. [4]

In order to provide fast enough processing for the suggestions we will create a "bag of words" containing all the distinct words contained inside the content column. For each one of them we will attach the metaphone key. When a search is done we will select the candidates by calculating the metaphone of every word inside the search string and select the matches inside the metaphone table. After that operation we will have about 3 to 8 possible words to choose from.

The Levenshtein distance is defined as the minimal number of characters you have to replace, insert or delete to transform str1 into str2 . The complexity of the algorithm is O(m*n), where n and m are the length of str1 and str2 (rather good when compared to similar_text(), which is O(max(n, m)**3), but still expensive).

In its simplest form the function will take only the two strings as parameters and will calculate just the number of insert, replace and delete operations needed to transform str1 into str2. [5]

We will use this algorithm to determine the closest form from the list of candidates by selecting the candidate with the smallest Levenshtein distance to the searched keyword. To be noticed that this algorithm only works for word spelling mistakes. It does not work for mistakes like missing space, or composed words.

Missing space example: "MySQLFultext" will not suggest "MySQL Fultext"
Composed words example: "Pine apple" will not suggest "Pineapple"

Although some of the suggestion results can be hidden from the users because of the use of metaphone equivalence, this approach gives an incredible boost of performance. With a suggestion table containing over 130,000 rows it would take way too long to determine the Levensthein distance between the searched form and the correct form regarding the fact that, as we have said before, the Levenshtein algorithm has a complexity O(m*n). In owner case we would have a complexity O(m*n)*NrRows, where NrRows exceeds 130,000. Also, by applying this direct approach, we wouldn't solve the problems we've illustrated above.

By applying owner algorithm we can reduce the candidates from 130,000 to approximately 4-7 words. This represents a significant improvement of performance with only little disadvantages. Furthermore, those disadvantages could be counteracted by completing the suggestion table according to the newly searched term if the suggestion we provided is wrong and if there

is a sufficient similarity between the previous and the actual search, similarity which can be determined by dividing the Levenshtein distance with the string length.

A schematisation of the proposed search suggestions algorithm is provided below:
1. Create a table containing all the words in the contents column and their attached metaphone keys.
2. The metaphone of the searched keyword is determined using the metaphone function [4].
3. The Levenshtein distance is calculated as in [5].
4. The word with the smallest Levenshtein distance is chosen.

6 Comparison with other similar technologies

In comparison with other similar technologies we have strong points and weak points at the same time. It is impossible to have all of them at the same time and there has to be a compromise between costs, speed, maintenance and quality of the results.

In comparison with MySQL basic full-text search we have the advantage of a better ordered result set. However, this comes with the price of an increased search time because of the extra computation needed for the relevance ordering, according to Fig. 4. Both methods have the advantages of being free, with easy maintenance (the update of the full-text index is done on the fly) and they don't require special permissions on the server.

If we compare it with the free open-source SQL full-text search engine Sphinx, we can mention other differences. The common part of the two solutions is that they are both free. The strong points of our solution are that we have a better ordered result set and easier maintainability. The strong point for Sphinx is the search speed making it suitable for searching bigger databases. Another week point for Sphinx is the maintainability. The full-text index is not able to be updated on the fly so that it has to be recomputed periodically, task that can take from 15 minutes for small databases to a few hours for large ones.

7 Conclusions

By using our implementation it is possible to offer a complete and powerful search option. For the future, we would like to improve the search speed by using a dedicated full-text search engine.

Another improvement we would like to implement would be a new parameter which will determine the popularity of a result based on determining if a user found what he wanted or not. This will use a combination of the time the user spent on that certain page divided by the content length, how many times the user returned to this page or if a page is or is not a stop page (which can indicate that the user found what he wanted to know).

Also, in addition to search, we would also like to implement high-speed insert and delete, allowing full-text search to be used in the same way as other types of database search in which data can be searched right after data is inserted [7].

Another idea for future improvement is handling scientific document searches, especially mathematical text and mathematical operations [8].

For the suggestion tool we want to improve the algorithm in such a way to be able to correct as many mistakes as possible.

Bibliography

[1] Fine-Tuning MySQL Full-Text Search - http://dev.mysql.com/doc/refman/5.0/en/fulltext-fine-tuning.html

[2] Full text search - by Wikipedia - http://en.wikipedia.org/wiki/Full_text_search

[3] MySQL's Full-Text Formulas - by Database Journals - `http://www.databasejournal.com/features/mysql/article.php/3512461/MySQLs-Full-Text-Formulas.htm`

[4] Metaphone - by Wikipedia - http://en.wikipedia.org/wiki/Metaphone

[5] The Levenstein distance - http://us2.php.net/levenshtein

[6] Atlam, E.-S., Ghada, E.-M., Fuketa, M., Morita, K., Aoe, J., A compact memory space of dynamic full-text search using Bi-gram index, Computers and Communications, 2004. Proceedings ISCC 2004. Ninth International Symposium

[7] Ikeda, T., Mano, H., Itoh, H., Takegawa, H., Hiraoka, T., Horibe, S., Ogawa, Y., "TRMeister: a DBMS with high-performance full-text search functions", Data Engineering, 2005. ICDE 2005. Proceedings. 21st International Conference

[8] Misutka, J., Galambos, L., "Mathematical Extension of Full Text Search Engine Indexer", Information and Communication Technologies: From Theory to Applications, 2008. ICTTA 2008. 3rd International Conference

[9] D. Zmaranda, G. Gabor, Issues on Optimality Criteria Applied in Real-Time Scheduling, *International Journal of Computers Communications & Control*, ISSN 1841-9836, Suppl.S, 3(S):536-540, 2008.

Int. J. of Computers, Communications & Control, ISSN 1841-9836, E-ISSN 1841-9844
Vol. V (2010), No. 5, pp. 744-754

Complex Computer Simulations, Numerical Artifacts, and Numerical Phenomena

D.-A. Iordache, P. Sterian, F. Pop, A.R. Sterian

Dan-Alexandru Iordache, Paul Sterian, Andreea Rodica Sterian
Physics Department, University Politehnica of Bucharest,
313 Splaiul Independentei,
Bucharest 060042, Romania
E-mail: {daniordache2003,paul.sterian,andreea_rodica_sterian}@yahoo.com

Florin Pop
Lecturer, Computer Science Department, University Politehnica of Bucharest,
313 Splaiul Independentei, Bucharest 060042, Romania
E-mail: florin.pop@cs.pub.ro

Abstract: The study of some typical complex computer simulations, presenting one or more Complexity features, as the: a) symmetry breaking, b) nonlinear properties, c) dissipative processes, d) high-logical depth, e) self-organizing processes, etc allows to point out some several numerical artifacts, namely the: (i) distortions, (ii) scattering, (iii) pseudo-convergence, (iv) instability, (v) mis-leading (false) symmetry-breaking simulations and others. The detailed analysis of these artifacts allowed clarifying the numerical mechanisms of some such artifacts, which can be named in following numerical phenomena, because their basic features can be exactly predicted.
Keywords: Computer Simulations, Numerical Artifacts, Numerical Phenomena, Self-organizing Processes.

1 Introduction

We live in a computerized world, our civilization being a "civilization of Computers". Taking into account that computers control the work of all-present complex installations and devices, the appearance (due to some numerical phenomena) of some important distortions of the simulated processes lead usually to major failures of the technical installations. Particularly, the events referring to the erroneous computer (numerical) simulation and design of the flight of the Patriot Missile which failed (with disastrous results) - during the Gulf War in 1991 - to stop a Scud Missile [1] and the self-destruction of the European Space Agency's Ariane 5 rocket, at 37 seconds after its launch [2], were both assigned to computer errors [3] and to their associated numerical artifacts/artefacts/phenomena.

Given being the computer simulations are considerably cheaper than the experimental studies and they allow the prediction of the physical systems behavior even in inaccessible conditions, the computer simulations are widely used in technical studies. Because the modern (optimized) technical systems are complex [4], the computer simulations are also complex, and - for this reason - they generate specific numerical artifacts. In fact, there is a huge number of publications reporting such phenomena (usually related to some complex numerical simulations), as it can be easily found consulting some search systems, e.g. the Google system. As it results from Table 1, even eliminating by the use of quotation marks (") the "parasitic" published works entitled as numerical simulations of... boiling *phenomena*, etc, there remain very large numbers of published works in these fields.

This work deals with the study of possibilities to discover the mechanisms of the artifacts intervening in some complex simulations and to predict quantitatively the basic parameters of

Topics	No quotation marks	With quotation marks, e.g. "Numerical Artifacts"
Numerical Artifacts	1,180,000	9,370
Numerical Artefacts	126,000	2,440
Numerical Phenomena	4,160,000	1,340
Complex Simulations		3,000

Table 1: Numbers of published papers found by the Google search system (beginning of 2009)

the computer generating errors, transforming so the observed numerical artifacts/artefacts in the so-called numerical phenomena. We have to underline from beginning that the discovery of these mechanisms belong to the field of Numbers Theory and that many problems in the field of Numbers Theory are extremely difficult. E.g., the statement of the (Pierre de) Fermat's last (greatest) theorem was published (after his death, by his eldest son - Clément Samuel Fermat) in 1670 [4], but its solution was found only in 1995 [5] by the professor Andrew Wiles [6].

We will focus mainly to the study of the main features of the classical [7], [8] and of the newly found [9], [10] numerical phenomena associated to the Finite Differences (FD) simulations [11] of the pulses propagation through media with sharp interfaces and attenuative character, as well as of other numerical methods (as the random walk method, the gradient one, etc), applied to the study of different physical processes, as diffusion [12], [13], solitary wave propagation, applications to the evaluation of the parameters of some physical systems, etc.

2 Symmetry breaking in some computing programs

2.1 Symmetry breaking of the wave equation in ideal media

The Finite Differences (FD) discretization of the wave equation in ideal media:

$$\frac{\partial^2 w}{\partial \tilde{t}^2} = V_\Phi^2 \frac{\partial^2 w}{\partial x^2} \rightarrow \frac{w_{t+1} + w_{t-1} - 2w}{\tau^2} = V_\Phi^2 \frac{w_{i+1} + w_{i-1} - 2w}{\epsilon^2} \tag{1}$$

which is symmetrical relative to the space steps $i-1$, i, $i+1$ and the time steps $t-1$, t, $t+1$, if the FD velocity and the wave propagation one are equal:

$$V_{FD} = \frac{\epsilon}{\tau} = V_\Phi \tag{2}$$

Defining the Courant's number [7] by means of the relation:

$$C = \frac{V_\Phi}{V_{FD}} \tag{3}$$

one finds easily that if: (i) $C > 1$, there will intervene *instabilities*, because the FD schema does not use all information received at the observation point ($V_\Phi > V_{FD}$), (ii) $C < 1$, there will intervene *distortions*, because for $V_\Phi < V_{FD}$, the FD schema uses more information than it receives, and this additional information acts as a jamming, (iii) $C = 1$, we have an ideal FD schema (stable and convergent), because this schema has all necessary physical information and nothing more! One finds that *the symmetry breaking for values different than 1 of the Courant number leads to the "classical" numerical phenomena.*

2.2 Symmetry breaking of the smoothing model of a sharp interface

Consider a sharp interface between 2 homogeneous elastic media. If the method of Finite Differences (FD) is used, then - in order to avoid the use of Dirac function - a certain smoothing of the sharp interface is necessary, spreading it over 2 or 3 FD nodes, whose indices i are denoted as $I-1$, I and $I+1$ (see Figure 1).

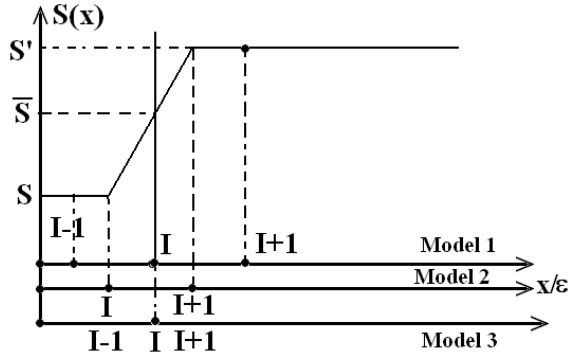


Figure 1: Smoothing models of the sharp 1-D interfaces (see also [14]-a).

Then the differential equation $\rho(x)\frac{\partial^2 w}{\partial t^2} = \frac{\partial}{\partial x}\left[S(x)\frac{\partial w}{\partial x}\right]$ of the elastic pulses propagation through an in-homogeneous medium becomes:

$$\tilde{\rho}_i \frac{w_{t+1} + w_{t-1} - 2w}{\tau^2} = \left\langle \frac{\partial S}{\partial x} \right\rangle_i \frac{w_{i+1} + w_{i-1}}{2\epsilon} + \tilde{S}_i \frac{w_{i+1} + w_{i-1} - 2w}{\epsilon^2}. \tag{4}$$

where $\tilde{\rho}_i = \langle \rho \rangle_i$, $\tilde{S}_i = \langle S \rangle_i$ and $\left\langle \frac{\partial S}{\partial x} \right\rangle_i$ are the chosen average values around the FD node i. The chosen expressions of these average values (see Table 2) can succeed or not to remake the (apparent) symmetry of the propagation medium in the frame of the FD simulation; e.g., one finds from the examination of Table 2, that all required expressions are symmetrical around the sites $I-1$, I and $I+1$ for model 1, and around the sites I and $I+1$ for model 2a, while this general symmetry is not kept for all average expressions of models 2b, 3a and 3b.

That is why - while the smoothing models 1 and 2a ensure always stable and convergent numerical simulations - for the smoothing models 2b 3a and 3b, respectively, there appear the basic types of usual numerical artifacts [7, 8, 14]: a) the instability, b) the pseudo-convergence. The plots of these numerical artifacts for the above-indicated 5 types of studied Finite Differences (FD) smoothing schemes (models), intended to the simulation of certain elastic pulses propagation through complex materials are presented by Figures 2 and 3 below (see also [15]).

i	$\tilde{\rho}_i$			$\tilde{S}_i$			$\left\langle \frac{\partial S}{\partial x} \right\rangle_i$		
	$I-1$	I	$I+1$	$I-1$	I	$I+1$	$\left\langle \frac{\partial S}{\partial x} \right\rangle_{I-1}$	$\left\langle \frac{\partial S}{\partial x} \right\rangle_I$	$\left\langle \frac{\partial S}{\partial x} \right\rangle_{I+1}$
1	ρ	$\frac{\rho'+\rho}{2}$	ρ'	S	$\frac{S+S'}{2}$	S'	0	$\frac{S'-S}{\epsilon}$	0
2a	ρ	$\frac{3\rho'+\rho}{4}$	$\frac{3\rho'+\rho}{4}$	S	$\frac{3S+S'}{4}$	$\frac{S+3S'}{4}$	0	$\frac{S'-S}{2\epsilon}$	$\frac{S'-S}{2\epsilon}$
2b	ρ	ρ	ρ'	S	S	S'	0	$\frac{S'-S}{2\epsilon}$	$\frac{S'-S}{2\epsilon}$
3a	ρ	$\frac{\rho'+\rho}{2}$	ρ'	S	$\frac{S+S'}{2}$	S'	$\frac{S'-S}{4\epsilon}$	$\frac{S'-S}{\epsilon}$	$\frac{S'-S}{4\epsilon}$
3b	ρ	$\frac{\rho'+\rho}{2}$	ρ'	$\frac{7S+S'}{8}$	$\frac{S+S'}{2}$	$\frac{S+7S'}{8}$	$\frac{S'-S}{4\epsilon}$	$\frac{S'-S}{\epsilon}$	$\frac{S'-S}{4\epsilon}$

Table 2: Expressions of the average values of the main elastic parameters for the basic smoothing models of one-dimensional (1-D) interfaces (see also [14]-a)

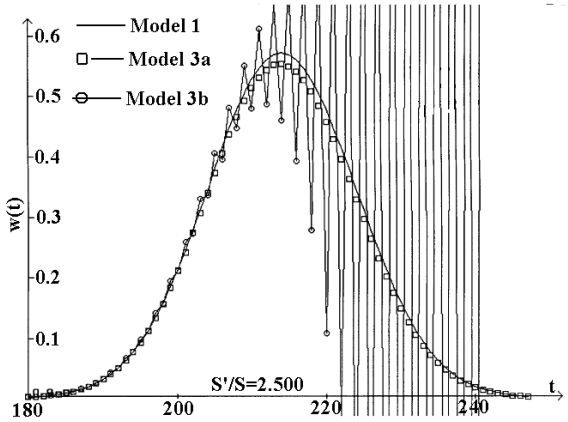


Figure 2: Plots of numerical simulations corresponding to different FD schemes (those of models 3a and 3b are pseudo-convergent, and unstable, respectively).

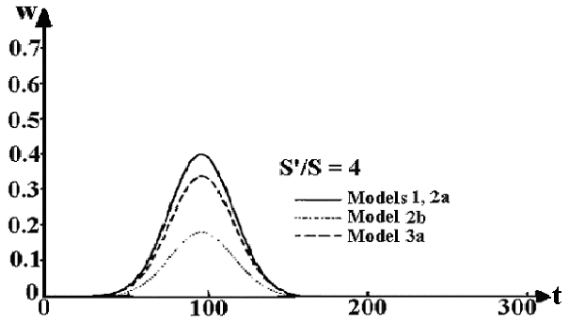


Figure 3: Convergent numerical simulations corresponding to models 1 and 2a, and pseudo-convergent ones for models 2b, 3a.

One finds that while the instabilities can be easily detected and eliminated, the pseudo-convergence is considerably more "dangerous", because: a) the pseudo-convergent simulations have a right shape, while: b) the corresponding wrong displacement values can be considerably more difficult observed, hence the pseudo-convergent simulations could be easily misleading.

3 Nonlinear properties of the propagation medium

It is well-known [7] that the computer rounding errors are amplified considerably in the frame of some non-linear equations, as those corresponding to certain solitary waves, leading due to some instability numerical artifacts (see Figure 4). Particularly, the Korteweg-de Vries equation:

$$\frac{\partial u}{\partial t} = -\nu_{00} u' - n u u' - d_1 u''' \tag{5}$$

can be discretized as:

$$f(i) = p(i) - \gamma \cdot [a(i+1) - a(i-1)] + \alpha \cdot a(i) \cdot [a(i-1) - a(i+1)] + \beta \cdot [a(i-2) - a(i+2)] \tag{6}$$

There were studied the numerical artifacts corresponding to the 2 main types of nonlinear solitary waves (which can propagate keeping their shapes): the bell-shaped (or breathers) and

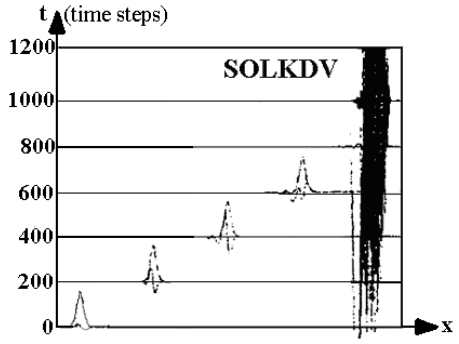


Figure 4: FD simulation of a Korteweg-de Vries (KdV) solitary wave breather propagation

the kink-shaped waves [16]. Figures 4 and 5 present the basic numerical artifacts intervening in the FD simulations of the breathers propagation (see also [17]).

While the artifacts intervening in the simulations of the KdV breathers propagation reduce to a monotonic increase of distortions up to the appearance of the instability (Figure 4), the use of the discrete version (DNLS) of the cubic nonlinear Schrödinger (NLS) equation to describe the wave-guide arrays with saturable nonlinearity leads in certain conditions to the artefact corresponding to the merging of two breathers with symmetry breaking (see Figure 5 and [17]).

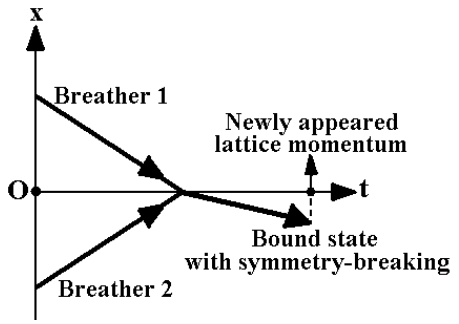


Figure 5: The symmetry breaking artefact intervening in the merging (bound state formation) of 2 symmetric SNLS breathers

4 Dissipative media

Because the modulus of the first solution of the attenuation-dispersion relation [15] is larger than 1: $|g_1| = e^{\epsilon E} > 1$, the FD schemes used for the simulation of the acoustic pulses propagation in dissipative media are always unstable. There were pointed out also: (i) the instability of the attenuated wave simulation, *even for absolutely exact initial conditions*, due to the generation of the amplified wave (mathematically possible, but without a physical meaning) by the stochastic local accumulation of some local "rounding" inaccuracies of the exact values corresponding to such waves, acting as a self-organising process in the computing program run, as well as: (ii) the extremely strong acceleration of the amplified wave generation when the complex wave-functions are used (stability and convergence radii of the magnitude order of 1 dB or even smaller). The introduction of some: (i) corrective measures (the use of some analytical expressions of some partial derivatives, particularly), (ii) properly chosen effective parameters, allows the weakening of these unpleasant numerical phenomena, ensuring stability and convergence radii of the magnitude order of 100 dB [13], which represent sufficiently high values for accurate descriptions, by

means of the finite difference method, of the cases of technical interest.

5 High Logical Depth

As it results from equation (1), the FD scheme of waves propagation in ideal media is not too intricate, even if the Courant's number is less than 1: $C < 1$. The repeated use [by a large number ($N > 10^5$) of successive iterations, e.g. for simulations of the ultrasonic non-destructive examinations of some industrial components] of this equation, leads however to the high logical depth Complexity feature of the used computer program, and consequently to several numerical artifacts indicated in the frame of Figure 6.

The expressions of the main limits are: $x_e = -Ct - 2\sqrt[3]{t}$, $x_p = 1 + C(t-1)$, $X_p = N + C(t-1)$, $x_n = Ct + N + 2\sqrt[3]{t}$, while for $C \approx 0.5$ and $N \approx 71$, the relative (to the incoming pulse one) amplitudes of the echo pulses have the magnitude orders: 0.1 for the rectangular pulse, 0.01 for the sine pulses, and 0.001 for the Gaussian pulses.

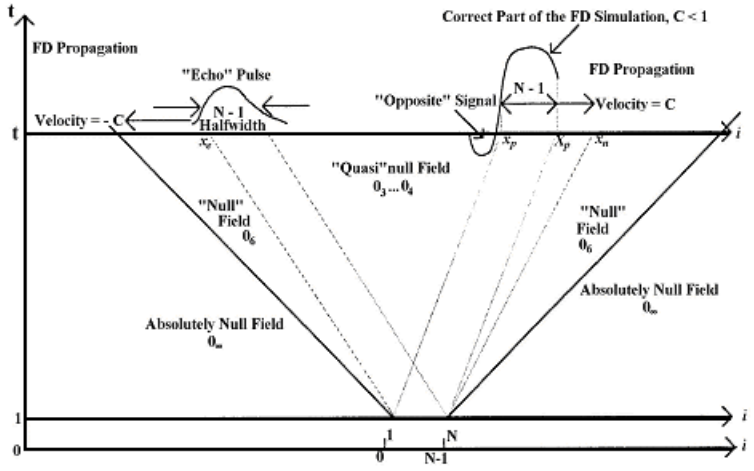


Figure 6: Structure of FD simulations of the propagation of pulses of different shapes

6 From the Numerical Artifacts to the Numerical Phenomena

6.1 Difficulties and main methods used to study the numerical artifact mechanisms

Because many problems in the field of Numbers Theory are extremely difficult, the aim of this study is to point out the main features of the mechanisms leading to some numerical artifacts intervening in the computer simulations of pulses propagation. The accomplished analysis pointed out that the main methods to study these numerical artifact mechanisms are the methods of the: a) FD transfer coefficients [18], b) Fourier's representation of the exact solutions of the discretized wave equation [19].

6.2 The method of transfer coefficients

In order to explain the results corresponding to the linear FD schemes, a partition of the incoming pulse in N components of amplitudes (in the order of their incoming on the studied

material) $s_1, s_2, \ldots, s_N$ is considered. The amplitudes of the same components in the previous time step are denoted by $s'_1, s'_2, \ldots, s'_N$.

The transfer coefficients k_{ti} are defined by means of the expressions (4) of the displacement w_{It} corresponding to the space site I at the moment (time step) t:

$$w_{It} = \sum_{j=1}^{N} k_{t,N+2+t-I-j} \cdot s_j - \sum_{j=1}^{N} k_{t,N+t-I-j} \cdot s'_j \tag{7}$$

A simplified definition of the transfer coefficients corresponds to the FD simulations with equal values of the real phase speed V_Φ and of the FD one: $V_{FD} = \frac{\epsilon}{\tau}$ (i.e. for the value 1 of the Courant number [7]: $C = \frac{V_\Phi}{V_{FD}}$), because then the pulse partition components at successive time steps coincide: $S'_i = s_i$ (for any $i = 1, 2, \ldots, N$). In the particular case of a sharp 1-D interface located in the site I, the transfer coefficients describing the transmitted wave are defined as [14]-a:

$$w_{I+1,t} = \sum_{j=1}^{t-2} k_j s_{t-j-1} \tag{8}$$

6.3 Fourier's representation method of the exact solutions of the discretized wave equation

The exact discrete solution of the wave equation is written by means of its Fourier expansion, as:

$$w_{j,t} = \sum_{k=-\infty}^{\infty} C_k \cdot [g(k)]^t \cdot e^{ik \cdot j\epsilon} \tag{9}$$

where $g(k)$ is named the amplification factor. Introducing this expression in the wave equation, one obtains the "attenuation-dispersion" relation:

$$g - 2 + \frac{1}{g} = F(k \cdot \epsilon, V_{FD}, w) \tag{10}$$

where $F(k \cdot \epsilon, V_{FD}, w)$ is a specific function of the considered wave.

According to von Neumann's theorem ([19]-a, p. 42), the considered FD scheme will be stable if both solutions of the algebraic equation (10) fulfil the requirement: $|g_{1,2}| \leq 1$, else this scheme will be unstable.

6.4 Applications to the study of mechanisms of some numerical artifacts

The method of transfer coefficients. Appplying this method to the problem of sharp interfaces (see Section 2a and Figures 1 - 3), we will find that the above indicated numerical artifacts belong to the class of numerical phenomena, because they can be identified and described starting from the values of the roots of the characteristic equation [15]:

$$\xi^2 - (t\_t_1 + 2t\_t_2 + t_0 t_1)\xi - t\_t_2 = 0. \tag{11}$$

where: $t_i = a_{i-1} b_i - 1$, while a_i and b_i are the coefficients of the FD wave equation (4): $w_{i,t+1} = a_i w_{i+1,t} + b_i w_{i-1,t} - w_{i,t-1}$. One finds so that as $|\xi| > 1$ or $|\xi| < 1$, the used FD scheme is unstable, or it is stable.

The method of Fourier's representation. Using the above presented method of the Fourier's representation of the exact solutions to the problem of Korteweg-de Vries solitons

(Section 3), one obtains the following expression [19]-b,c of the upper threshold (for the numerical scheme stability) of the time step:

$$\tau_{max} = \frac{\epsilon}{\alpha A + \frac{4\beta}{\epsilon^2}} \tag{12}$$

Our study [15] pointed out both the validity of the Vliegenthart condition (13), as well as the monotonic improvement of the FD simulations accuracy as the representative point of the FD steps ϵ, τ tends to the Vliegenthart's boarder of the stability and instability regions.

One finds so that the stability field of the FD simulations of KdV solitons propagation is rather broad. Because the size and borders of the stability domain depend on the: a) strongly (e.g., in the case of some exponential dependencies) or weakly (as in the above studied case) character of the nonlinear dependence, b) number of interacting system components.

7 Stability and Convergence Radii of Different Numerical Schemes

The accomplished numerical studies [20], [21] have pointed out that, for given values of the wave frequency (or wavelength) and of the tangent of mechanical losses, beginning from a certain number of space (or time) steps x_{lim}, one finds usually the appearance of large oscillations of the simulated displacements, which lead quickly to instability. Because the instability is determined by the value of the factor e^{Ex} and: $E = k \tan \frac{\delta}{2}$, while the wave intensity is proportional to the square of displacement: $I \propto w^2$, one finds that the measure (in deci-Bells) of the intensity level corresponding to the stability field is:

$$\langle L_{I,stab} \rangle_{dB} = 2 \langle L_{w,stab} \rangle_{dB} = 20 E x_{lim} = 20 k x_{lim} \tan \frac{\delta}{2} = 40\pi \frac{x_{lim}}{\lambda} \tan \frac{\delta}{2}. \tag{13}$$

Of course, the decrease of the wave intensity corresponding to the stability field (limit) is:

$$\frac{I_{lim}}{I_0} = e^{-2Ex_{lim}} = e^{\frac{\langle L_{I,stab} \rangle}{10}} \tag{14}$$

Table 3 synthesizes the obtained numerical results.

Type of the wave equation	No. of numerical interactions	$\langle L_{I,stab} \rangle_{dB}$	t_{lim} (stability, life, steps)
Complex stiffness equation	12	0.0321	102
Complex stress relaxation time	10	4	12732
Complex wave-vector equation	8	20	63622
Space evolution Equation	6	40.476	128839
Real wave, equation	5	80.130	255062

Table 3: Stability radii and mean life of different numerical simulations [15].

8 Analysis of the Obtained Results for Different Studied Numerical Schemes and Physical Processes

The obtained results (Tables 1 and 2) concerning the stability and convergence radii of different numerical schemes intended to the computer simulation of certain physical processes

(acoustic pulse propagation, diffusion with drift, absorption, etc) indicate the "accessible" logical depths [22] of the specific studied physical problems, for each of the used numerical schemes. These results present also a considerable importance for the choice and optimization of the numerical schemes [23]. Certain numerical schemes, e.g. that corresponding to the complex stiffness $\bar{S}$ symmetric wave equation of the propagation in dissipative media:

$$\rho \frac{\partial^2 \bar{w}}{\partial t'^2} = \bar{S} \frac{\partial^2 \bar{w}}{\partial x^2} \tag{15}$$

allow multiple solutions; using the FD descriptions: $t' = t\tau$ and: $x = I\epsilon$ (in terms of the time τ and space ϵ steps) of the real time t and space coordinate x, these solutions can be written as:

$$\bar{w}_{I,t} = A \cdot e^{\pm i\omega t\tau} \cdot e^{\pm(E+ik)I\epsilon} \tag{16}$$

Even if the initial conditions launch only the "direct" wave:

$$\bar{w}^{dir.}_{I,t} = A \cdot e^{-EI\epsilon} \cdot e^{i(\omega t\tau + kI\epsilon)} \tag{17}$$

some random accumulations of the rounding errors intervening in the evaluation of the partial derivatives produce a local ("spontaneous") generation of the inverse wave:

$$\bar{w}^{inv.}_{I,t} = A' \cdot e^{EI\epsilon} \cdot e^{i(\omega t\tau + kI\epsilon)} \tag{18}$$

leading to the sudden apparition of instabilities.

One finds so that the numerical simulations of the waves propagation through dissipative media lead to a typical problem of self-organizing systems, with a spontaneous symmetry breaking. This symmetry breaking corresponds to the "spontaneous" local generation of the inverse wave, launched by the random accumulation of the "garbage" rounding errors and followed by the transition between the attenuated wave and the apparently amplified wave, corresponding to the "inverse" wave. The accomplished study (see Tables 1 and 2) points out that the "speed" of this self-organization process crucially depends on the number and intensity of the numerical "interactions" between the components (the values $w_{I,t}$ of the displacement in different sites I, t of the FD grid) of the simulation process.

Because such numerical "interactions" are achieved mainly by the FD approximate expressions of the partial derivatives, the "spontaneous" breaking of the symmetry appears quicker for (in the decreasing order of importance):

- a) large numbers of displacement components involved in the expressions of partial derivatives, e.g. when their expressions with 2 previous time steps (instead of those using an only one previous time step) are used[1]:
  $\dot{f}(0) = -f(2\tau) + 8f(\tau) - 8f(-\tau) + f(-2\tau)/12\tau$,
  $\ddot{f}(0) = -f(2\tau) + 16f(\tau) - 30f(\tau) + 16f(-\tau) - (-2\tau)/12\tau^2$ when the instabilities appear after only few tens of iterations,
- presence and repeated "mixture" of the values of both real and pure imaginary parts of the complex wave function (displacement) $\bar{w}$,
- more parasitic solutions,
- more partial derivatives involved in the expression of the differential equation of the acoustic pulse propagation.

[1]The formulae in more points are considerably more accurate for rather small numbers of iterations, but they give rise later to spurious solutions and instability (see Table 1).

For these reasons, the highest "accessible" logical depth [22] is reached (for the simulations of the acoustic pulse propagation through attenuative media) for the numerical scheme using the real wave function equation (see table reftab02), with the usual FD approximations of the first 2 order derivatives:

$$\dot{f}(0) = \frac{f(\tau) - f(-\tau)}{2\tau}, \ddot{f}(0) = \frac{f(\tau) - 2f(0) + f(-\tau)}{\tau^2}. \tag{19}$$

9 Conclusions and Future Works

The obtained results concerning the different numerical phenomena associated to the complex computer simulations present a considerable importance for the choice and optimization of these numerical schemes [23]. It was also found that some numerical simulations (e.g, those of the acoustic pulse propagation through attenuative media) allow the study of some features of the self-organizing systems (the "spontaneous" symmetry breaking, the influence of the interactions between the system components on the "accessible" logical depth, etc).

Acknowledgments

The authors acknowledge the financial support from the National Center for Programs Management (CNMP) of the Romanian Ministry of Education, Research, Youth and Sports, under the Contract No. D11-044/2007-QUANTGRID.

Bibliography

[1] R. Skeel, SIAM News, 25(4), p. 11, 1992.

[2] SIAM News, 29(8), pp. 1, 123, 13, 1996, http://www.siam.org/siamnews/general/ariane.htm

[3] a) D. W. McClure "Computer Errors", in D. A. Iordache, D. W. McClure, Selected Works of Computer Aided Applied Sciences, vol. 2, Printech Publishing House, Bucharest, 2002, p. 535; b) D. W. McClure "Computer errors", Basic notions (chapter 9), Applications (chapter 10), in the frame of textbook E. Bodegom, D. W. McClure et al (D. Iordache, Fl. Pop, C. Roşu - editors) "Computational Physics Guide", Politehnica Presss, Bucharest, 2009.

[4] Cl. S. Fermat "Diophantus' Arithmetica containing (48) observations by P. de Fermat", Toulouse, 1670.

[5] A. Wiles "Modular elliptic curves and Fermat's last theorem", Annals of Mathematics, 142, 443-551(1995).

[6] S. Singh "Fermat's Enigma: the Epic Quest to Solve the World's Greatest Mathematical Problem", Walker Publishing Company, New York, 1997.

[7] R. Courant, K. Friedrichs, H. Lewy, Math. Ann., 100, 32(1928).

[8] P.P. Delsanto, T. Whitcombe, H.H. Chaskelis, R.B. Mignogna, Wave Motion, 16, 65(1992).

[9] D. Iordache, P. Delsanto, M. Scalerandi "Pulse Distortions in the FD Simulations of Elastic Wave Propagation", Mathl. Comp. Modelling, 25(6) 31-43, 1997.

[10] D. Iordache, M. Scalerandi, C. Ruginǎ, V. Iordache "Study of the Stability and Convergence of FD Simulations of Ultrasound Propagation through Non-homogeneous Classical (Zener's) Attenuative Media", Romanian Reports on Physics, 50(10) 703-716, 1998; b) D. A. Iordache, M. Scalerandi, V. Iordache, Romanian Journal of Physics, 45(9-10) 685(2000).

[11] J. C. Strikwerda "Finite Difference Schemes and Partial Difference Equations", Wadsworth-Brooks, 1989.

[12] P. P. Delsanto, G. Kaniadakis, M. Scalerandi, D. Iordache, Comp. Math. Applic., 27(6) 51-61(1994).

[13] P.P. Delsanto, G. Kaniadakis, M. Scalerandi, D. Iordache, Mathl. Comp. Modelling (UK), 19(9) 1-8 (1994).

[14] a) P. P. Delsanto, D. Iordache, C. Iordache, E. Ruffino "Analysis of Stability and Convergence in FD Simulations of the 1-D Ultrasonic Wave Propagation", Mathl. Comp. Modelling, 25(6) 19-29, 1997; b) D. Iordache, Şt. Puşcă, C. Toma "Numerical Analysis of some Typical FD Simulations of the Waves Propagation through Different Media", Lecture Notes on Computer Sciences, 3482, 614-620, 2005.

[15] D. Iordache "Contributions to the Study of Numerical Phenomena intervening in the Computer Simulations of some Physical Processes", Credis Printing House, Bucharest, 2004.

[16] A. V. Porubov, M. G. Velarde "Strain kinks in an elastic rod embedded in a viscoelastic medium", Wave Motion, 35, 189-204, 2002.

[17] J. Cuevas, J. C. Eilbeck "Discrete soliton collisions in a waveguide array with saturable nonlinearity", Physics Letters A, 358(1) 15-20, 2006.

[18] D. Iordache, M. Scalerandi, C. Iordache "Mechanisms of Some Numerical Phenomena Specific to the Finite Differences Simulations of the Ultrasound Propagation", Proc. 25th Congress of the American-Romanian Science Academy, Cleveland (US), 2000, pp. 263-266.

[19] a) J. C. Strikwerda "Finite Differences Schemes and Partial Difference Equations", Wadsworth-Brooks, 1989; b) A. C. Vliegenthart, J. Eng. Math., 3, 81-94, 1969; c) A. C. Vliegenthart, J. Eng. Math., 5, 137-155, 1971.

[20] a) P. P. Delsanto, M. Scalerandi, V. Agostini, D. Iordache, Il Nuovo Cimento, B, 114, 1413-26(1999); b) D. Iordache, M. Scalerandi M., C. Ruginǎ, V. Iordache, Romanian Reports on Physics, 50(10) 703-716 (1998).

[21] D. Iordache, M. Scalerandi, V. Iordache, Romanian J. of Physics, 45(9-10) 685-704 (2000).

[22] M. Gell-Mann, Europhysics News, 33(1) 17-20 (2002).

[23] D. Iordache, V. Iordache, Romanian Journal of Physics, 48(5-6) 697-704 (2003).

Int. J. of Computers, Communications & Control, ISSN 1841-9836, E-ISSN 1841-9844
Vol. V (2010), No. 5, pp. 755-767

Boundary Control by Boundary Observer for Hyper-redundant Robots

M. Ivanescu, D. Cojocaru, N. Bizdoaca, M. Florescu
N. Popescu, D. Popescu, S. Dumitru

Mircea Ivanescu, Dorian Cojocaru,
Nicu Bizdoaca, Mihaela Florescu, Sorin Dumitru
Mechatronic Department, University of Craiova
E-mail: ivanescu, cojocaru, nicu, mihaela, dumitru@robotics.ucv.ro

Nirvana Popescu, Decebal Popescu
Department of Computer Science,
Politehnica University Bucharest,
E-mail: nirvana.popescu, decebal.popescu@cs.pub.ro

Abstract: The control problem of a class of hyper-redundant arms with continuum elements, with boundary measuring and control is discussed. First, the dynamic model of the continuum arm is presented. The measuring systems are based on the film sensors that are placed at the terminal sub-regions of the arm. The observers are proposed in order to reconstruct the full state of the arm. A back-stepping method is used to design a boundary control algorithm. Numerical simulations of the arm motion toward an imposed position are presented. An experimental platform shows the effectiveness of the proposed methods.
Keywords: hyper-redundant system, distributed parameter system, observer, control.

1 Introduction

The hyper-redundant arms are a class of manipulators that can achieve any position and orientation in space. A special class of these robots is represented by the mechanical structures with continuum elements described by distributed parameter model. The control of these systems is very complex and a great number of researchers have tried to offer solutions. In [2, 3], Gravagne analyzed the kinematic models. Important results were obtained by Chirikjian and Burdick [4], which laid the foundations for the kinematical theory of hyper-redundant robots. Their results are based on a "backbone curve" that captures the robot's macroscopic geometric features. Mochiyama has also investigated the problem of controlling the shape of an HDOF rigid - link robot with two-degree-of-freedom joints using spatial curves [5]. In other papers [6, 7], several technological solutions for actuators used in hyper-redundant structures are presented and conventional control systems are introduced. In [8] control problem of a class that performs the grasping function by coiling is discussed. A frequential stability criterion for the grasping control problem is proposed in [9].

In this paper, control problem of a class of hyper-redundant arms with continuum elements, with boundary measuring and control is discussed. The development of feedback controllers and compensators for these models is a very complex problem. The difficulty is determined by the complexity of the dynamic models expressed by partial differential equations and by the observability problems in distributed parameter systems. An essential part of designing feedback controllers for these models is designing practical controllers that are implementable. Standard feedback control design assumes full-state feedback with measurements of the entire state. Recent

advances in distributed sensor technology, as Polyvinylidene Fluoride film sensors [10], allow ensuring a good quality of position measuring in distributed systems. However, the use of these sensors on all surface of continuum arms, is not practical due to mechanical constraints. In fact, the sensors are placed on the boundary of the elements. In this case, the development of the state-feedback controllers needs to design state observers. The observability problems are solved by an approach derived from the Luenberger observer type and the "back-stepping method" developed in [11].

The paper is organized as follows: section II presents technological and theoretical preliminaries, section III studies the dynamic model, section IV presents the control by boundary observer, section V verifies the control laws by computer simulation and section VI presents an experimental model.

2 Technological and Theoretical Preliminaries

The hyper-redundant technological models are complex structures that operate in 3D space, but the control laws of the elements can be infer from the planar models. For this reason, the model discussed in this paper is a 2D model.

The technological model basis is presented in Fig.1. It consists of a number (N) of continuum segments, each segment having a layer structure that ensures the flexibility, the driving and position measuring (Fig.2).

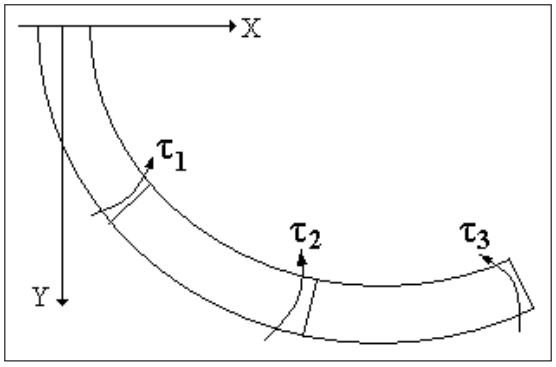


Figure 1:

The high flexibility is obtained by an elastic non-extensible backbone rod with distributed damping and negligible shear effects.

The driving system consists by two antagonistic cable actuators that are connected at the end of each segment and determine the bending of the arm.

The position measuring of the segment is obtained by an electro-active polymer curvature sensor that is placed on the surface at the terminal sub-regions of each segment. These sensors can measure the curvature on the boundary of the segment (s=0 or s=l). The essence of the segment i is the backbone curve C_i.The length of each segment is l. The independent parameter s is related to the arc-length from origin of the curve C_i, $s \in \Omega, \Omega = [0, l]$. The curvature of the segment is [13] (Fig.3)

$$\chi = \frac{d\phi}{ds} \tag{2.1}$$

where

$$\phi = \frac{s}{R_c} \tag{2.2}$$

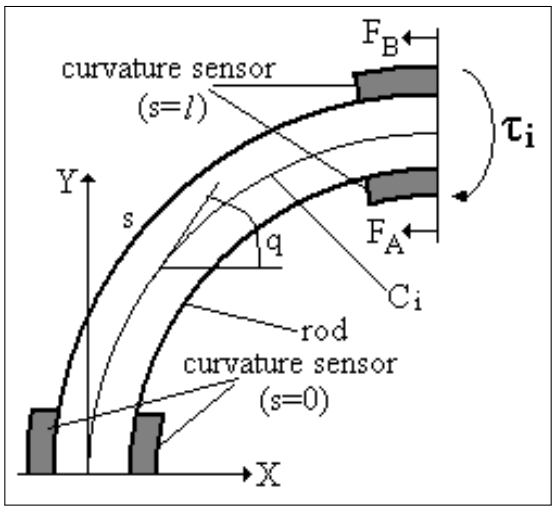


Figure 2:

represents the angle of the current position and R_C is the radix of the arc. We denote by τ the equivalent moment at the end of the segment (s=l) exercised by the cable forces F_A and F_B. The position of a point s on curve C_i is defined by the position vector r=r(s), $s \in [0, l]$. For a dynamic motion, the time variable will be introduced, r=r(s,t) . The segment has the elastic modules E, the moment of inertia I, the bending stiffness EI, the linear mass density ρ and rotational inertial density I_ρ.

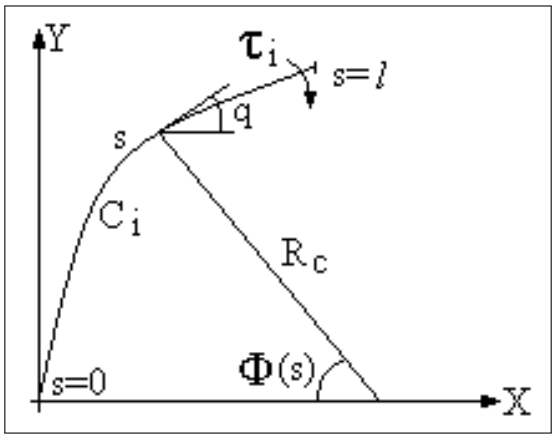


Figure 3:

3 Dynamic model

The dynamic model of a segment can be derived from the Hamiltonian principle. Using the same procedure as in[3] yields the partial differential equations of the arm segment,

$$I_\rho \ddot{q} + b_1 \dot{q} - EIq_{ss} + c_1 q = 0 \tag{3.1}$$

with the initial and boundary conditions

$$\dot{q}(0, s) = 0 \tag{3.2}$$

$$EIq_s(t,l) = \tau \tag{3.3}$$

$$q_s(t,0) = 0 \tag{3.4}$$

where q=q(t,s), $\dot{q}$, q_s, q_{ss} denote $\frac{\partial q(t,s)}{\partial t}$, $\frac{\partial q(t,s)}{\partial s}$, $\frac{\partial^2 q(t,s)}{\partial s^2}$, respectively, b_1 is the equivalent damping coefficient and c_1 characterizes the elastic behavior.

The equations (3.1) - (3.4) can be rewritten as:

$$\ddot{q} = aq_{ss} + b\dot{q} + cq \tag{3.5}$$

$$q_s(t,0) = 0 \tag{3.6}$$

$$q_s(t,l) = d \cdot \tau \tag{3.7}$$

$$\dot{q}(0,s) = 0, s \in [0,l] \tag{3.8}$$

where

$$a = \frac{EI}{I_\rho}; b = -\frac{b_1}{I_\rho}; c = -\frac{c_i}{I_\rho}; d = \frac{1}{EI} \tag{3.9}$$

The input of the system is represented by the moment τ applied at the boundary s=l of the arm. The output is determined by the angle values measured by the sensor,

$$y(t) = q(0,t) \tag{3.10}$$

or

$$y(t) = q(l,t) \tag{3.11}$$

4 Control by boundary observer

We shall analyze two cases: 1) the measurement system allows to measure the angle at the bottom end s=o; 2) the measurement system allows to measure the angle at the upper end s=l. For the both cases, a regional boundary observer is introduced in order to reconstruct the state in the domain and generate a full-state feedback.

4.1 Problem 1: q(t,0) is available for measurement (Fig.4)

The following observer is proposed:

$$\ddot{\hat{q}} = a\hat{q}_{ss} + b\dot{\hat{q}} + c\hat{q} + k_1(s)(q(t,0) - \hat{q}(t,0)) \tag{4.1}$$

$$\hat{q}_s(t,0) = k_0(q(t,0) - \hat{q}(t,0)) \tag{4.2}$$

$$\hat{q}_s(t,l) = d \cdot \tau \tag{4.3}$$

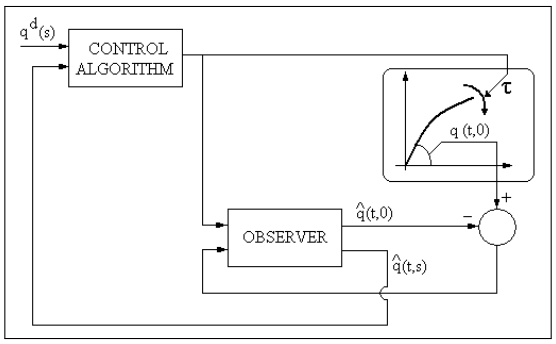


Figure 4:

$$\dot{\hat{q}}(0,l) = 0 \tag{4.4}$$

where $\hat{q}$=$\hat{q}$(t,s) is the observer state and $k_1(s), k_o$ are a function and a constant, respectively, that define the observer parameters. The objective is to determine these parameters in order to reconstruct the state in the domain, i.e., to find $k_1(s)$ and k_o such that $\hat{q}$ converges to q as time goes to infinity.

An error variable $\tilde{q}$ is introduced

$$\tilde{q} = q - \hat{q} \tag{4.5}$$

and the error system will be:

$$\ddot{\tilde{q}} = a\tilde{q}_{ss} + b\dot{\tilde{q}} + c\tilde{q} - k_1(s)\tilde{q}(t,0) \tag{4.6}$$

$$\tilde{q}_s(t,0) = -k_0\tilde{q}(t,0) \tag{4.7}$$

$$\tilde{q}_s(t,l) = 0 \tag{4.8}$$

$$\dot{\tilde{q}}(0,l) = 0 \tag{4.9}$$

where

$$\lim_{t\to\infty} \tilde{q}(t,s) = 0, s \in [0,l] \tag{4.10}$$

We consider that the desired states of the arm motion are given by the curve C_d,

$$C_d : (q^d(s).s \in [0,l]) \tag{4.11}$$

The control problem is to find the moment control law τ in order to achieve the desired state.

Control algorithm 1. The closed loop control law of the arm (3.5) - (3.8) with the boundary observer (4.1) - (4.4) is given by

$$\tau(t) = EI\left(-k(l,l)(\hat{q}(t,l) - q_d(l)) + q_s^d(l) - \int_0^l k_s(l,z)(\hat{q}(t,z) - q^d(z))dz\right) \tag{4.12}$$

where k(s,z) is the solution of the following partial differential equations,

$$-ak_{ss}(s,z) + ak_{zz}(s,z) + ck(s,z) = 0 \tag{4.13}$$

$$k_{ss}(s,s) = \frac{c}{2a}(l-s) \tag{4.14}$$

with the boundary condition

$$k(l,z) = 0, z \in [0,l] \tag{4.15}$$

and the observer parameters are defined by the equations

$$k_0 = -k(0,0) \tag{4.16}$$

$$k_1(s) = -ak_z(s,0) - \int_0^s k_1(z)k(s,z)dz \tag{4.17}$$

Proof. See Appendix 1.

4.2 Problem 2: q(t,l) is available for measurement (Fig.5)

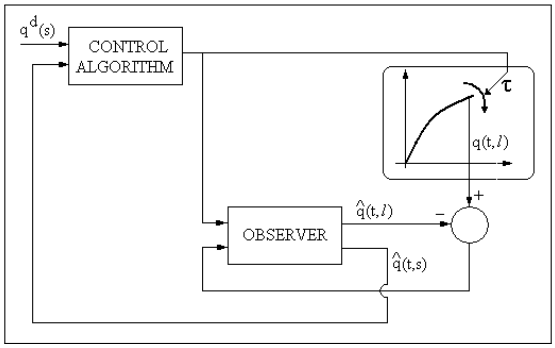


Figure 5:

The observer will be

$$\ddot{\hat{q}} = a\hat{q}_{ss} + b\dot{\hat{q}} + c\hat{q} + k_1(s)(q(t,l) - \hat{q}(t,l)) \tag{4.18}$$

with the boundary conditions

$$\hat{q}_s(t,0) = \hat{q}_0 \tag{4.19}$$

$$\hat{q}_s(t,l) = k_0(q(t,l) - \hat{q}(t,l)) + d\tau \tag{4.20}$$

$$\dot{\hat{q}}(0,0) = 0 \tag{4.21}$$

The error system has the form

$$\ddot{\tilde{q}} = a\tilde{q}_{ss} + b\dot{\tilde{q}} + c\tilde{q} - k_1(s)\tilde{q}(t,l) \tag{4.22}$$

$$\tilde{q}_s(t,0) = -\hat{q}_0 \tag{4.23}$$

$$\tilde{q}_s(t,l) = -k_0\tilde{q}(t,l) \tag{4.24}$$

$$\dot{\tilde{q}}(0,0) = 0 \tag{4.25}$$

Control algorithm 2. The closed loop control law of the arm (3.5) - (3.8) with the boundary observer (4.17) - (4.20) is given by

$$\tau(t) = EI(q_s^d(l) - k(l,l)(q^d(l) - \hat{q}(t,l))) \tag{4.26}$$

where k(s,z) is the solution of the following equations

$$ak_{ss}(s,z) - ak_{zz}(s,z) + ck(s,z) = 0 \tag{4.27}$$

$$k(s,s) = \frac{c}{2a}s \tag{4.28}$$

with the boundary conditions

$$k(0,z) = 0, z \in [0,l] \tag{4.29}$$

The observer parameters are obtained by solving the following equations

$$k_0 = k(l,l) \tag{4.30}$$

$$k_1(s) = ak_s(s,l) + \int_s^l k_1(z)k(s,z)dz \tag{4.31}$$

Proof. See Appendix 2.

5 Simulation

A hyper-redundant manipulator control with continuum segments is simulated. The parameters of the arm were selected as: bending stiffness EI=1, rotational inertial density $I_\rho = 0.001kg \cdot m^2$, damping ratio 0.35 and elastic coefficient 4.8. These constants are realistic for long thin backbone structures. The length of each segment is l=1.

Problem 1. **q(t, 0) is available for measurement**

The observer parameters $k_0, k_1(s)$ are computed. First, a numerical solution for k(s,z) is obtained by the integration of partial differential equation (4.13) with boundary conditions (4.14) and (4.15). The result is presented in Fig.6.

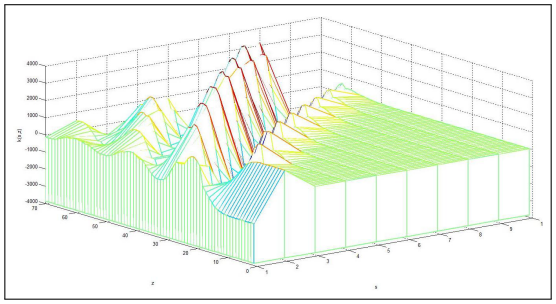


Figure 6:

The parameter k_0 is obtained from (4.16) as k_0=-k(0,0)=12.4

The parameter k_1 is determined from the integral equation (4.17). The numerical solution is presented in the Table 1.

Table 1. Parameter $k_1(s)$ for the Problem 1

S	0	0.2	0.4	0.6	0.8	1.0
$k_1(s)$	-3.30	-3.80	-0.52	0.14	0.06	0.00

A desired trajectory defined as

$$q^d(s) = \pi s^2, s \in [0, 1] \tag{5.1}$$

is proposed and the control law (4.12) is used A MATLAB simulation of the observer arm system is implemented. The result is presented in Fig.7.

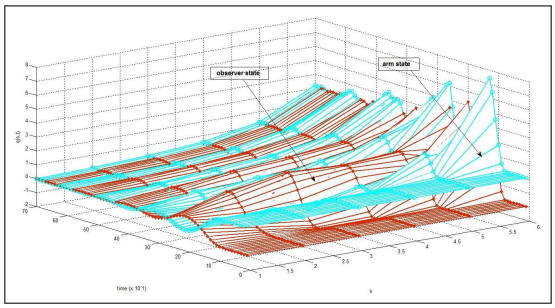


Figure 7:

We can remark the convergence of the estimated state of the observer to the system state and the quality of evolution on the trajectory to the desired state.

Problem 2. **q(t, l) is available for measurement**

The numerical solution of k(s,z) is obtained from the equation (4.28) with the boundary conditions (4.28), (4.29). The result is presented in Fig.8.

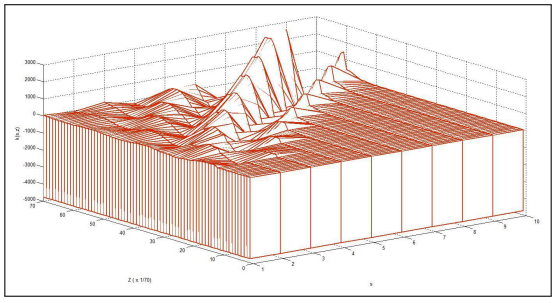


Figure 8:

The observer parameters are obtained from (4.30), (4.31), k_0=k(1,1)=12.4 and $k_1(s)$ is represented in Table 2.

Table 2. $k_1(s)$ for the Problem 2

S	0	0.2	0.4	0.6	1.0
$k_1(s)$	4.83	8.64	2.15	-0.01	0.00

A desired trajectory

$$q^d(s) = 1.4s^2, s \in [0, 1] \tag{5.2}$$

is proposed.

The control law (4.26) is applied and a simulation in MATLAB of the global system, as presented in Fig.5, is implemented. The result is shown in Fig.9.

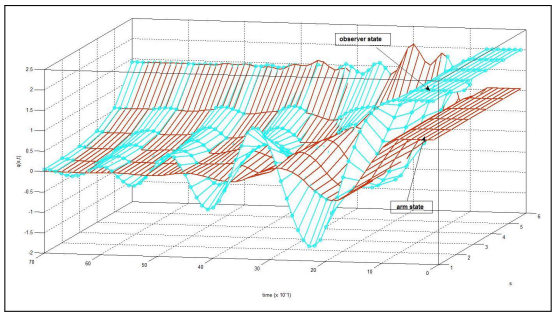


Figure 9:

6 Experimental results

In order to verify the suitability of the control algorithm, a platform with a 3D hyper-redundant arm has been employed for testing. The arm consists of three continuum segments with a flexible backbone rod. Three antagonistic cable actuators for each segment ensure the actuation system (Fig.10, Fig.11). The force in each cable is determined by the DC motors and a transmission system.

e

Figure 10:

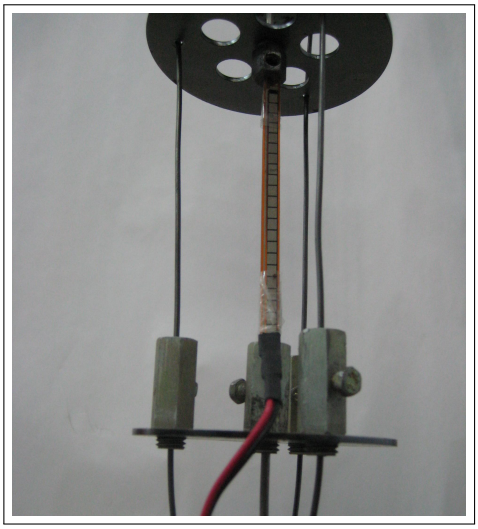


Figure 11:

A polymer thick film layer is placed on the upper level of the rod on each segment (s=l=0.3m). A sensor exhibits a decrease in resistance when an increase of the film curvature is used. A Wheatstone bridge system is used to measure the variation of the resistance. A Quancer based platform is used for control and signal acquisition. A control law (4.27) with $q_d(s)=40s^2$ is implemented. The result is presented in Fig.12.

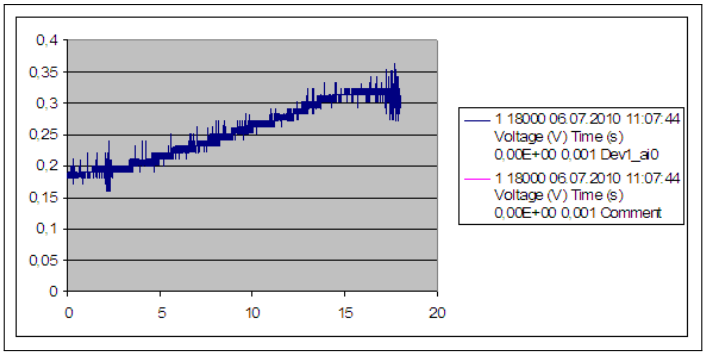


Figure 12:

7 Conclusions

The paper treats the control problem of a class of hyper-redundant arms with continuum elements. The observability problems for these models described by partial differential equations are analyzed. The measuring systems are based on the sensors placed on the boundary of the arm. Several observers are proposed for reconstructing the full state of the arm. A back-stepping technique is used in order to design a boundary control algorithm. The numerical simulations and an experimental platform illustrate the effectiveness of the method.

Acknowledgement
The research presented in this paper was supported by the Romanian National University Research Council CNCSIS through the IDEI Research Grant ID93 and by CNCSIS - UEFISCSU, project number PNII - IDEI code 289/2008.

Appendix 1

The control algorithm is derived by using the back-stepping method developed in [11]. The coordinate transformation

$$w(t,s) = \tilde{q}(t,s) - \int_0^s k(s,z)\tilde{q}(t,z)dz \qquad (A.1.1.)$$

transforms the error system (4.6) - (4.9) into a stable (b<0) target system

$$\ddot{w} = aw_{ss} + b\dot{w} \qquad (A.1.2.)$$

$$w_s(t,0) = 0 \qquad (A.1.3.)$$

$$w_s(t,l) = 0 \qquad (A.1.4.)$$

$$\dot{w}(0,s) = 0, s \in [0,l] \qquad (A.1.5.)$$

where

$$\lim_{t\to\infty} w(t,s) = 0, s \in [0,l] \qquad (A.1.6.)$$

From (A.1.1) we obtain

$$\dot{w} = \dot{\tilde{q}} - \int_0^s k(s,z)\dot{\tilde{q}}(t,z)dz \qquad (A.1.7.)$$

$$\ddot{w} = \ddot{\tilde{q}} - \int_0^s k(s,z)(a\tilde{q}_{ss} + b\dot{\tilde{q}} + c\tilde{q} - k_1(z)\tilde{q}(t,0))dz \qquad (A.1.8.)$$

$$w_{ss} = \tilde{q}_{ss} - \int_0^s k_{ss}(s,z)\tilde{q}(t,z)dz - k_s(s,s)\tilde{q}(t,s) - \frac{dk(s,s)}{ds}\cdot\tilde{q}(t,s) - k(s,s)\tilde{q}_s(t,s) \qquad (A.1.9.)$$

If we substitute (A.1.7), (A.1.8) in (A.1.2) and integrate by parts we get

$$\begin{aligned}&\int_0^s(-ak_{zz}(s,z) + ak_{ss}(s,z) + ck(s,z))\tilde{q}(t,z)dz + \\ &+\left(c + ak_z(s,s) + ak_s(s,s) + a\frac{dk(s,s)}{ds}\right)\tilde{q}(t,s) + \\ &+\left(-ak_z(s,0) - k_1(s) - \int_0^s k_1(z)k(s,z)dz\right) = 0\end{aligned} \qquad (A.1.10)$$

For left hand side to be zero, the condition (4.13), (4.14) can be easily inferred.
From (A.1.7), the velocities at t=0, s=l will be

$$\dot{w} = \dot{\tilde{q}}(0,l) + \int_0^l k(l,z)\dot{\tilde{q}}(0,z)dz \qquad (A.1.11.)$$

and by the boundary and initial conditions (3.8), (4.9), (A.1.5), we obtain

$$k(l,z) = 0, z \in [0,l] \qquad (A.1.12.)$$

From (A.1.1) we get

$$w_s(t,s) = \tilde{q}_s(t,s) - \int_0^s k_s(s,z)\tilde{q}(t,z)dz - k(s,s)\tilde{q}(t,s) \qquad (A.1.13.)$$

and by using the boundary condition (4.7) this relation becomes

$$-k_0\tilde{q}(t,0) - k(0,0)\tilde{q}(t,0) = 0$$

or

$$k_0 = -k(0,0) \qquad (A.1.14.)$$

Also, if we consider that the desired position is defined by $q^d(s)$, $s \in [0,l]$ and use the boundary conditions (3.7), (A.1.4) in the relation (A.1.13), the control law (4.12) is easily obtained.

Appendix 2

The back-stepping transformation is chosen as [11, 12]

$$w(t,s) = \tilde{q}(t,s) - \int_s^l k(s,z)\tilde{q}(t,z)dz \qquad (A.2.1.)$$

with the target system defined by (A.1.2) - (A.1.5). By using the same procedure as in Appendix 1, we obtain

$$\begin{aligned} &\left(c - ak_z(s,s) - ak_s(s,s) - a\frac{dk(s,s)}{ds}\right)\tilde{q}(ts) + \\ &+\left(-k_1(s) + \int_s^l k_1(z)k(s,z)dz + ak_z(s,l)\right)\tilde{q}(t,l) + \\ &+\int_s^l (ck(s,z) - ak_{zz}(s,z) + ak_{ss}(s,z))\tilde{q}(t,z)dz = 0 \end{aligned} \qquad (A.2.2)$$

or

$$ak_{ss}(s,z) - ak_{zz}(s,z) + ck(s,z) = 0 \qquad (A.2.3.)$$

$$k(s,s) = \frac{c}{2a}s, (for k(0,0) = 0) \qquad (A.2.4.)$$

$$k_1(s) = ak_s(s,l) + \int_s^l k_1(z)k(s,z)dz \qquad (A.2.5.)$$

From (A.2.1) the velocities at t=0, s=0 will be

$$\dot{w}(0,0) = \dot{\tilde{q}}(0,0) + \int_0^l k(0,z)\dot{\tilde{q}}(0,z)dz \qquad (A.2.6.)$$

and the boundary conditions (3.8), (4.26), (A.1.5) require

$$k(0,z) = 0 \qquad (A.2.7.)$$

From the back-stepping transformation (A.2.1) we obtain

$$w_s(t,s) = \tilde{q}_s(t,s) - \int_s^l k_s(s,z)\tilde{q}(t,z)dz + k(s,s)\tilde{q}(t,s) \qquad (A.2.8.)$$

By using the boundary conditions for s=l, (4.25), (A.1.4), it results

$$-k_0\tilde{q}(t,l) + k(l,l)\tilde{q}(t,l) = 0$$

or

$$k_0 = k(l,l) \qquad (A.2.9.)$$

From (A.2.8), for s=1 and the condition (4.20), the control law (4.27) is obtained.

Bibliography

[1] Hemami, A., *Design of light weight flexible robot arm*, Robots 8 Conference Proceedings, Detroit, USA, June 1984, pp. 1623-1640.

[2] Gravagne, Ian A., Walker, Ian D., *On the kinematics of remotely - actuated continuum robots*, Proc. 2000 IEEE Int. Conf. on Robotics and Automation, San Francisco, April 2000, pp. 2544-2550.

[3] Gravagne, Ian A., Walker, Ian D., *Kinematic Transformations for Remotely-Actuated Planar Continuum Robots*, Proc. 2000 IEEE Int. Conf. on Rob. and Aut., San Francisco, April 2000, pp. 19-26.

[4] Chirikjian, G. S., Burdick, J. W., *An obstacle avoidance algorithm for hyper-redundant manipulators*, Proc. IEEE Int. Conf. on Robotics and Automation, Cincinnati, Ohio, May 1990, pp. 625 - 631.

[5] Mochiyama, H., Kobayashi, H., *The shape Jacobian of a manipulator with hyper degrees of freedom*, Proc. 1999 IEEE Int. Conf. on Robotics and Automation, Detroit, May 1999, pp. 2837- 2842.

[6] Robinson, G., Davies, J.B.C., *Continuum robots - a state of the art*, Proc. 1999 IEEE Int. Conf. on Rob and Aut, Detroit, Michigan, May 1999, pp. 2849-2854.

[7] Ivanescu, M., Stoian, V., *A variable structure controller for a tentacle manipulator*, Proc. IEEE Int. Conf. on Robotics and Aut., Nagoya, 1995, pp. 3155-3160.

[8] Ivanescu, M., Florescu, M.C., Popescu, N., Popescu, D., *Position and Force Control of the Grasping Function for a Hyperredundant Arm*, Proc.of IEEE Int. Conf.on Rob. and Aut., Pasadena, California, 2008, pp. 2599-2604.

[9] Ivanescu, M., Bizdoaca, N., Florescu, M., Popescu,N., Popescu, D., *Frequency Criteria for the Grasping Control of a Hyper-redundant Robot*, Proc.of IEEE International Conference on Robotics and Automation, Anchorage, Alaska (ICRA 2010), May 3 - 8, 2010, pp. 1542-1549.

[10] Miller, D.W., Collins, S.A., Peltzman, S.P., *Development of Spatially Convolving Sensors for Structural Control Applications*, the 31st AIAA Structures, Structural Dynamic and Materials Conference, 1992, paper 90 - 1127.

[11] Krstic, M., Smyshlyaev, A., *Boundary Control of PDEs: A Short Course on Backstepping Design*, VCSB, 2006.

[12] Krstic, M., *Compensation of Infinite - Dimensional Actuator and sensor Dynamic*, IEEE Control Systems, February 2010, vol. 30, no. 1, pp. 22 - 41.

[13] Camarillo, D., Milne, C., *Mechanics Modeling of Tendon - Driven Continuum Manipulators*, IEEE Trans. On Robotics, vol. 24, no. 6, December 2008, pp. 1262 - 1273.

Int. J. of Computers, Communications & Control, ISSN 1841-9836, E-ISSN 1841-9844
Vol. V (2010), No. 5, pp. 768-774

Towards the implementation of Computer-Aided Semiosis

A.E. Lascu, S.C. Negulescu, C. Butaci, V. Cret

Alina E. Lascu, Sorin C. Negulescu
Lucian Blaga Univeristy of Sibiu, Romania
E-mail: alina.lascu@ulbsibiu.ro, sorin.negulescu@ulbsibiu.ro

Casian Butaci, Vasile Creţ
Agora University, Oradea and R&D Agora Ltd.
Cercetare Dezvoltare Agora Oradea, Romania
E-mail: casian12@yahoo.com, vcret@univagora.ro

Motto: *In the beginning was the Word*
Bible, John 1:1

Abstract: Computer-Aided Semiosis (CAS) is a concept coined by a team of researchers a couple of years ago. Since it is a promising domain due to the fact that responds to actual trans-cultural communication needed in the broad-band society - where often the message behind the words does not come clear - the subject ought being inquired more detailed as promised in other papers of the same authors. This interesting idea was inspired from Eco's theory of communication which states that the receiver "fills the message with significance"; hence it is vital for any communication and is strongly dependent on the cultures involved. In line with Eco's theory, the research in this area must be trans-disciplinary and anthropocentric. In the intention of narrowing the existing gap between the technological offers and user expectations the macro-architectural feature is that translation will progress from textual, semantically correct, to multimodal, culturally adequate, based on common concepts and "grammar" (rules to combine them into meaningful sentences); thus, this paper will present possible approaches towards the implementation of CAS. Given the fact the ontologies are considered to be the pillars of Semantic Web but also a key tool in implementing CAS, both will be a subject of this paper in the light of finding an implementation solution.
The paper is structured on five sections: the first will present the defining aspects of the concept relating it with previous research; the second section will deal with CAS approach and architecture, following with the state of the art regarding ontologies and their relation with Semantic Web. Among the conclusions, one is already noticeable: CAS could not be possible without a trans-cultural ontology.
Keywords: Computer-Aided Semiosis (CAS), Human-Computer Interaction (HCI), Ontologies, Semantic Web, Interfaces.

1 Introduction

The key to effective online cross-cultural communication is a well-designed transcultural ontology which to help in disambiguating between concepts that seem alike but their meaning differs from culture to culture; this tool is intended to be developed in the following years by a team of young researchers in order to substantiate and implement a new and innovative concept in the field of HCI: computer-aided semiosis (CAS).

University of Colorado, USA developed a study regarding cross-cultural communication strategy and came up with the conclusion that often intermediaries who are familiar with both cultures can be helpful in cross-cultural communication situations. They can translate both the substance and the manner of what is said. For instance, they can tone down strong statements that would be considered appropriate in one culture but not in another, before they are given to people from a culture that does not talk together in such a strong way. They can also adjust the timing of what is said and done. Some cultures move quickly to the point; others talk about other things long enough to establish rapport or a relationship with the other person. If discussion on the primary topic begins too soon, the group that needs a "warm up" first will feel uncomfortable. A mediator or intermediary who understands this can explain the problem, and make appropriate procedural adjustments [4].

The results of this study can be as well applied in ICT due to the fact that online cross-cultural communication could also use virtual intermediaries which to have access to a transcultural ontology assisting thus the user grasping the right meaning of a certain message, i.e. in written (chat), spoken (voice) and/or visual form.

In the light of the earlier scan, this paper will present possible approaches towards the implementation of CAS. Since the ontologies are thought to be the pillars of the Semantic Web but also an important tool in implementing CAS, both will be key-subjects of this paper in the quest of finding an implementation solution.

The paper is structured on five sections: the second will present the defining aspects of the concept relating it with previous research; the third will deal with CAS rational and approach, following with the state of the art regarding ontologies and their relation with Semantic Web. Among the conclusions, one is already easily remarked: at the online level, communication can be impaired not only by the cultural differences but also by a wide range of differences such as race, age, sex, profession, religion or disabilities; in this regard, if the transcultural ontology will prove its efficiency in disambiguating cultural concepts, consequently other ontologies could be also implemented for aiding the online communication process between different users.

2 Defining the Concepts in line with History

According to modern paradigms, the goal of using ICTs is "obtaining a service from a huge palette of available ones" and the means is "interacting with an entity". The "entity" is either a human (e.g. when speaking via mobile phones) or a device (e.g. when buying travel documents via computers) [1]. Therefore there are three possible ways of communication as described in [1]: a) "face to face", b) "face to interface", c) "interface to interface" in the near future context of semantic web, domain ontologies and so forth.

very agent metaphor. Thus, when users employ agents (in whatever domain of activity) they expect: a) personalization (agents act considering the specific momentary needs of their clients); b) authorization (agents act on behalf of their clients, within the limits stipulated by the hiring agreement); and c) competence. In short, the agent metaphor suggests that "I hire an agent when I do not have enough time or lack competence to handle the problem myself" [1]. As regards the interface agents, the emphasis is on the interface, entailing that the agent remains hidden (i.e., the users perceive just a "smarter functionality", no "pseudoavatars" [10] intervening in a human-to-human dialogue) [1]. Considering these features, one can say that an agent, or more specific an interface agent can act like a smart mediator which to have access to knowledge (i.e. ontology) and by using it efficiently to "translate" to the user only the meaningful messages, saving a lot of time that otherwise would have been wasted. In the 21st century time is money therefore it is vital mostly in the business sector but not restricted to.

Online communications nowadays means more than a showy website and a newsletter. If

new web-based technologies are joint with a society that is rapidly getting to think of online interactions as just as authentic as face-to-face ones, one has the possibility of radically easing the communication even in difficult (but frequent) circumstances, like those involving cross-cultural interaction ([2], [6], [7], [8], [9], [10], [11], [12], [13]) as debated and probed in previous paper of the authors.

In our previous researches we approached the transcultural interfaces' subject which could be explained in human-to-human communication as a progression from textual (semantically correct) conversions to multimodal (culturally adequate) ones [2], based on the concept of CAS. Since at this research stage the experimental models presented in the related papers are not agent-oriented, the reference to "agents" is found only in [1] and regards conceptual aspects as well as future development.

Thus far, the emphasise in our research was put on transcultural which represents the ability of people belonging to different cultures to communicate efficiently preserving in the same time their cultural identity [2]. This concept will be further on related with ontology which together, i.e. transcultural ontology, will embody the tool by which CAS could be validated as a possible new and challenging domain of HCI.

3 Rationale and Approach

As depicted from the title this paper intends to draw the sketch of what it should be the onward study on the implementation/ validation of CAS. Since the prime motives because of which we started this endeavour was already stated in the previous papers of this team, it is redundant to re-state them; instead this paper is set to focus on the approach and methods.

Since CAS was designed from an anthropocentric perspective, meaning to provide an assistance (i.e. an interface agent which to access a transcultural ontology, transfering the user only meaningful messages), lessening the linguistic hurdles (such as the traduttore-traditore effect), the logocratic pressure of (spoken or written) text, response time criticality, as well as the danger of distortions and noise, via a major upgrade in communication granularity: (one) idea instead of (many) words [9].

The ongoing study should be approached from a trans-disciplinary perspective, in respect with both humanists (i.e. linguists, psychologists) and technologists (i.e. ontology, interface designers and so on). When creating the ontology, the designer should bear in mind the way users and agents may "think"; how an user creates meaning from a piece of image of some sings (i.e. words) and how an interface agent does the same job effortlessly on the contrary. Though, do an agent depict the meaning same accurate as a human does? Or maybe detecting the true meaning of a message is even more endangered by human's scrambled mind which is very much contextual, in opposition with an interface agent which will act and respond based only on the given ontology and some very specific rules, and hence the probability to fail giving the expected meaning will be smaller. These are all questions which can not be answered at this point, some answers will be empirically uncovered, some may prove to be exactly contradictory with those thought in the first place. After all, this is what academically we call exploratory research. The approach will be adapted step by step base on the further findings.

4 Ontology. State of the art

People are able to use the Web in order to complete tasks such as finding the Icelandic word for "alphabet", reserving a plane ticket, or searching for a low price for an e-book. Still, computers cannot perform the same tasks without human direction because web pages are designed to be

read by people, not machines. The semantic web is a vision of information that is understandable by computers, so that they can perform more of the dull work involved in finding, sharing, and combining information on the web [3].

The vision of Tim Berners-Lee regarding semantic web [3]: "I have a dream for the Web (in which computers) become capable of analyzing all the data on the Web - the content, links, and transactions between people and computers. A 'Semantic Web', which should make this possible, has yet to emerge, but when it does, the day-to-day mechanisms of trade, bureaucracy and our daily lives will be handled by machines talking to machines. The 'intelligent agents' people have touted for ages will finally materialize" is related in this study with Eco's semiosis theory (which states that the receiver "fills the message with significance" [5]), meaning that on the same features' as semantic web's CAS could be implemented but using instead a dedicated/personalized ontology based on which to develop the "translation process".

An ontology is a formal representation of a set of concepts within a domain and the relationships between those concepts [17]. The ontology envisioned by this research team is based on the idea of Maya script (a logographic type of script which used both logograms and syllabic characters [16]) by replacing, where possible, the words with images as in the catchphrase: "a picture it is worth a thousand words". Figure 1 illustrates the way the correspondence between a word-based ontology and an image-based ontology can be created, of course, the example given in this paper is very much simplified. The demarcation line was traced in order to separate the abstract and the concrete synonyms of the word apple. On one side of the line there can be easily remarked every-day interpretations of the word apple, which can without any doubt be recognized by anyone. On the other side, the other interpretations depict more abstract representations of the word apple which require a higher level of knowledge in order to be grasped, i.e. the apple polisher, the temptation (religiously), Wilhelm Tell representation (historically), Newton's gravitation theory (physical sciences), "an apple a day keeps the doctor away" (health idiom); the point is that the ontology must be trans-disciplinary and trans-culturally created in order to gather all the possible meanings and definitions for a word/concept. Disambiguation can further on continue and the figure presented above be expanded with other concepts and relation between them. An example at hand for expanding the ontology would be adding other characteristics for the fruit apple such as variety name (e.g. Granny Smith, Pink Lady, Red Delicious, Golden Delicious and so on), colour (e.g. red, green, pink, yellow), taste (e.g. sour, sweet, etc.). Also, in order to disambiguate AppleŽ's logo there can be made a separation between the old and the new logos. An important disambiguation which must be made in the first place is between English word apple and Swedish word äpple which happens to mean the same thing but they belong to different cultures which can generate later on other misunderstanding problems.

Briefly, the implementation of CAS will follow this ontology framework where the icon position has syntactic role (in line with ontology rules) and semantic role for CAS (to reduce the differences between "intentio auctoris" and "intentio lectoris") [2]. This kind of ontology based on visual rules can be further enhanced by using animations instead of images.

Considering this idea, the role of the interface agent will be to "translate". In German, the verb "to translate" means "übersetzen", the reason for this association is that in its most basic visualization, the German word means "to carry something from one side of the river to the other side of the river". In opposition the English word "to translate" does not immediately evoke the same image in the mind. In German, one can say: "mit der Fähre den Fluss übersetzen" which would literally mean: "to translate the river on a ferry boat". Anyway the translator so to name it must take in consideration both cultures from which to which the translation is being made; one must know that ad literam translation doesn't apply especially when a more subtle message (culturally dependent) must be transmitted, therefore the visualisation of the word "übersetzen" (to move from one side of the river to the other side of the river) leads to

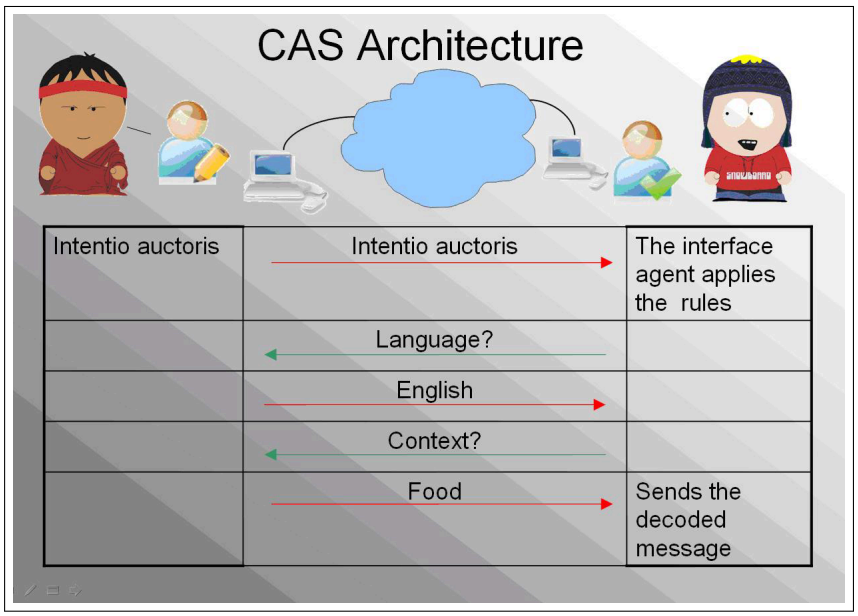


Figure 1: Example of architecture for CAS[1]

several important insights into the nature of translation. In this context, the role of the agent is carrying something across the river, whether it is from here to there or from there to here. The cargo can have a multiplicity of shapes: the description of a technological object, a cultural or historical phenomenon, a poetic image, a metaphorical expression, or a human emotion, to name only a few; the parameters of each word are quite fragile; no two people will take the exact same meaning from a word [14], so the interface agent will have the hardest ever job, to extract the meaning of a message taking into consideration the transcultural ontology in line with the cultural background of both users involved in the communication process.

In order to validate CAS concept as a possible new subdomain of HCI, the researches have to focus from now on, on the implementation by creating a transcultural ontology on the framework presented in the previous section involving in this process preferable a trans-disciplinary team of researchers (linguists, psychologists, anthropologists, designers, programmers and so on) - by keeping in mind the fact that an ontology deals with the nature of existence so it is a too impressive domain to be approached only by a team of thrilled researchers, willing to take the burden of exploratory research.

The future work will consider refining the ontology framework and hopefully in three-year time span (during the PhD studies of the first author) the objectives to be fulfilled involving European teams of researchers interested in this kind of projects.

5 Conclusions and Future Work

In order to validate CAS concept as a possible new subdomain of HCI, the researches have to focus from now on, on the implementation by creating a transcultural ontology on the framework presented in the previous section involving in this process preferable a trans-disciplinary team of researchers (linguists, psychologists, anthropologists, designers, programmers and so on) - by keeping in mind the fact that an ontology deals with the nature of existence so it is a too impressive domain to be approached only by a team of thrilled researchers, willing to take the burden of exploratory research.

The future work will consider refining the ontology framework and hopefully in three-year

time span (during the PhD studies of the first author) the objectives to be fulfilled involving European teams of researchers interested in this kind of projects.

Acknowledgements:

This research is fully supported by the POSDRU/88/1.5/S/60370 project which is co-financed by the European Social Fund through the Sectoral Operational Programme for Human Resources Development 2007-2013.

We thank Mr. Cătălin Boaru from Apple IMC Romania for kindly granting us the permission to use the Apple® logo.

Bibliography

[1] Bărbat, B.E. (2009). Interface Agents for Transcultural Communication: A Framework . In Sapio, B. (Ed.), The Good, the Bad and the Challanging. The user and the future of information and communication technologies (pp.666-675). Copenhagen: Cost Action 298.

[2] Bărbat, B.E., Negulescu, S.C., Lascu, A.E., Popa, E.M. (2007). Computer-Aided Semiosis. Threads, Trends, Threats. In Mastorakis, N.E. (Ed.), Proc. of the 11th WSEAS International Conference on Computers (pp.269-274). Agios Nikolaos, Crete: ICCOMP '07.

[3] Berners-Lee, T. (1999). Weaving the Web: The Original Design and Ultimate Destiny of the World Wide Web by its Inventor. New York: HarperOne.

[4] Conflict Research Consortium (1998). Cross-Cultural Communication Strategies. Retrived 2010, from University of Colorado, USA. Web site: http://www.colorado.edu/conflict/peace/treatment/xcolcomm.htm.

[5] Eco, U. (2005). The Limits of Interpretation. Bloomington, SUA: Indiana University Press.

[6] Georgescu, A.V., Lascu, A.E., Bărbat, B.E. (2008). Protensity in Agent-Oriented Systems. Role, Paths, and Examples. Int. J. of Computers, Communications & Control, III, 304-309.

[7] Lascu, A.E., Fabian, R. (2007). e-Semiotics for Romanian-German trans-cultural interfaces. In Sapio, B. et al (Ed.), The Good, the Bad and the Unexpected: The User and the Future of Information and Communication Technologies (pp.on CD). Moscow, Russian Federation: COST Action 298 Participation in the Broadband Society.

[8] Lascu, A.E., Georgescu, A.V. (2009). From Extensity to Protensity in CAS: Adding Sounds to Icons. In Esposito, A. et al (Ed.), Multimodal Signals: Cognitive and Algorithmic Issues (pp.130-137). Vietri sul Mare, Italy: Springer.

[9] Lascu, A.E., Moisil, I., Negulescu, S.C. (2009). Computer-Aided Semiosis mirrored in Creolization. Rational and Approach . International Journal of Advanced Statistics and IT&C for Economics and Life Sciences, 1(1), 38-43.

[10] Lascu, A.E., Negulescu, S.C., Cioca, M., Zerbes, M.V. (2009). Interface Agents as Virtual Tutors. Conference Proceedings of Balkan Region Conference on Engineering and Business Education & International Conference on Engineering and Business Education, 2, 626-629.

[11] Lascu, A.E., Negulescu, S.C., Kifor, C.V. (2009). Different Time Perception in Creolization Mirrored in Transcultural Interface for "Immediate". In Sapio, B. et al (Ed.), The Good, the Bad and the Challenging - The user and the future of information and communication technologies (pp.661-666). Copenhagen, Denmark: COST Action 298 Participation in the Broadband Society.

[12] Negulescu, S.C., Lascu, A.E., Oprean, C. (2009). Cultural Differences in Decision-Making. A Transcultural Interface for Gambler's Fallacy. In Sapio, B. et al (Ed.), future of information and communication technologies (pp.656-661). Copenhagen, Denmark: COST Action 298 Participation in the Broadband Society.

[13] Prundurel, A., Negulescu, S.C., Lascu, A.E. (2007). Mini-Ontology for Trans-Cultural Interfaces. In Sapio, B. et al (Ed.), The Good, the Bad and the Unexpected: The User and the Future of Information and Communication Technologies (pp.on CD). Moscow, Russian Federation: COST Action 298 Participation in the Broadband Society.

[14] Rainer Schulte (1999). The translator as mediator between cultures. Retrived 11.2009, from Translation Studies. Web site: http://translation.utdallas.edu/translationstudies/mediator_essay1.html.

[15] Russell S., Norvig P. (2003). Artificial Intelligence: A Modern Approach. NJ: Prentice Hall.

[16] Simon Ager (2009). Mayan script. Retrived 2010, from Omniglot - writing systems and languages of the world. Web site: http://www.omniglot.com/ writing/mayan.htm.

[17] Wikipedia (2009). Ontology (information science). Retrived 11.2009, from Wikipedia. Web site: http://en.wikipedia.org/wiki/Ontology_(information_ science).

Int. J. of Computers, Communications & Control, ISSN 1841-9836, E-ISSN 1841-9844
Vol. V (2010), No. 5, pp. 775-782

Tool Support for fUML Models

C.-L. Lazăr, I. Lazăr, B. Pârv, S. Motogna, I.-G. Czibula

Codruţ-Lucian Lazăr, Ioan Lazăr, Bazil Pârv,
Simona Motogna and István-Gergely Czibula
Department of Computer Science
Babeş-Bolyai University, Cluj-Napoca, Romania
Romania, 400084 Cluj-Napoca, 1 M. Kogălniceanu
E-mail: {clazar, ilazar, bparv, smotogna, istvanc}@cs.ubbcluj.ro

Abstract: In this paper we present a tool chain that aids in the construction of executable UML models according to the new Foundational UML (fUML) standard. These executable models can be constructed and tested in the modeling phase, and code can be generated from them towards different platforms. The fUML standard is currently built and promoted by OMG for building executable UML models. The compatibility of the executable models with the fUML standard means that only the UML elements allowed by fUML should be used for the abstract syntax and the extra constraints imposed by the fUML standard should be considered. The tool chain we propose is intregrated with the existing UML tools of Eclipse modeling infrastructure.
Keywords: Class Diagram, fUML, Action Language, Code Generation, Eclipse.

1 Introduction

The executable models are models that can be executed and tested without having to generate code from them and test them in a specific platform. Creating executable models in the process of developing an application is considered to be a good approach, because the business model and functionality can be implemented in the modeling phase, while the decisions regarding the implementation in the specific platform can be delayed to the phase for code generation from the model. The executable models have the advantage of not being poluted with code that is not related to the business logic, keeping the model and functionality much more compact and clear. And they also have the advantage of not being tied to a specific platform or technology.

The Foundational UML (fUML) [1] is a computationally complete and compact subset of UML [2], designed to simplify the creation of executable UML models. The semantics of UML operations can be specified as programs written in fUML. We introduced in a previous paper [3] an action language based on fUML, with a concrete syntax that follows the principles of the structured programming, which is supported by the modern languages like Java and C++.

In this paper we describe a tool chain that is aimed at building and testing executable UML models, as well as generating code towards different target platforms. The generated code is meant to be complete, with no code placeholders for the developer to fill out.

The tools are built on top of the Eclipse Modeling Framework Project (EMF) and some other projects from Eclipse that are part of the Eclipse Modeling Tools distribution. These tools are integrated with the Eclipse modeling infrastructure and with each other.

The remainder of the paper is organized as follows: section 2 presents the infrastructure needed to build executable models and section 3 presents the research problem. Then, section 4 describes the tool chain we propose. Section 5 presents the existing work related to ours and section 6 gives the conclusions of this paper.

2 Background

The UML Class Diagrams are widely used to create the structure of a model. They are intuitive and easy to use. However, the UML behavior diagrams (Activity Diagrams and State Machines) are not easy to use for larger models. The fUML standard provides a simplified subset of UML Action Semantics package (abstract syntax) for creating executable UML models. It also simplifies the context to which the actions may be applied. For instance, the structure of the model will consist of packages, classes, properties, operations and associations, while the interfaces and association classes are not included.

However, creating executable fUML models is difficult, because the UML primitives intended for execution are too low level, making the process of creating reasonable sized executable UML models close to impossible.

A concrete textual syntax is needed, because it enforces a certain way of constructing models. This means that a lot of elements that need to be created explicitly in the graphical UML Activity Diagram can be implicitly derived from the syntax and created automatically.

OMG issued an RFP for a concrete syntax for an action language based on fUML [4]. Because there is no standardized action language at this moment, we proposed an action language of our own [3] as part of our framework: ComDeValCo (Framework for Software Component Definition, Validation, and Composition) [5–7].

The fUML standard also specifies how to create a virtual machine that can execute fUML executable models.

To generate code from MOF compliant models, OMG created MOF Model to Text Transformation Language [8], which is suitable to generate code from fUML models.

3 Research problem

The research problem is to investigate the creation of a tool chain that aids in the construction of fUML models that can be created and tested in the modeling phase, and from which code can be generated towards different platforms.

The tehniques mentioned below represent the core of a tool chain that can change the development experience for the better, by simplifying the process and allowing the developer to take decisions in the proper stage of development.

Model creation. To create the structure of the model, the usual UML Class Diagrams should be used. These Class Diagrams, however, must restrict the elements that can be used to those included in the fUML standard.

To create the part of the model corresponding to the behavior of the operations, an Action Language based on fUML needs to be used. This is because it is close to impossible to use UML Activity Diagrams for this task, as the user will need to create, configure and relate too many elements.

The Class Diagram editor needs to be integrated with the Action Language textual editor, which is used to create the behavior for each operation. The integration refers to the ability to select a class or an operation from the model and, with a simple action (double-click or some key shortcut), the action language editor for the behaviors can be opened and used for the selected element. The action language editor must be able to load the existing behavior under the selected element and display it properly. Also, on save action, it should properly update the model under the selected element.

An action language that follows the principles of the structured programming is to use by many programmers familiar with structured programming languages. Thus, we consider this

an important factor in choosing the action language, even though the fUML standard allows non-structured control flow.

Because many programmers are familiar with object-oriented languages like Java or C++, having a similar syntax will make the language much easier to be used. Also, the effort for learning the new language will be very much reduced. The action language will need to create the abstract representation for the statements and control structures as provided by the programming languages mentioned above, it must support complex expressions and easy access to parameters and variables.

Model execution. After creating the model, or parts of it, there is a need to simulate the execution from certain starting points. This can be achieved with a virtual machine that knows how to execute fUML models. This means that it can execute the activities built with the action language editor described above.

Code generation. After creating and testing the model, the only thing left is to generate code to a specific platform. A code generation tool should take into considerations any tags (stereotypes) placed on the model elements, and adjust the code generation process accordingly.

If the resulted action language has a well structured abstract representation, it should be possible to apply model transformations and generate code to structured programming languages with little effort. Thus, the executable models for which the behavior is created with such an action language can be converted to a multitude of platforms.

4 Tool Chain

In this section we present the tools we use to create fUML based executable models. The tools are integrated in the Eclipse modeling infrastructure. The UML Class Diagram Editor and the fUML Execution Framework exist and they only need to be integrated with the other tools. The fUML based Action Language Editor is a tool proposed by us. The Code Generation Utility exists, but the templates used to generate code are written by us.

We use a Point-of-Sale (POS) example model, presented in fig. 1. The example model consists of a `POS` class, which contains a list of `Products`. The user can make a new `Sale` (stored in `POS` as the `currentSale`) by invoking the `makeNewSale` operation and by adding `SaleItems` to the current sale in the form of product code and quantity. The `POS` component finds the `Product` associated with the given product code and, if present, passes the product and quantity to the `currentSale`. The `Sale` creates a new `SaleItem` instance and adds it to its list of sale items.

Fig. 2 and 3 present the implementation of the operations, as it is written using our action language. Fig. 4 shows the abstract representation in fUML of `Sale.addItem`'s behavior.

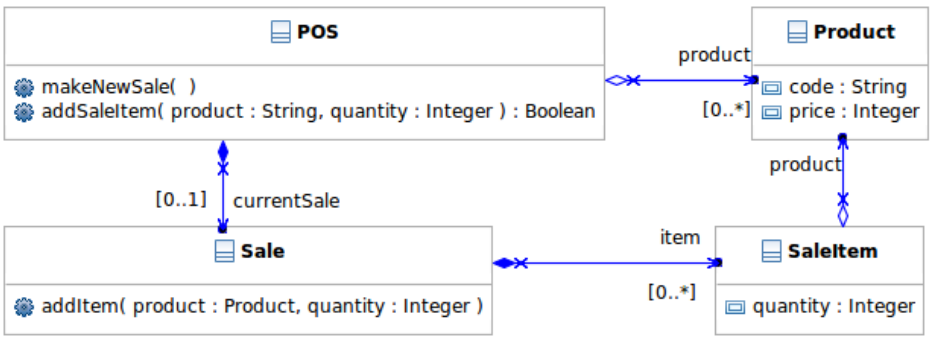


Figure 1: POS Example Model - Class Diagram Editor

4.1 UML Class Diagram Editor

The UML meta-model is provided by the Eclipse UML2 project [9], which is part of the larger Model Development Tools (MDT) project. UML2 is built on top of EMF(Core) [10] (which is part of the bigger Eclipse Modeling Framework project - EMF) with some adaptations, as UML has a structure that is not compatible with EMF.

The UML2 project provides the Java classes that correspond to the UML model, so that a UML model can be represented in Java. The model can be saved in files with the "uml"extension in XMI format, as it is standardized by OMG.

The UML2 project comes with a tree-based editor that allows the user to manipulate an "uml" file and build the UML models using a tree. This editor is not quite easy to use, but, for building only the structure of classes of a model, it is enough.

The Eclipse UML2 Tools [11], which is also part of MDT, provides a Class Diagram editor. This project allows the user to attach UML Class Diagrams to the UML models and build the UML models using a Class Diagram editor. The editor is quite mature and easy to use, and most programmers are used to building models using Class Diagrams, so this editor is the best choice for a tool to build the class structure of a model (fig. 1).

Because we want to build fUML based models, it is important that only the elements allowed by fUML are used. The Interface, for instance, was excluded from fUML, so the user should not include interfaces in the model. To simulate an interface, the programmer can use abstract classes with public abstract methods. A specialized editor for Class Diagrams that allows only fUML elements to be used, is considered for the future.

4.2 fUML based Action Language Editor

We introduced in a previous paper [3] an action language based on fUML, with a concrete syntax similar to the concrete syntax of the modern programming languages like Java or C++. This action language follows the principles of the structured programming.

We have also built an Eclipse textual editor for this action language using the Xtext project [12], which is the only remaining project from the bigger Textual Modeling Framework (TMF) project from Eclipse. This textual editor can take the textual representation for the functionality of an `Operation` (fig. 2 and 3) and convert it to an `Activity` with all the UML `Actions` and other elements necessary to provide the same functionality in UML (fig. 4). This `Activity` is added to the `Operation`'s classifier (of `Class` type) and set as the behavior of the `Operation`. Only the elements allowed by fUML are used to create this UML model of the `Activity`.

The textual editor for the Action Language is integrated with the Class Diagram editor, so that when the user wants to edit the behavior of an Operation, s/he only needs to double-click, or use a context menu item of the Operation, or use a key shortcut to open the textual editor. After the behavior is created, the user needs to save it by pressing Ctrl+S, at which point the main "uml" model file is updated to contain the model for the Operation's behavior.

4.3 fUML Execution Framework

The fUML standard specifies how a virtual machine for fUML models should work, and there is a reference implementation in progress from ModelDriven.org [13]. We managed to integrate this tool in the Eclipse workbench, so that we can test our fUML models. We can either pass the activities created with our action language, along with the needed parameters, directly to the execution framework to be executed, or we can write test activities with our action language and execute the tests.

```
public makeNewSale() {
    self.currentSale := new Sale;
}

public addSaleItem(code:String, quantity:Integer) :  Boolean {
    def product:Product := null;
    foreach (prod in self.product) {
        if (code = prod.code) {
            product := prod;
        }
    }
    if (product = null) {
        return false;
    } else {
        self.currentSale.addItem(product,quantity);
        return true;
    }
}
```

Figure 2: POS::makeNewSale and POS::addSaleItem Activity Concrete Syntax

4.4 Code Generation Utility

To generate code, we used the Acceleo project [14], which is part of the bigger Model To Text (M2T) project from Eclipse. This project implements the MOFM2T standard from OMG. It allows the user to create templates, which can later be used to generate code from the models.

In our case, these templates need to work on the elements included in fUML. Generating the structure of classes is straightforward. However, generating code for the behavior of the `Operations` is more complex, because the structure of the elements and the way the actions are connected needs to be considered. Because we took the decision to follow the structured programming principles, the UML model resulted for the `Activities` is well structured, so we are able to generate code with ease to languages like Java or C++. If we wouldn't have taken this decision, then the code generation step would have been a real problem, as fUML allows the user to create interactions between actions that might be impossible to represent in the languages mentioned above. Also, due to the resemblance in concrete syntax between the action language and the target languages, we are able to generate compact code in the target languages.

An important note is that the templates are specifically written for `Activities` constructed with an editor compatible with our Action Language. This is because the way the elements are structured and the way they interact is very important. If these aspects are not followed, the templates will produce a bad output.

For a fUML `Activity` that does not respect these constraints and which might be constructed using a different fUML based Action Language, a special set of templates need to be written. This is not necessarily an issue, because the Action Languages can be used in conjunction to construct UML models and because the templates only need to be written once. Our Action Language is a general purpose language and might be a bit hard to specify and build a proper editor for it. But a different Action Language that is specialized on a more concrete domain could have a simpler syntax and the code generation templates might be easier to be written.

Fig. 5 shows a snippet from the Acceleo templates we used to generate Java code. It shows how the statements are iterated, considering the `StructuredActivityNodes` that contain the

```
1  public addItem(product:Product, quantity:Integer) {
2      def newItem :  SaleItem := new SaleItem;
3      newItem.product := product;
4      newItem.quantity := quantity;
5      self.item.add(newItem);
6  }
```

Figure 3: Sale::addItem Activity Concrete Syntax

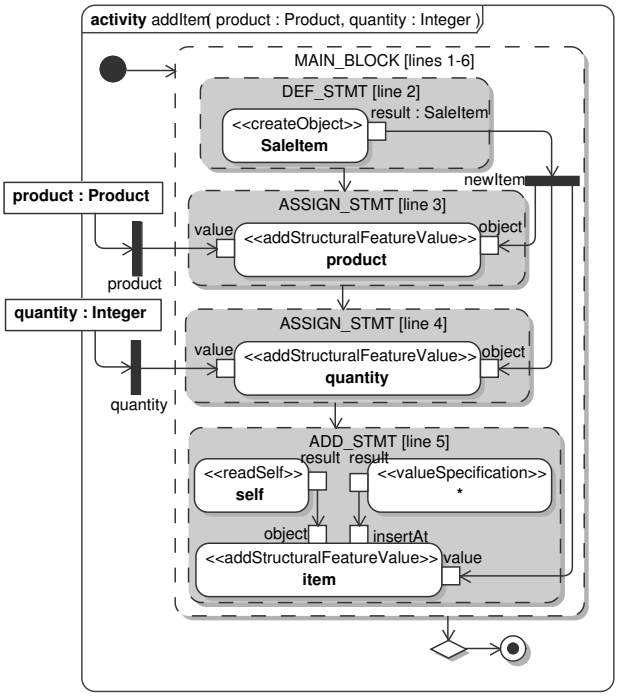


Figure 4: Sale::addItem Activity - fUML Abstract Syntax

actions for each statement, and the edges that enforce the sequential flow between these nodes.

The `operationActivity` template selects the only `StructuredActivityNode` it contains, which is the node containing all the statement nodes (*MAIN_BLOCK* node from fig. 4). The `activityBlock` template selects the statement node with no incoming edges, which represents the node corresponding to the first statement (*DEF_STMT [line 2]* from fig. 4).The `iterateBlockStatements` template prints the text for the statement node by calling the `blockStatement` template and, if the node has an outgoing edge to the next statement node, calls recursively the `iterateBlockStatements` for the next statement node.

5 Related Work

Only a few frameworks for building executable models exist and some of them are proprietary: Mentor Graphics' BridgePoint product with its Object Action Language (OAL) [15], Kennedy Carter's iUML product with its Action Specification Language (ASL) [16] and others. Some of these frameworks have their action languages based on the Action Semantics package from UML, but none is based strictly on fUML, as the standard is still in Beta version.

```
[template public operationActivity(a : Activity)]
[a.node->any(oclIsTypeOf(StructuredActivityNode)).oclAsType(StructuredActivityNode).activityBlock('    ')/]
[/template]

[template private activityBlock(blockNode : StructuredActivityNode, ident : String)]
{
[let firstNode : ActivityNode = blockNode.node->any(incoming->isEmpty())]
[blockNode.iterateBlockStatements(firstNode.oclAsType(StructuredActivityNode), ident)/][/let]
[ident/]}
[/template]

[template private iterateBlockStatements(parentNode : StructuredActivityNode,
                          stmtNode : StructuredActivityNode, ident : String)]
[stmtNode.blockStatement('    '.concat(ident))/]
[if (stmtNode.outgoing->notEmpty())]
[let nextNode : ActivityNode = stmtNode.outgoing->any(true).target]
[if (parentNode.node->includes(nextNode))]
[parentNode.iterateBlockStatements(nextNode.oclAsType(StructuredActivityNode), ident)/][/if]
[/let]
[/if]
[/template]
```

Figure 5: Acceleo Template Snippet for Java

6 Conclusions and Further Work

In this paper we presented a tool chain that can be used to create, execute and generate code from fUML executable models. The tool chain integrates an UML Class Diagram editor, a fUML Action Language editor, a fUML Execution Framework and a Code Generation Framework.

We plan to further work on this tool set, to better integrate the tools, in order to improve the user experience with the tool set as a whole. A specialized editor for Class Diagrams that allows only fUML elements to be used, is considered as future work. Also, we plan to investigate the possibilities of applying transformations to the models created by our action language, and to generate code in other programming languages.

Acknowledgment

This work was supported by the grant ID 546, sponsored by NURC - Romanian National University Research Council (CNCSIS).

Bibliography

[1] *Semantics of a Foundational Subset for Executable UML Models*, Object Management Group Standard, Rev. 1.0, Beta 2, October 2009. [Online]. Available: `http://www.omg.org/spec/FUML/1.0/Beta2/PDF/`

[2] *UML Superstructure Specification*, Object Management Group Standard, Rev. 2.2, February 2009. [Online]. Available: `http://www.omg.org/spec/UML/2.2/Superstructure/PDF/`

[3] C.-L. Lazăr, I. Lazăr, B. Pârv, S. Motogna, and I.-G. Czibula, "Using a fUML Action Language to construct UML models," *Proceedings of the 11th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing*, 2009.

[4] *Concrete Syntax for a UML Action Language*, Object Management Group Request For Proposal, 2008. [Online]. Available: `http://www.omg.org/docs/ad/08-08-01.pdf`

[5] B. Pârv, S. Motogna, I. Lazăr, I.-G. Czibula, and C.-L. Lazăr, "ComDeValCo - a Framework for Software Component Definition, Validation, and Composition," *Studia Universitatis Babeş-Bolyai, Informatica*, vol. LII, no. 2, pp. 59–68, 2007.

[6] I. Lazăr, B. Pârv, S. Motogna, I.-G. Czibula, and C.-L. Lazăr, "An Agile MDA Approach for Executable UML Structured Activities," *Studia Universitatis Babeş-Bolyai, Informatica*, vol. LII, no. 2, pp. 101–114, 2007.

[7] C.-L. Lazăr and I. Lazăr, "On Simplifying the Construction of Executable UML Structured Activities," *Studia Universitatis Babeş-Bolyai, Informatica*, vol. LIII, no. 2, pp. 147–160, 2008.

[8] *MOF Model to Text Transformation Language (MOFM2T)*, Object Management Group Standard, Rev. 1.0, January 2008. [Online]. Available: `http://www.omg.org/spec/MOFM2T/1.0/PDF/`

[9] *UML2*, The Eclipse Foundation, 2010. [Online]. Available: `http://www.eclipse.org/modeling/mdt/?project=uml2`

[10] *Eclipse Modeling Framework (Core)*, The Eclipse Foundation, 2010. [Online]. Available: `http://www.eclipse.org/modeling/emf/?project=emf`

[11] *UML2 Tools*, The Eclipse Foundation, 2010. [Online]. Available: `http://www.eclipse.org/modeling/mdt/?project=uml2tools`

[12] *Xtext - a programming language framework*, The Eclipse Foundation, 2010. [Online]. Available: `http://www.eclipse.org/Xtext`

[13] *Foundational UML Reference Implementation*, ModelDriven.org, 2009. [Online]. Available: `http://portal.modeldriven.org/project/foundationalUML`

[14] *Acceleo*, The Eclipse Foundation, 2010. [Online]. Available: `http://www.eclipse.org/modeling/m2t/?project=acceleo`

[15] *Object Action Language Reference Manual*, Mentor Graphics, 2009. [Online]. Available: `http://www.mentor.com/products/sm/techpubs/object-action-language-reference-manual-38098`

[16] *UML ASL Reference Guide*, Kennedy Carter Limited, 2003. [Online]. Available: `http://www.ooatool.com/docs/ASL03.pdf`

Int. J. of Computers, Communications & Control, ISSN 1841-9836, E-ISSN 1841-9844
Vol. V (2010), No. 5, pp. 783-791

An Algorithm for Customer Order Fulfillment in a Make-to-Stock Manufacturing System

D. Lečić-Cvetković, N. Atanasov, S. Babarogić

Danica Lečić-Cvetković, Nikola Atanasov, Sladan Babarogić
University of Belgrade, Faculty of Organizational Sciences
Jove Ilića 154, 11000 Belgrade, Serbia
E-mail: {danica,nikola.atanasov,sladjan}@fon.rs

Abstract: In the competitive environment, many manufacturers are increasingly focusing on designing the systems that help them to manage variable demand and supply situations. Dynamic allocation of demands is very important in case of customer order allocations. Order promising and allocation can be based on the simple sequence that enables a manufacturing company to receive orders unless there are some other priority orders. Manufacturing company can also manage allocations of supply to key customers and channels, thereby ensuring that they can meet contractual agreements and service levels in the priority that yields better profit. This paper will focus on a Make-to-Stock order fulfillment system facing random demand with random orders from different classes of customers. Available-to-promise (ATP) calculating from master production schedule (MPS) exhibits availability of finished goods that can be used to support customer order allocation. This order allocation system is adapted in MTS (make-to-stock) production model and all orders are treated according to maximization of customer service policy. It allows incoming purchase orders as well as existing inventory on hand to be selected and allocated to customer sale orders and back orders. The system then automatically allocates the available stock to the selected sales orders. We developed an integrated system for allocation of inventory in anticipation of customer service of high priority customers and for order promising in real-time. Our research exhibits three distinct features:
(1) We explicitly classified customers in groups based on target customer service level;
(2) We defined higher level of customer selection directly defined according to company strategy to develop small and medium customers;
(3) We considered backorders that manufacturing company has to fulfill in order to maximize overall customer service for certain customers.
Keywords: manufacturing system, order allocation, customer service.

1 Introduction

Production companies are facing the dilemma on whether to increase or decrease the capacities of production lines and/or the capacities of production plants. Everyday, production managers are facing the dilemma on whether to buy a new machine, to increase the number of shifts within the production process, or to employ new workers, in spite of the fact that such decisions also affect the middle-term and long-term production planning, and should not be made hastily. Sales forecasting and identification of an increase in demand represent the starting point for production capacity planning, and also for making a decision on capacity expansion. Until the available capacities are increased, the production company has a goal of satisfying the market

demand by means of available production capacities. The allocation of available finished products to customer orders requires an efficient system of allocation aimed at improving the overall operating efficiency. Operating efficiency is directly related to quantities of goods produced and the profit from sales. In addition to profit-oriented decision on selection of orders to be fulfilled, one needs to consider the customer service, since the overall operating of a production company depends on customers. Customers, who account for a large share in the sales results, require special attention, and the fulfillment of every order. There is also a group of customers who constantly increase their orders, and thus expect adequate and better service. Small but numerous customers, thus shape the overall sales of a production company. Some of them represent the future potential for the increase in sales and also for the increase in incomes of the production company. The abovementioned facts highlight the importance of making a decision while selecting the orders to be fulfilled, and the necessity of having an algorithm that will efficiently and effectively perform the allocation of limited quantities of available finished products.

This paper is structured as follows: the second chapter presents relevant approaches of other authors in solving the observed problem. The third chapter defines the problem of allocation of limited resources in production companies. The problem of allocation refers to the distribution of limited amount of production to customer orders, with the goal of maximizing the percentage of fulfilled orders. Aiming to define efficient and effective allocation, the fourth chapter describes the procedure of the development of algorithms for fulfilling the orders in the make-to-stock production system. In their conclusion, authors cited the main advantages of the proposed algorithm for allocation of limited production, as well as the possible directions of development and upgrading the algorithm aimed at improving the performances of the production company.

2 Related work

Providing the ordered quantities of products represents the key function in planning operations within the entire supply chain. The system of providing stocks is recognized as a key challenge in production companies. The basic goal of the process of providing stocks ATP/CTP (ATP – available-to-promise, and CTP – capable-to-promise) is providing a reliable response to customer demands, taking into consideration the wide range of information and limitations that exist in the distribution channel network. Key measures of performances within production companies, according to [6] are recognized in measuring the level of customer service, and customer satisfaction. Traditional approaches to order provision, adjusted to MTS production system, were described in [1] while considering the available stocks of finished products for order fulfillment in line with the principle First Come – First Served (FCFS), without awarding priority to either customers or orders.

The priority of received orders also represents a significant factor in providing stocks for order fulfillment. Optimization of orders is carried out based on the maximization of profit for the entire operating, through the system of production planning represented in [9]. Priority is given to orders in the following way: (1) orders are given the priority if they are already in the forecast of sales, i.e. (2) orders are given high-priority if the delivery of such orders should bring higher profit to operating of a production company. The general application of ATP/CTP decision system was implemented with an aim to improve the profit and system performances, in line with [6]. By considering the Advanced Planning Systems (APS), it is possible to identify various approaches in defining the priority of customers, based on which it will be possible to fulfill the incoming orders. The essential idea of the approach described in [9] is the segmentation of customers with the goal of increasing the overall income of the production company through order acceptance, and delivery of orders with the greatest profit. ATP model allows for such system, since it provides quantities in advance in accordance with certain customer segments, by

satisfying only the orders of priority customers.

ATP system is based on mixed-integer programming, with the goal of optimizing the utilization of limited production capacities, in order to provide timely information regarding the fulfillment of a customers order. In addition, the abovementioned problem is according to [2] classified amongst the dynamic models of order management with limited capacities based on the profitability analyses. If a production company has limited production capacities, it is clear that the company will decide to reject some of the incoming orders, which will directly affect the profitability of operating. The decision on rejecting orders is based on the comparison of orders, since company rejects those who yield lower profit.

ATP system that is based on providing quotes for important customers in a defined timeframe, in order to provide timely information on the delivery date, is described in [9]. ATP defined in a way that orders yielding higher profit for the producer must be fulfilled with priority, in spite of the fact that there are less profitable orders still 'on hold'. Meyr focused ATP on MTS system through the assumption that the supply of finished products is fixed with available stocks and the ongoing production, which will be available in a short period of time. The model of allocation of limited quantities of finished products through customer segmentation, developed by Kilger and Meyr, is based on profit assessment that is generated by fulfillment of orders in accordance with their priority. New orders could be fulfilled by the allocation of quantities granted to the group of customers they belong to, i.e. if there is no 'free stock' in that group from the quantities granted to lower-priority groups. Such method prevents orders of low-priority customers to be fulfilled prior to orders of high-priority customers, who help in generating higher profit. The ATP allocation system is based on defining the priority classes with the goal of maximizing the overall profit of the production company. On the grounds of various researches, according to [9], it was determined that the FCFS system for customers orders bring the best results with limited production capacities, if case the production is realized based on the forecast accuracy. If there are classes defined in accordance with the priority of customers, the quantity awarded to certain class within a cycle, if not allocated, stays booked for the same class in the following cycle. The same author in [10] presents that the provisional allocation of stocks for certain customer classes within a defined period of time could give a significant contribution if the customers demand within the class could be anticipated with guaranteed accuracy.

3 Order fulfillment problem definition

Production companies operating on markets with irregular demand often face the problem of insufficient stock of finished products. Until it makes a decision on the expansion of production capacities, a production company has the goal of fulfilling as much orders as possible, only with available production capacities and quantities of finished products. The allocation of finished products to customer orders is realized through the process of allocation within the MTS (make-to-stock) production system. The problem of allocation of available products to incoming orders is listed among problems of the allocation of limited resources. In case the incoming customer orders within one cycle do not exceed the available stocks of finished products, the allocation is complete and all orders are fulfilled. In the opposite case, when the sum of overall orders is greater than the quantity of available stock, there is a need to define the way of product allocation, i.e. rules must be introduced to help the allocation of products to incoming customer orders. With the goal of maximizing the effect of allocation of available products, the need for the development of allocation algorithm has been perceived, with which to enable the fulfillment of orders in line with their priority. The fulfillment of orders of high-priority customers contributes to the maintenance of preferred service for those customers who rank high because of their share in incomes and the profit of the production company. The order fulfillment algorithm refers to

products of FMCG (Fast Moving Consumer Goods) industry, and with minor modifications, it is possible to apply it to the problem of allocation of products in other industries too.

4 Algorithm

The problem of allocation of limited quantities of finished products to customer orders could be solved by creating an algorithm that systematizes the allocation process. The proposed allocation algorithm tends to maximize the customers service that is expressed by the number of fulfilled orders and the percentage of order fulfillment, all this through the processing of orders in accordance with previously defined groups and partitions of customers. The criterion of dividing customers into groups is based on the consideration of revenue per customer within the company, profit per customer, development potential, service rate per customer, strategic partnership with the customers, etc. With the goal of providing strategic potential for the development of a customer network, customers are classified into partitions that are provided with guaranteed quantity of products for certain groups of customers, which would otherwise be marginalized. The proposed algorithm has polynomial complexity.

Table 1: Variable descriptions

Variables	Description
PT_l	Total available production in l-iteration ($l = 1, ..., r$)
OT_l	Total order in l-iteration ($l = 1, ..., r$)
P_k	Partition ($k = 1, ..., p$)
G_j	Group ($j = 1, ..., m$)
C_i	Customer ($i = 1, ..., n$)

Algorithm is applicable only when $OT_l > PT_l$. In the beginning of the year, or at the beginning of iteration:

1. Form the list of customers with parameters ($i = 1, ..., n$)
The list comprises basic data on customers and their business indicators for the previous year, for the last three years, and last five years.

2. Define the number of groups ($j = 1, ..., m$)
Every identified group has its priority. Smaller ordinal number of a group brings higher priority. The recommendation is to form three groups, in spite of the fact that the algorithm functions even if only two groups are present.

3. Clustering customers into groups
At the beginning of the year, a company disposes of the list of customers, and it revises business results from the previous period for every respective customer, which is actually the base for classification into groups (Figure 1a). Every customer within the same group is equal with other members of the group when it comes to products' allocation. Clustering (classification) into groups could be done based on results of applying methods such as ABC method of prioritizing.

$$C_i \in G_j, \quad \text{for } i \in \{1, ..., n\} \text{ and } j \in \{1, ..., m\}$$

Information on which customer belongs to what group are kept in the matrix customer group (MCG), as shown on Figure 1b.

If a new customer emerges at the very beginning of new iteration, he is been added to the system and attached to some of the existing groups.

$$C_{n+1} \in G_j, \quad \text{for } j \in \{1, ..., m\}$$

4. Define the number of partitions with protective percentage quotes ($k = 1, ..., p$)
Partitions are being introduced with an aim to protect small customers with an opportunity to growth in the future. Granting a quote to the partition with groups containing small customers will provide a certain quantity of products for small customers who have great potential for development and also for the increase in quantities to be ordered in the future. The protective quote, i.e. the percentage portion of the overall production in l-iteration is defined for every partition, and it will be available to the observed partition.

$$0 \leq KP_k \leq 1, \text{ for } k \in \{1, ..., p\}, \quad \sum_{k=1}^{p} KP_k = 1$$

In addition, the number of partitions is smaller or equal to the number of groups. The concept of partition will lose sense (purpose) if only one partition has been defined.

5. Clustering groups into partitions
Every defined partition must have at least one group, even though such case counts as if every group has been awarded the protective quote (Figure 1a).

$$G_j \in P_k, \quad \text{for } j \in \{1, ..., m\} \text{ and } k \in \{1, ..., p\}$$

Information on which group belongs to what partition are kept in the matrix group partition (MGP), as shown on Figure 1c.

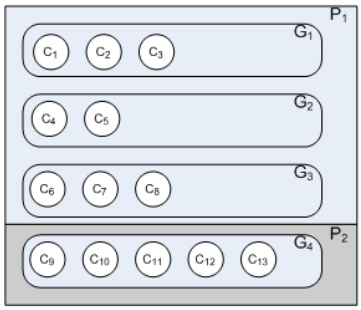


Figure 1a: Customers, groups, partitions

	G1	G2	G3	G4
C1	1	0	0	0
C2	1	0	0	0
C3	1	0	0	0
C4	0	1	0	0
C5	0	1	0	0
C6	0	0	1	0
C7	0	0	1	0
C8	0	0	1	0
C9	0	0	0	1
C10	0	0	0	1
C11	0	0	0	1
C12	0	0	0	1
C13	0	0	0	1

Figure 1b: Matrix customer group

	P1	P2
G1	1	0
G2	1	0
G3	1	0
G4	0	1

Figure 1c: Matrix group partition

If we mark the customer i demand with OC_i, then we could say that the total order for products in l-iteration is:

$$OT_l = \sum_{i=1}^{n} OC_i \qquad \text{or} \qquad OT_l = \sum_{j=1}^{m} \sum_{i=1}^{n} OC_i \cdot MCG_{ij}, \quad \text{for } l \in \{1, ..., r\}$$

Variable AP_k is then introduced with the goal to keep the amount of products allocated to k-partition in l-iteration:

$$AP_k = PT_l \cdot KP_k, \qquad PT_l = \sum_{k=1}^{p} AP_k, \quad k \in \{1, ..., p\} \text{ and } l \in \{1, ..., r\}$$

where PT_l is the total production in l-iteration.

Table 2: Variable descriptions

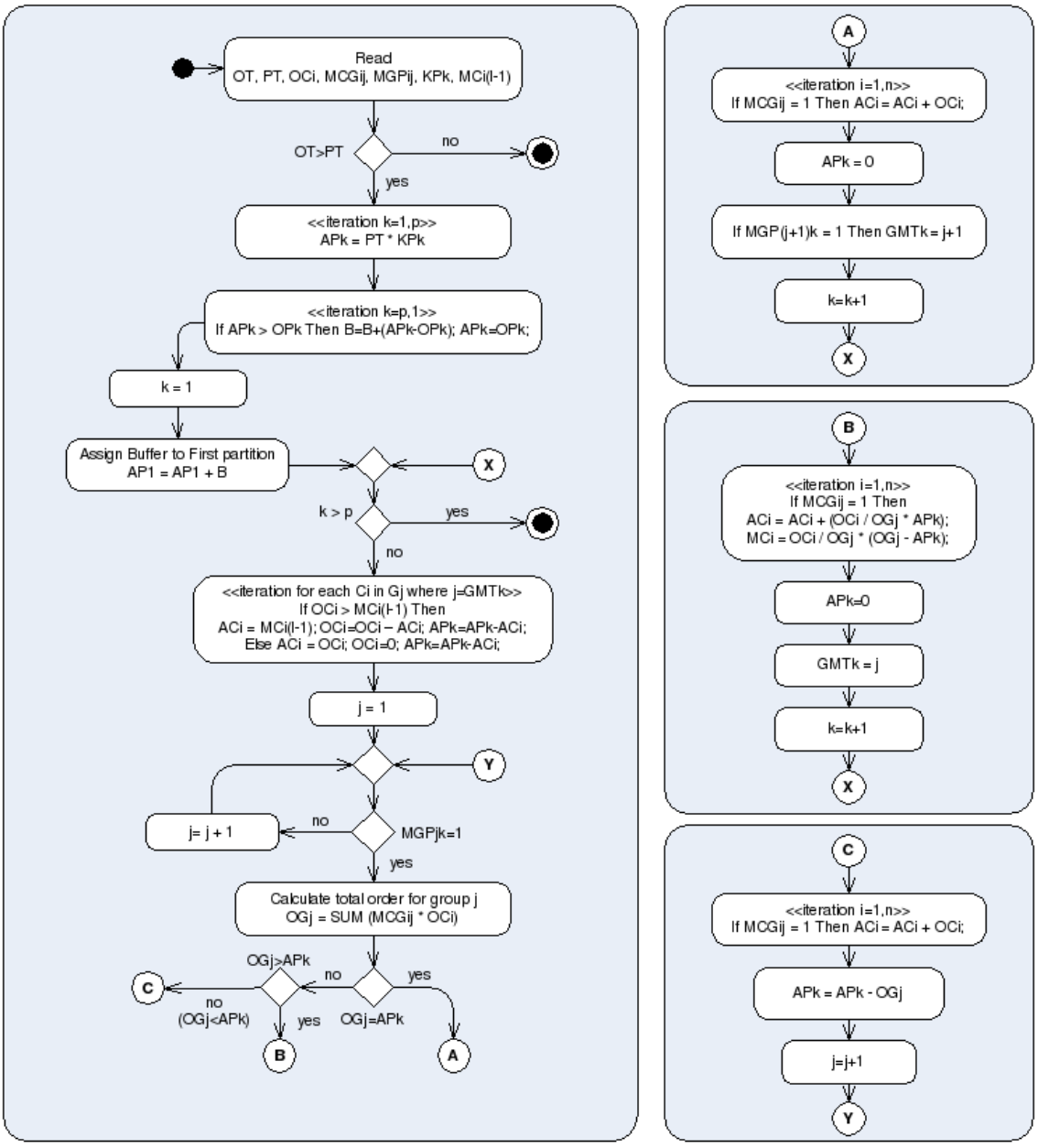


Figure 2: Algorithm represented as UML 2.0 Activity Diagram

Variables	Description
OC_i	Order of customer i in l-iteration
AC_i	Allocated quantity of customer i in l-iteration
$NC_j = \sum_{i=1}^{n} MCG_{ij}$	Number of customers in group j
$OG_j = \sum_{i=1}^{n} MCG_{ij} \cdot OC_i$	Total order of group j
$NG_k = \sum_{j=1}^{m} MGP_{jk}$	Number of groups in partition k
$OP_k = \sum_{j=1}^{m} \sum_{i=1}^{n} OC_i \cdot MCG_{ij} \cdot MGP_{jk}$	Total order of all customers in partition k
B_l	Buffer in l-iteration
MC_i	Memory - remembered quantity of unfulfilled part of the order of customer i in iteration l
GMT_k	Order number of group in partition k where token is set

The diagram on Figure 2 shows that the allocation algorithm in every iteration starts from checking whether $AP_k > OP_k$. Checking goes from the last partition to the first one, and if this condition has been met, then all allocated surplus is transferred to a joint buffer, whereas the allocated amount of finished products transferred to k-partition now equals the overall demand in given partition (first set $B_l = B_l + (AP_k - OP_k)$, and then assign $AP_k = OP_k$).This situation might happen in partitions that gather customers with small orders in case that higher quote than necessary is given in the beginning. This is a sign that one has to reaccess the analysis of partition of 'small' customers, and award smaller and more reasonable protective quote *(fine tuning)*. Within the algorithm, all amounts transferred to buffer are added to the allocated amount of finished products of the first partition ($AP_1 = AP_1 + B_l$).

One of the most important concepts of the proposed algorithm is group memory token (GMT). It is introduced in order to designate a group of customers whose order has not been fulfilled during an observed iteration. There could be only one group memory token (pointer) in each partition. In the end of one allocation iteration, token stays in a certain group with unfulfilled demand, which gives a priority to that group in the following iteration, so that the customers would get the unfulfilled order in the following iteration as a priority. In order to remember the unfulfilled order of group of customers with GMT, the variable MC (Memory of customer) is introduced. We need to stress that MC obtains the value for every customer from the groups

that kept token in the observed iteration, and it is used only during the allocation process in the following iteration.

Formula for calculating the unfulfilled part of an order is:

$$MC_i = \frac{OC_i}{OG_j} \cdot (OG_j - AP_k) \cdot MCG_{ij}$$

where MC_i is remembered quantity of unfulfilled part of an order of i-customer in l-iteration that is being transferred to the following iteration.

In case GMT in several iterations stays in the same group of the first partition of important customers, then the necessity of expanding capacities becomes clear.

5 Conclusion

The goal of this paper was to create an algorithm for solving the problem of allocation of limited stocks to incoming orders. Order fulfillment is carried out by letting the allocation algorithm grant stocks in line with the priority of customers, with the goal of providing high level of customer service. The algorithm proposed in this paper provides following: (1) classification in groups will provide the order of allocation with primarily focus to satisfy customers that are important for the company, in accordance with previously defined criteria, (2) application of partitions will help certain groups of lower priority within protected partitions to be involved in allocation so that the low-ranked customers would be at least partially satisfied, which keeps all the customers in the system and sends a useful signal to company's management saying that there is still unsatisfied demand, and they need to enlarge the production capacities, (3) introduction of GMT allows all customers within a group to be delivered the backorders from the previous cycle, with the extended delivery lead time, with which they improve the overall customer service.

The main differences between the allocation algorithm presented in this paper, and the models described in the related work are: (1) this allocation algorithm directly affects the long-term operating results by accomplishing the customer service as the primary goal, unlike other models that focus only on short-term profit generation, and (2) there is a tendency of bounding customers with different priority for a long-term period, i.e. the tendency of keeping the customers within the system. In addition, there are also differences in basic goals of models: the primary goal of advanced algorithm is the maximization of customer service, unlike other systems that are primarily orientated towards profit maximization.

Further research will be focused on testing and improving algorithm based on real data obtained in other companies and also from different industries. In addition, there are possibilities for further improvement of the allocation algorithm by creating a report system that would support management in making timely decisions on the change in operating strategy.

Bibliography

[1] Cederborg O., Rudberg M.: Customer Segmentation and Capable-to-promise in a Capacity Constrained Manufacturing Environment, 16th International Annual EurOMA Conference, Göteborg, Sweden, June 14-17, 2009.

[2] Chan F.T., Chung S.H.: A modified multi-criterion genetic algorithm for order fulfillment in manufacturing network, Proceedings of the 9th Asia Pacific Industrial Engineering & Management System Conference, APIEMS, Indonesia, 2008.

[3] Chen J.H., Lin J., Wu Y.S.: Order Promising Rolling Planning with ATP/CTP Reallocation Mechanism, IEMS, APIEMS, Vol. 7, No.1, 2008.

[4] Kaschel H., Bernal L.M.S.: Importance of Flexibility in Manufacturing Systems, International Journal of Computers, Communications & Control, Vol. I, No. 2, 2006.

[5] Lawrence G.: Introducing APS: Getting Production in Lock Step with Customer Demand, Automotive Manufacturing&Production, Vol. 110, Issue 5, 1998.

[6] Lin J., Chen J.H.: Enhance order promising with ATP allocation planning considering material and capacity constraints, Journal of the Chinese Institute of Industrial Engineers, Vol.22, No.4, 2005.

[7] Lupse V., Dzitac I., Dzitac S., Manolescu A., Manolescu M.-J., CRM Kernel-based Integrated Information System for a SME: An Object-oriented Design, *International Journal of Computers Communications and Control*, ISSN 1841-9836, Suppl. S, 3(S): 375-380, 2008.

[8] Makatsoris H.C., Chang Y.S., Richards H.D.: Design of a distributed order promising system and environment for a globally dispersed supply chain, Int. J. Computer Integrated Manufacturing, Vol. 17, No. 8, 2004.

[9] Meyr H.: Customer segmentation, allocation planning and order promising in make-to-stock production, OR Spectrum, Vol 31, No 1, 2009.

[10] Rudberg W., Wikner J.: Mass customization in terms of the customer order decoupling point, Production Planning & Control, Vol. 15, No. 4, 2004.

[11] Wikner J., Rudberg M.: Introducing a customer order decoupling zone in logistics decision-making, International Journal of Logistics: Research and Applications, Vol 8, No. 3, 2005.

Int. J. of Computers, Communications & Control, ISSN 1841-9836, E-ISSN 1841-9844
Vol. V (2010), No. 5, pp. 792-798

The Development of Students' Metacognitive Competences. A Case Study

D. Mara

Daniel Mara
"Lucian Blaga" University of Sibiu
E-mail: danielmara11@yahoo.com

Abstract: In the information society metacognitive competencies are essential. Based on some activities from the Enrichment Instrumental Program elaborated by professor Reuven Feuerstein we have designed a program for developing the students capacities of selfcontrol, selfknowing and intelectual learning strategies. The case study presents the formation of students' metacognitive competences at the "Lucian Blaga" University of Sibiu, "Hermann Oberth" Faculty of Enginereeing, Department of Computers Sciences. A Web based application has been developed in order to enable students to self-evaluate their metacognitive competencies and to acquire self-regulatory abilities.
Keywords: metacognitive competences, instrumental enrichment program, computer science, higher education, relational competences, motivation for didactic career, web based application.

1 Introduction

Metacognitive skills are a must for students preparing to have a career in the information/-knowledge society. For those that want to embrace a didactic career that is essential. Therefore the Department for Teaching Staff Training [8] has started a special training program consisting of two modules. The education plans for the first module consists of the following courses: psychology of education, pedagogy 1 (foundations of pedagogy, curriculum theory and methodology), pedagogy 2 (instruction theory and methodology, evaluation theory and methodology), specialty didactics, teaching practice, optional courses, final evaluation-didactic portfolio [9]. In the second module the following courses are included: curriculum area didactics, class management, counseling and vocational guidance, computer-assisted instruction, psychology of education, optional I (1 of 4: intercultural education, educational politics, contemporary pedagogical doctrines, management of school organization), optional II (1 of 4: psycho-pedagogy of adults, foundations of special psycho-pedagogy, sociology of education, research methodology in the sciences of education), final evaluation-project, teaching probation (42 hours - for those who did not teach during the period between the attendance of the first module and the enrollment at the second module).

2 Developing Students' Metacognitive Competences

Metacognitive skills development is an important formative intellectual object in education of the students, as reaching this level involves a route through effective education, appropriate to each one in particular [6]. Metacognitive skills suppose that students are aware of their own cognitive activity, i.e. learning activity, and self-adjustment mechanisms consisting in cognitive controls (rules, procedures, strategies).

2.1 The Background

The development of metacognitive skills goes in the same direction with the strategies used in developing cognition. The main steps in the formation, in the affirmation of conscience gripping meta-cognition are:

• affirmation of trust and intuition (AH Schoenfeld's model) in solving the problems, or tasks, based on knowledge, on previous experiences; in this step the trainer looks to identify in the student a sense of referral tasks, intuition, a way of understanding and finding solutions taking into account all possibilities;

• personal reflection on the knowledge involved; the student must become aware of the solutions found, the instruments used, her/his capacity of analysis and comparison, the way to analyse the difficulties of other methods previously used;

• self-awareness or awareness of effective solutions addresses the solving style, based on self-observation, analysis of results and of the ways to solve, progress and cognitive acting.

The development stages show that metacognitive skills are associated with knowledge from management, and construction and that these are the conditions in which the knowledge appears. Cognition managerial approach reveals the fact that metacognitive includes: awareness of how to understand the problem and how to solve it, planning the processes and finding the pathways necessary, monitoring the application solutions, the resources used, constraints, necessary instruments, decisions and analysis of results [4].

F.P. Büchel considers that training of metacognitive competences is more efficient if working in groups, in a climate of cooperation and confrontation, because there is the possibility of mutual evaluation. In self-training, the individual student is more concerned about solving the problem itself, bout the acquisition of knowledge and s/he is less concerned by the understanding of how knowledge is acquired, how solutions were found or decisions taken.

Researchers have built a hierarchical model of the criteria-assessment questions in the classroom climate in function of different elements: diversity of awareness, respect for others' style, commitment, encouragement, student-teacher relationship, student-group-class relationship, learning with pleasure, and sense of humour, comfortable participation and freedom of expression.

Studies that explore the effects of attitudes and emotions on learning indicate that stress and constant fear, at any age, can circumvent the brain's normal circuits. A person's physical and emotional well-being is closely linked to the ability to think and to learn effectively [1].

2.2 The Instrumental Enrichment Program

The Instrumental Enrichment Program is composed of a set of exercises divided into 14 tools that are used as means for developing mental capacities. The exercises do not concern the acquisition of specific knowledge, but the acquisition of mental skills, of ways to use concepts in different situations [2]. Each instrument is focused on specific cognitive functions and provides means for developing cognitive capacities necessary for solving tasks that require a high level of abstraction.

The Instrumental Enrichment Program components are:

- Organization Points;
- Spatial Orientation I.
- Comparisons;
- Analytical Perception;
- Pictures;
- Spatial Orientation II,
- Classification,
- Temporal Relations,

- Instructions,
- Family Relations,
- Numerical Progression;
- Syllogism;
- Transferable Relations;
- Outlines.

The exercises have images and temporal relations that are organized differently and provide a gradual increase in difficulty. In this way the student is encourage to a progressive acquisition of skills necessary to solve the problems or tasks, thus strengthening motivation, the feeling of competence and autonomy in organizing intrinsic work. Subjects become aware of the importance and need for discussion about the work done and to make transfers on the basis of principles/rules/patterns formulated during activity. The development of principles/rules/patterns and the implementation of transfers are very important elements. All the details of the page of an exercise must be caught and analyzed and a synthesizing valid principle must be identified and expressed in a concise sentence. A principle is important because it can highlight a complex problem, newly learned information, or a necessary element to solve the exercise.

The transfer is created as a link between the principles/patterns/rules resulting from the reflection necessary for understanding and addressing new events [3]. During an activity, two or more instruments are used in order to avoid monotony of using for a long period the same type of exercises, or the feeling of failure resulted from difficulties in solving an exercise. Students are lead to use different instruments and to learn to choose the right ones. An activity is made up of elements called "pages". A page contains a story, illustrated by images. Each instrument begins with a picture page (cover or homepage), which is used for placing the instrument, creating a horizon for motivation and development through the following pages.

Any learning (instructional) form may be tackled from the point of view of the general systems theory, distance learning forms included. A system is defined by a set of elements that interact and work together in order to achieve an objective [7, 8]. Cover pages have certain features that remain unchanged from instrument to instrument to highlight the continuity of work, but each instrument is different from others. The mediator/trainer oriented subjects to consider the symbol on the cover to deduce the exercises that will solve the issues and that they will discuss.

2.3 Organize An Activity Of The Instrumental Enrichment Program

An activity of the Instrumental Enrichment Program is organized respecting some rules and some key moments: the introduction, individual work, discussion and conclusions [5]: In the following I will briefly present the key moments and the rules to follow.

Introduction. By going through this phase the mediator wants to awake the group interest in the work that will be developed and to define the problems they will have to solve. The introduction begins with revision, i.e. data from previous lessons. The mediator shall ensure that requirements and concepts were well understood, and that the vocabulary necessary to solve the task is assimilated. Students will learn to analyze the page autonomously. The trainer guides students in observing and identifying objectives [10].

Individual Work. In this stage students will be asked to solve an individual task, after which they will be involved in a discussion aimed at highlighting possible strategies for solving the exercises from the page. Students must understand that it is important not to finish quickly the exercises of a page. Is important to understand how to solve a task and how they form and develop certain abilities. An activity based on reflection, even if not fully effective, may be more useful, more fruitful than the one done in a hurry, because it is based on a more deep analyze the processes that formed it.

Discussion. When most students have completed the individual task the trainer may start the discussion stage. Being particularly interested in mental processes that led to finding the solution, it is appropriate to insist on correct answers and to explain the wrong ones, to understand the mental processes through which solutions were found [12]. At first it is recommended that the mediator identifies the link between work and other applied situations, then the students will gradually create these connections between the instruments and the surrounding reality. Each transfer is built on a solid and appropriate explanation of the type of connection between the examples and the proposed developments.

Conclusion. At the end of each lesson there should be a revision of the whole activity. Even if it is short it should highlight the steps taken to achieve the objective, the new words acquired, targets and strategies set out above for achieving the aim of the lesson. It is possible to encourage valorisation activities to determine individually or in small groups the utility obtained by applying different tools.

3 The Case Study

Development activities of the metacognitive skills students were conducted by applying the tools instrumental enrichment program developed by Reuven Feuerstein a group of 75 students from the Faculty of Engineering "Hermann Oberth", Section Computer Sciences, of the University "Lucian Blaga"

Principles: *One event can't be observed by itself , it has to be seen in the whole context, before and after. We have to make a difference between opened eyes dreams and reality, between what is possible and impossible. We have to be aware about our goals, about their importance and about the risks they implied.*

The 75 students have bee enrolled in a trening program aimed to develop their metacognitive skills. One group (40) has worked only in the classroom and one group (35) has used also the web based application designed to support them in developing metacognitive comptences. The web based application is preparing the student for the training program, the students becoming familiar with the kind of exercises used in the Fuerstein program.

At the end of the training program the overall scores of the students that have used also the web based application was significant higher than the score of the group that worked only in the classroom. Gender has not a significant effect on students' perception of their metacognitive skills.

A rather uncomfortable conclusion is that more than 75% of the students (no gender significant differences) have difficulties in expressing in words their thoughts and experience. The group that has a pre-training with the Cogitino web based application was significant more rapid in solving different tasks, but has the same difficulties in expressing in words their thoughts and experience as the control group.

4 The Web Application

In order to help students to better understand their thoughts and experiences a support software - Cogitino - has been designed and implemented. Cogitino is a web based application (fig.1) that offers a set of resources concerning meta-cognitive skills and acts as an adviser for the student that is enrolled in the metacognitive skills development training program. Cogitino is a multi-agent system (fig.2) that through its Profiler agent determines the student's level of meta-cognitive competences and then recommend different training paths (fig.3).

The system is rating students' metacognitive competences based on the answers to several questionnaires and problems' results.

Before applying the Instrumental Enrichment Program students are asked to solve several problems similar to those that they will have to solve during the class.

Principle of e-learning applications have been observed [1, 7]

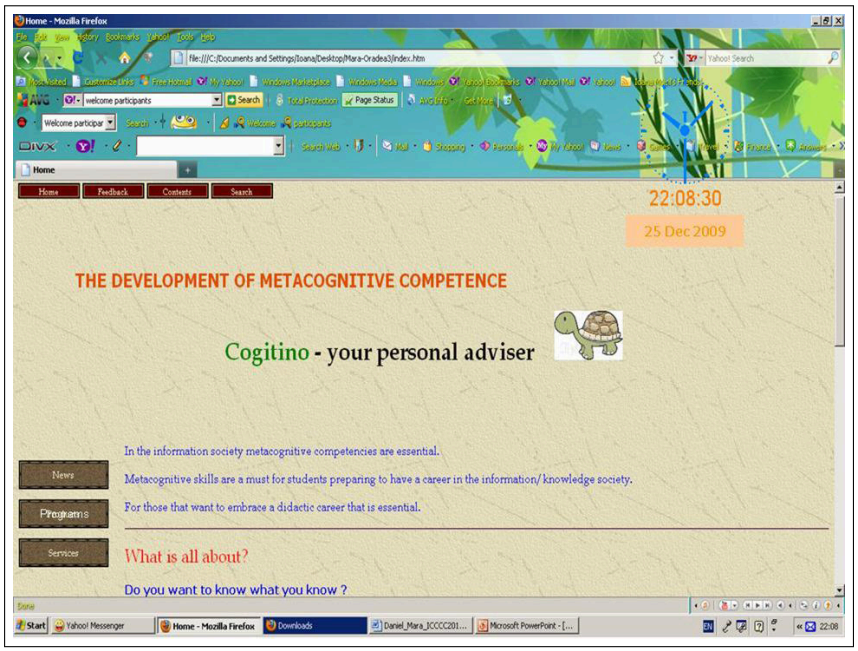


Figure 1: The Home page of the Cogitino application

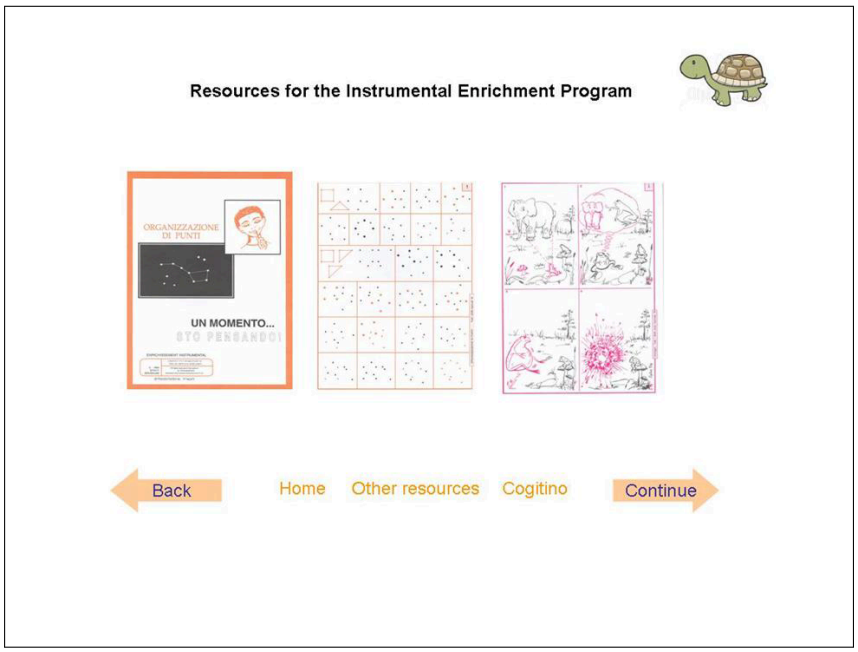


Figure 3: Cogitino - Training resources

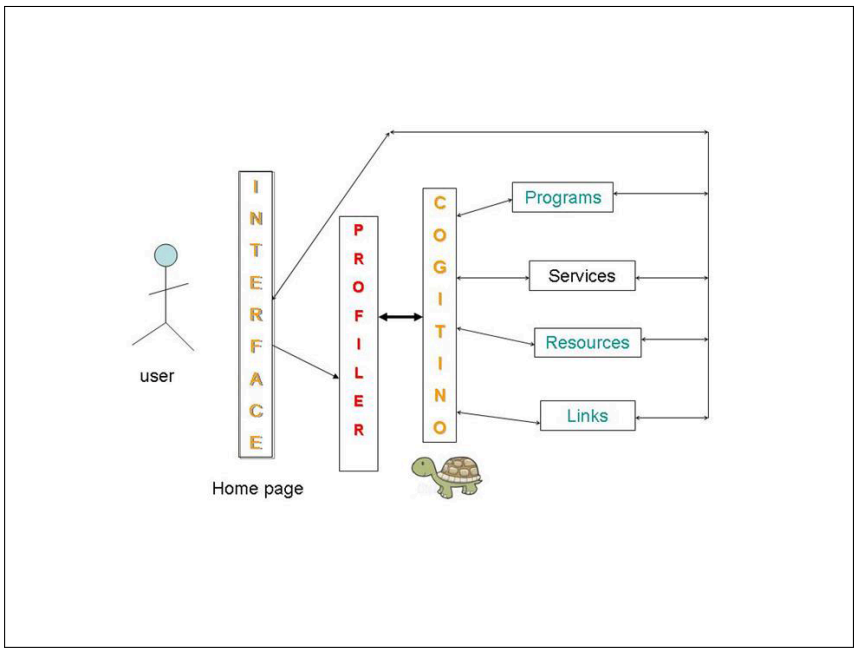


Figure 2: Cogitino - General structure

5 Conclusions

Metacognitive skills are mandatory for today students. They must be aware and must know their mental processes and they must be able to self-monitor, regulate, and direct their actions to their global aim. Metacognitive training becomes an important and basic tool also in business and management efficiency, skill and competences.

The result of the research carried out leads to evidence of at least three essential aspects in the development of students' meta-cognition competences. First, students balance their attention in preparation, implementation and evaluation of the educational and training process itself, but their qualitative analysis is poor. They have a reduce vocabulary and therefore a difficulty in explaining their experience and performance. Secondly, being enrolled in a technical program they feel, at least at the beginning of the training program, that they do not need to express themselves in words. And last but not least, the lack of general culture is an obstacle in understanding some of the tasks and problems they have been asked to solve. Meta-cognition components are usually observed only in the final stage of evaluation.

Another conclusion is that the web based application has been appreciated by students as very helpful. Considering this aspect and that in educational practice meta-cognition principles can be developed and applied efficient by students following a training program that included theoretical aspects and practical-application, my future work will consist in enriching (with the help of colleagues from the Computer Science Department of our university) Cogitino with two new modules : one module that will automatically generate explanations after a task has been solved, showing to the students how s/he has proceed, and another one that will be a "vocabulary training" for the student.

Acknowledgment
This work was supported by CNCSIS-UEFISCSU, project number 882/19/01/2009 PNII - IDEI, code 471/2008.

Bibliography

[1] Chen, Z., Learning about learners: system learning in virtual learning environment. International Journal of Computers, Communications & Control, Vol. III (2008), No. 1, pp. 33-40.

[2] Feuerstein R., Rand J., Rynders J.E., Non accettarmi come sono, Sansoni, Milano, Italy, (1995).

[3] Feuerstein, R., Rand, Y., Hoffman, M. B., Miller, R., Instrumental Enrichment: an intervention program for cognitive modifiability, Baltimore University Park Press, Baltimore, USA, (1980).

[4] Joiţa, E., Educaţia cognitivă. Fundamente. Metodologie, Editura Polirom, Iaşi, Romania, (2002).

[5] Kopciowski Camerini, J., L'apprendimento mediato. Orientamenti teorici ed esperienze pratiche del metodo Feuerstein, La Scuola, Brescia, Italy, (2002).

[6] Miclea, M., Psihologie cognitiv?. Modele teoretico-experimentale, Editura Polirom, Iaşi, România, (1999).

[7] Moise, G., A Formal Description of the Systemic Theory based e-Learning. International Journal of Computers, Communications & Control, Vol. III (2008), No. 1, pp. 90-102.

[8] Moisil, I., A model of the student behavior in a virtual educational environment, International Journal of Computers, Communications & Control, Vol. III (2008), Suppl. issue: Proceedings of ICCCC 2008, pp. 108-115

[9] Neculau, A., (coord.), (1996), Psihologie socială. Aspecte contemporane, Editura Polirom, Iaşi.

[10] Neculau, A., (coord.), (1997), Universitatea: valorile şi actorii săi, în "Câmpul universitar şi actorii săi", A. Neculau (coord.), Editura Polirom, Iaşi.

[11] Vanini, P., Potenziare la mente? Una scommessa possibile: L'apprendimento mediato secondo il metodo Feuerstein, Vannini Editrice, Gussago (Brescia), Italy, (2003).

[12] Vanini, P., Il metodo Feuerstein: una strada per lo sviluppo del pensiero, IRRSAE Emilia Romagna, Bologna, Italy, (2001).

Int. J. of Computers, Communications & Control, ISSN 1841-9836, E-ISSN 1841-9844
Vol. V (2010), No. 5, pp. 799-812

Discussion of the Analysis of Self-similar Teletraffic with Long-range Dependence (LRD) at the Network Layer Level

G. Millán, H. Kaschel, G. Lefranc

Ginno Millán, Héctor Kaschel
Universidad de Santiago de Chile
Departamento de Ingeniería Eléctrica
Avda. Libertador Bernardo O'Higgins #3363.
Estación Central. Santiago - Chile
E-mail: {ginno.millan,hector.kaschel}@usach.cl

Gastón Lefranc
Pontificia Universidad Católica de Valparaíso
Escuela de Ingeniería Eléctrica
Avda. Brasil #2147. Valparaíso - Chile
E-mail: glefranc@ucv.cl

Abstract: Traffic streams, sources as well as aggregated traffic flows, often exhibit long-range-dependent (LRD) properties. This paper presents the theoretical foundations to justify that the behavior of traffic in a high-speed computer network can be modeled from a self-similar perspective by limiting its scope of analysis to the network layer, since the most relevant properties of self-similar processes are consistent for use in the formulation of traffic models when performing this specific task.
Keywords: Long-range-dependent, network layer, network traffic, self-similar process.

1 Introduction

Still rooted in the genesis of the design of present day high speed computer networks is the trend to scalable development with a base prepared for the primary support of processing applications which, although requiring a reliable transport service, are not demanding in terms of other quality of service (QoS) parameters such as delay, flow rate, latency, and loss rate. It is a reality that is subordinated to financial justifications, unable to reflect both the behavior and the operation of present day network environments, most of them characterized not only by their scalability and support of services and added value applications with high band width and availability requirements, but also by their convergence, complementarity, and interoperability. On the other hand, sustained development in the fields of optical, nanometric, and quantum technologies with greater emphasis has allowed the evolution of computer networks, providing them with the capacity needed to satisfy simultaneously the requirements of diverse traffic, creating the scenario inherent to the appearance of new services and applications for which this characteristic is essential; we are in the presence of services that involve real time traffic and which, because of their nature, have highly demanding needs to and from the available bandwidth. Therefore, the new high speed networks must be capable of providing a service that not only ensures the availability of the resources, but is also provided under quality conditions that are well defined, parameterizable, adaptable, and dynamic in their assignment, because the requirements of present day real time applications and services cannot be satisfied using the high-level protocols if the carrier networks do not offer the necessary guaranties. It is crucial, therefore, to make quantitative analyses that evaluate the service quality offered by the new technologies, leaving

aside the unsubstantial bases, arguments, and assumptions. The main problem that appears at the time of making a rigorous evaluation of the performance of a communications network is that of modeling the input traffic to the network. In fact, to many authors, traffic modeling is the most critical problem related to the evaluation of the performance of communications networks, because the success of the analyses depends directly on how representative of reality are the traffic models used. Historically, traffic modeling had its origin in conventional telephony systems and has been based almost exclusively on assumptions of independence between the times of arrival of successive grids and exponential durations in the use of the resources. Concretely, the acceptance of both assumptions implies a restriction toward the stochastic processes so that they obey a universe of Poisson or Markov processes. In that respect, the usefulness of their use by network designers as well as by systems analysts for planning capacities and predicting performance is not questioned [1]. However, in a wide range of real world cases it has been verified that the results predicted from the analysis of tails differ significantly from the actual observed performance, and this marked discrepancy has its origin in the fact that traffic processes often present LRD on many or all the temporal scales, while Poisson or Markov models, which have no memory or show short-range dependence (SRD), present traffic flow over much shorter time scales. As a result, there is a tendency to produce highly optimistic forecasts of performance due to the use of distributions with finite variance for characterizing the periods with the presence and absence of burst packets. In view of the above arguments, the following working hypothesis is proposed: *"It is completely feasible to restrict the evolution of a statistically self-similar process to a well defined application setting without altering its nature and its more important properties, in that way highlighting the validity of its postulates and giving greater plausibility to its physical interpretation."*
In this context, the plausibility refers to the action of conferring an admissible character, therefore worthy of consideration, to one or several parameters that are components of an analytical model, whose interpretations do not constitute only a mathematical idealization. This paper presents a detailed discussion of the theoretical bases that justify the fact that the behavior of a high speed computer network traffic in the presence of long-range dependence can be modeled from a self-similar perspective limiting its analysis setting to the network layer level, stressing that all the most relevant properties of the self-similar processes are consistent for use in the formulation of traffic models when this distinction is made, since the need for its concept is justified to describe the traffic that is registered in the settings of present day computer networks.

2 Bibliographic Discussion

Since Kleinrock's publication [2], later expanded in [3], which establishes the mathematical theory that governs networks with packet switching, the existence of temporal dependence in the performance of the different types of data traffic flows has become an exciting field of research with countless discoveries, with the huge impact and influence that they have on the performance of tail systems standing out among them. The latter fact accounts for both the current existence and coexistence of a wide variety of input traffic models that show structures with rather complex correlations, which are applied to cases in which the models of the communications systems that are being studied allow adequate analytic handling. In any case, these models, basically Markovian, neglect the temporal correlations starting from a given temporal separation, even if the latter can be increased arbitrarily at the expense of complicating the models with additional parameters. Since 1993 an increasing number of published studies have documented that the data traffic pattern is well modeled by self-similar processes in a wide range of real world and network situations [1], with their top reference point and foundational work in the research of Leland, Willinger, Taqqu, and Wilson [4], presented originally at ACM SIGCOMM '93, and

then amended and extended in [5]. In spite of the existence of some previous papers that provide informal descriptions of this performance, such as [6]- [8], and to an exception from Mandelbrot himself [9], no one had submitted the idea of self-similarity applied to the analysis of data traffic in itself, and that paper shattered the illusion that a simple tail analysis, based on the assumption that the traffic follows a Poisson distribution can model adequately all the network traffic [1], showing that traffic in Ethernet has a self-similar or fractal nature and therefore requires new modeling and analysis statements. In this respect, the methodology followed by the authors involves a massive compilation of traffic samples from 1898 through 1992, from different Ethernet LANs of the research and engineering center of Bellcore in Morristown, USA [5], which resulted in a detailed temporal high resolution collection totaling more than 100 million packets with 10 μs precision, grouped in four sets of measures available in [10], and in the application of a rigorous and exhaustive statistical analysis based on the modeling of the traffic sources using hyperbolic tail distributions, in particular Pareto's, comparing the results with the behavior of the traffic flow of the real traces, and in the observation of the estimated value of Hurst's parameter (H) for each of the four sets of traffic samples, expressed for processing as a series of ordered pairs of data composed of the time of arrival and the size of the Ethernet packet, as well as for each level of temporal traffic aggregation considered.
A complete analysis of both the statement and the methodology followed by the authors is found in [11] and [12], and their proofs in [13]. Specifically, this research shows that:

- It is possible to model Ethernet traffic producing results similar to those of real Ethernet traffic using few parameters (parsimony), and with the fundamental added value of being physically plausible.
- Ethernet LAN traffic can be modeled through the superposition of many sources that vary between a state of burst transmission and one of inactivity, using for their characterizations infinite variance distributions. In particular it is proved using a Pareto distribution.
- Ethernet traffic is statistically self-similar regardless of the place and time at which it is checked.
- The degree of self-similarity measured in terms of Hurst's parameter (H) is a function of the use factor of Ethernet and can be used to get the magnitude of the traffic bursts.
- Traditional traffic models are unable to capture the property of self-similarity.

These results and their deep implications, as can be noted from the preceding paragraphs, produced a host of researchers seeking to observe that same behavior associated with the largest variety of communications and applications scenarios.
What follows is an exhaustive literature review of research results in relation to their field of application, that approaches the treatment of systems communications systems and applications from a self-similar perspective. It should be noted that the idea is not only to show the application of this view, but also to present results that dissent from it. The self-similar or fractal behavior of traffic in WAN networks is shown in [14]- [16], pointing out the failure of Poisson models to represent the strong correlations that exist at different temporal scales. Evidence and conclusions on this behavior in traffic due to the WWW are provided in [17]- [19], considering interconnection scenarios as well as traffic patterns in browsers. On the other hand, [20] and [21] point out the fractal nature of the data flow of the protocols that compose the signaling system 7 in common channel signaling networks [22], showing that the traditional methods are not adequate to interpret their behavior, and the duration of the calls are better characterized if hyperbolic tail distributions are used. In another setting, [9] and [23]- [25] show that the LRD

is a characteristic inherent to VBR video traffic that does not have any relation with the type of codec or the number of special effects that contain the recorded scenes. Specifically, the VBR video traffic flow transmitted through B-ISDN, ATM, and Internet networks are studied, showing that the behavior of the distribution's tail that represents the marginal band width can be described exactly if hyperbolic tail distributions (like Pareto) are used; that the self-correlation function of the video sequences decays hyperbolically (this is equivalent to long-range dependence) and it can be modeled using self-similar processes, and finally, that the use of models that only capture the SRD is inappropriate for characterizing this kind of traffic because it overestimates the performance, leading to insufficient resource assignments, which finally derives in poor perceptions by the users of the networks when difficulties appear to achieve the service quality expected by them. In the field of wireless communications, in [26] it is shown that traffic in CDPD networks has an LRD behavior. Using the R/S and Variance-Time methods, it estimates values for Hurst's parameters of $H = 0.8$ and $H = 0.9$, respectively, thus disregarding the use of predictive models based on Poisson arrival processes. In [27] an investigation is made of the impact of mobility on the added traffic in wireless networks in the city of Bristol, UK, and of whether the applications of voice and data together produce self-similar traffic. It is concluded that the added traffic generated by the mobile users that use voice and data services together show a self-similar LRD behavior that has no relation with the rate of penetration of the services. They warn on the drastic changes that wireless multimedia service implementations must undergo in terms of the traffic profiles used in their models to be able to capture that characteristic. From another perspective, in [28] it is shown that traffic in wireless networks with ad hoc topology is self-similar and forecastable as a consequence of the fact that subjacent temporal self-similar series are, in essence, predictable. The required data are captured using a wireless test network with ad hoc topology. The analysis of traffic and the design of wireless IP networks describing TCP traffic as dominant in present day Internet is approached in [29], indicating that its statistical nature shows the same behavior over all the temporal scales. It also presents an analysis of traffic traces which shows the statistically self-similar nature of the traffic due to WWW and to the VBR video over these types of networks. In [30] develops a new model for traffic in wireless networks that has its origin in the alternating fractal renewal processes (AFRP) proposed as traffic models in [14], and in the wide band network traffic model using the extended alternating fractal renewal processes (EAFRP) proposed in [31]. With the incorporation of a limiting rate for alternation between the two states, called extended limiting rate, the Rate-Limited EAFRP model is formulated, which assumes an advance with respect to the wireline models used traditionally for model traffic in wireless networks, since it takes care of its two main deficiencies: omission of the effects of the LRD temporal correlations, and inability to make reliable performance forecasts due to the high dependence on short-range processes. In [32] there is an extensive discussion of the problem of modeling the data traffic that flows from and to the wireless networks with respect to the Internet, taking care of large scale wireless communication structures. Based on the methodology presented in [33] and tested extensively in [34], [35], it is concluded that the circulation of traffic flows cannot be treated using Poisson models, and that their behavior is statistically self-similar. In [36] two solutions are proposed to create an interconnecting bridge between WiMAX and WiFi links. The former is based on maintaining a certain level of QoS from one end to the other regardless of the wireless technology used, while the latter is aimed at the reduction of the complexity in its physical implementation at the expense of not providing any QoS guaranty. In both cases the performance of the system is compared with computer simulations that consider real time traffic with long-range dependence that is manifested trough a polynomial type decay of the self-correlation function. To model the traffic generated by the many terminals within the WLAN, recourse is made to the On/Off methodology presented in [14], supporting its foundations on [37]. Finally, in the same context,

but in another field of application, [38] deals with the use of fractal geometry in the antenna design process, while in [39] a general synthesis methodology is proposed that covers the efficient design of fractal antennas and their WiMAX applications. In the field of optical networks, [40] deals extensively with the use of self-similar traffic models, in particular that proposed in [15], as the only way to represent reality faithfully and to evaluate the performance of Ethernet passive optical networks. In [41] a new protocol is proposed for labeled switching in optical networks with optical burst switching (OBS), with time/space labeling that allows to keep the signaling always joined with the addressing functions. The results obtained are based on simulations under the premise of accepting the self-similarity of the traffic in the networks with wavelength division multiplexing. Finally, [42] reports on the implications of reducing the self-similarity of IP traffic by the OBS assembly algorithms. In spite of what all the above arguments imply in terms of putting in evidence the merits and advantages of the use of parsimonious models and which also provide a plausible physical interpretation to its parameters, the question arises on the degree of prevalence of these self-similar traffic patterns and on which are the conditions for the analysis of performance to depend critically on considering self-similarity. In this sense it is valid to inquire not only about the origin of the data that have been analyzed with respect to the synthetic traces generated, but also on what is the context or the setting in which these comparisons have been made, and where are the results aimed at. It is no less true that in the light of all the research presented, it seems even irrelevant to think of a traditional analysis of tails to represent data traffic flow in present day high speed networks. But this is neither categorical nor restrictive; this methodology cannot be disregarded flatly, and neither is its proper to think that the former is the unified solution by simply arguing it ubiquitousness in all the temporal scales, because this, analyzed from a higher perspective, and even though the presence of the correlation in the traffic is not under discussion, brings about the question of whether the correlation structure alone is sufficient to characterize traffic over self-similar processes. The context described in this way is very extensive, and therefore the results must be limited.

There are several reports that put in evidence the lack of consensus on the application field of self-similar models and of the impact of LRD on the performance of communications systems, and even though their number is much more limited than that of those whose results support exactly the opposite, their conclusions must be analyzed carefully because they bring forth, in essence, a critical and fundamental matter in common: "since the traditional tail models are unable to put in evidence the self-similarity characteristic, its validity for predicting performance would be supported if it is shown that self-similarity does not have a measurable impact on performance," and even more so if it is shown that the models based on self-similar stochastic processes fail when considering the impact of important characteristic parameters in each particular network case. Precisely based on this last point, [43] presents a detailed analysis of a fault detected in self-similar models by considering that they are incapable of reflecting the impact of the range of temporal scales of interest for evaluating the performance and the prediction of problems, and those first order statistics such as the marginal distribution of the process. Based on traces generated by a JPEG encoder of an NTSC television channel and the traces of [6] it is reported that:

- There is a correlation horizon such that the rate of loss does not affect performance beyond it.
- The correlation level considered for evaluating performance depends not only on the structure of the source traffic correlation, but also on the temporal scales belonging to the system that is being studied.
- The scale factor considered has a considerably greater impact on the rate of loss than Hurst's parameter or the size of the buffers.

- Increasing the size of the buffers helps reduce the rate of loss if we are dealing with SRD traffic. On LRD traffic it does not have a considerable impact.

The in-depth analysis of this work shows that the impact of the LRD is dealt with over the contribution of the single server model in relation to the size of the input buffer, using for that purpose a model of fluid [44], [45], that presents a hyperbolic drop down to a given cut coefficient from which it drops to zero. Based on the results from many simulation experiments using video traces as well as Ethernet traces for different values of Hurst's parameter, cut coefficients and buffer sizes, and a wide range of marginal distributions, the authors discovered the existence of a critical cut coefficient that they call "correlation horizon," such that the rate of loss is not affected if the cut coefficient is increased above it. Therefore, the correlation horizon separates the relevant from the irrelevant correlation coefficients with respect to the rate of loss. Finally, in their conclusions the authors argue that because of the existence of a finite cut horizon (which in the case of the finite buffers is a function of their size), any model that captures the correlation structure up to that horizon will be valid to represent the system. On the contrary, if the correlation horizon is infinite, i.e., the system's time scale cannot be determined clearly, then self-similar models must be used. Based on the above point, [46] states that in the case of considering a finite buffer, the effects of the LRD are detectable only if it makes the occupation periods become sufficiently long, since the behavior of their tail is affected largely by the characteristics of the traffic that arrives during those periods. In that respect, this appraisal is based on [47], where through the definition of the concept of "relevant temporal scale" as the typical duration at which all the arrivals at a tail interact and affect collectively their behavior, it is deduced that for the large buffers the size of the tail can be large, so many arrivals interact in the tail and cause the long-range correlation to cause more losses than those predicted by the models that are not capable of considering it or do not consider it. However, in the case of small buffers, where few arrivals interact, the effect of the LRD is imperceptible. Therefore, the standpoint on the appraisal of the effects of the LRD in terms of its impact on the occupied periods is depicted by the authors through the concept of "reset effect," which involves that as the buffer in question is emptied, the system forgets everything. In this way, in the case of VBR video servers, since they are sensitive to the delay and to the loss of meshes, the intensity of the traffic flow will not be very large, giving rise to short periods of occupation and a very pronounced reset effect in the regions of practical operation, an effect that will also be reinforced in the case of finite buffers, due to the truncating effect of these types of buffers. The latter is due to the fact that an occupied period in which there is overflow is shorter than the corresponding one in an infinite buffer model, or similarly, there can be several periods occupied in the finite buffer version before the corresponding occupation period ends in the infinite buffer version. Then, through the use of Markov models, the authors finally report that when the SRD is strong and parameter H is moderate, the LRD has no impact on the occupation of the buffer, and therefore their models give rise to good estimations, and in the case in which the SRD shows a slightly pronounced behavior and parameter H has a high value, the truncating effect is sufficiently strong for their models to estimate well the rate of loss, even though it is admitted that for a high traffic intensity and a large buffer size the estimation of the mean size of the tail is bad. In [48], a research having characteristics similar to those of the previous one, the concept of critical temporal scale (CTS) is introduced as follows. Given the size of the buffer and the marginal distribution of mesh sizes, the CTS of a VBR video source is defined as the number of mesh correlations that contribute effectively to the rate of loss of cells. Using models of the video traces of [9], the authors state that for Markovian models as well as with LRD, the CTS is finite and decreases with the size of the buffer. Consequently, and under the assumption that the size of the buffer required to multiplex a large number of VBR video source is typically small due to the restrictions that correspond to real time applications, it is concluded that for buffer sizing scenarios in ATM

networks it is not necessary to capture the correlations with the LRD of a video source even if the traffic shows this behavior markedly, because for all practical cases the SRD has a dominant impact. The following can be argued with respect to this last conclusion and to the research in general:

- The result is based on an analysis that relies completely on marginal Gaussian distributions, and even though it can be explained from the perspective of keeping an adequate analytic treatment, it is not sufficient for validating absolute generalizations of the *"in all cases"* type nor to attenuate the fact of working with distributions independent of the behavior of others.
- The behavior of the models if hyperbolic tail distributions are used or if they are tested with distributions without margins is not reported.

An extended version of this research is [49], where not only these two considerations are taken up, but attention is also given to the relevance of the SRD and the LRD in real time VBR video traffic in wide band networks (particularly ATM) and in the integrated Internet services, and through a theoretical and simulations approach, the authors tackle the problem of determining admissible ranges for the probability of cell loss and its relation with the size of the buffers in terms of the maximum delay. It should be noted that this research takes place within the context of validating the use of Markovian models and SRD for video applications, and for that purpose the authors show basically that:

- The long-range correlations have no impact on the probability of cell loss.
- An adequately implemented Markov model that captures the relevant range of correlation provides good predictions of performance.
- The capture of the LRD by itself can lead to an underestimation of the necessary resources in the network.
- The CTS explains the strong relation existing between the probability of cell loss in ATM and the SRD at the expense of the LRD. That is, for the applications of interest, the CTS is small and more sensitive to the short-range than to the long-range traffic correlations.
- Analytically simple Markovian models are feasible and have the capacity to harmonize marginal distributions and correlations over a given critical temporal scale.

As a last example of the already mentioned lack of consensus in the field of video traffic, [61] presents a model of a VBR traffic source that uses finite states Markov chains, and states that although the original model presented in [26] is good in terms of its parsimony, it does not lend itself adequately to analytic studies. Summarizing, six reports have been presented so far with a common denominator: dissenting on the blind applicability of the self-similar model and its implications in performance, considering as examples systems that involve video traffic, mainly of the VBR type. But as expected, this is not the only field in which discrepancies appear, and more acutely and in direct relation with the working hypothesis stated there is a number of reports that question the validity of the use of Hurst's parameter as a single descriptor to characterize the LRD of a self-similar stochastic process. But before dealing with this topic a last group of reports are mentioned that have a critical position toward self-similar models in areas other than video traffic. In the field of wireless networks and their associated technologies, [51] studies the behavior of the self-similarity characteristic of traffic when it goes from a wired to a wireless network through a gateway, concluding that the device can change the traffic's degree of self-similarity as a direct consequence of the reassembly and repacking operations on the self-similar

input traffic, even reaching its annulment. In [52] the above behavior is reaffirmed considering the study of the influence of the MAC mechanism of IEEE 802.3e on LRD traffic when it goes through one or several links, suggesting that the traffic transported through a WLAN interface undergoes deep structural changes in its statistical model, and showing that the fractional Gaussian traffic model is inadequate to describe its behavior. Finally, in the field of OBS networks, [53] reports the development of an algorithm for the assembly of bursts that has the purpose of reducing the degree of self-similarity of IP traffic. It is admitted that this is a characteristic inherent to WWW traffic, but its presence causes an important disadvantage in terms of the performance of the tails, so it must be reduced in favor of a random SRD traffic. The first paragraph of this page mentions the existence of some research that questions the use of Hurst's parameter as a single indicator to capture the self-similarity characteristic of a stochastic process that boasts of being such, in addition to stressing the importance that this fact has to prove the stated work hypothesis. In this respect, and recalling its statement, it is specified that the idea is not to validate exhaustively the self-similarity parameter or Hurst's parameter, but only to obtain an indicator of the presence of the characteristic in representative traffic series. Consider the following definition: The valuated real process $X(t), t \in \mathbb{R}$ is self-similar with $H > 0$ if for all $a > 0$ the finite-dimensional distributions of $X(at), t \in \mathbb{R}$ are identical to the finite-dimensional distributions $a^H X(t)$, $t \in \mathbb{R}$, i.e.,

$$\{X(at), t \in \mathbb{R}\} =_d \{a^H X(t), t \in \mathbb{R}\} \qquad \forall a > 0 \tag{2.1}$$

where $=_d$ means equality for all the finite-dimensional distributions [54], [55].
The property defined by (1), is usually known as the scaling property, and a direct consequence of its definition is that a self-similar process preserves its distribution, and thereby its statistics, since it is subjected to a temporal scaling. Also, from the same standpoint parameter H, or Hurst's, is known as the self-similarity parameter for the stochastic process $X(t)$ to which it is associated. The first report that recognized and approached the need to have additional parameters to characterize the variability of the traffic is [26], where although it is accepted that H is necessary for that purpose, it is not sufficient. Through a detailed statistical analysis of VBR video samples, the authors conclude that the self-correlation of the VBR video sequences decays hyperbolically, equivalent to the LRD. But since the LRD is related to the frequency of the components of the process and not to the distribution of the bandwidth requirements, if the marginal distribution is compressed because the coefficient of variability (coefficient between the mean bandwidth and the standard deviation) tends to zero when the number of multiplexed input sources that give rise to the traffic tend to infinity, the traffic, as the number of sources increases, is confined within narrow statistical limits, and although within these frontiers the behavior continues to be long range (result confirmed through $H = 0.7$), in the range in which the standard deviation is much less than the product of the mean of the bandwidth distribution and the number of multiplexed input sources, the traffic does not depend on H. Therefore, H is necessary to characterize the variability, but it is not sufficient. So an adequate characterization of video traffic must consider at least the following four parameters: the mean of the bandwidth distribution, the number of multiplexed input sources, the standard deviation, and the coefficient of variability derived from them. The authors make it clear, finally, that these results are valid only if they follow the central limit theorem [56], i.e., when the standard deviation is finite. Although the above result may seem to be expected because the parsimony can lead to imprecision in both the interpretations and the results, it is not so in relation to the effect produced by assuming that if the detailed behavior of the components of a given stochastic process, $P1$, that shows some degree of self-similarity $H1$ is not known, then a biunivocal correspondence between $H1$ and $P1$ is clearly established. In other words, processes that show clearly differentiated behaviors are possible, but their correlation structures must be characterized by the same H parameter. In this

respect, [57] approaches this problem considering the asymptotic behavior of an unlimited buffer of a multiplexor under different self-similar input traffic models, in particular, Cox's infinite server models, or $M/G/\infty$ [58], and fractional Gaussian noise models [59], [60]. In this way the authors report getting two completely different behaviors for the buffer's probability tail, namely that while with the former the drop is mostly hyperbolic, with the latter it presents a Wiebull asymptotic behavior [61], simply showing that Hurst's parameter is insufficient as a single descriptor to characterize the LRD in input traffic models [62], [63]. In [64] it is also shown how synthetic traces that have identical self-similarity parameters and means differ significantly from one another. Finally, in [65] the insufficiency of Hurst's parameter by itself as a precise descriptor of the long-range dependence of the traffic in an Ethernet network is reported. In this research it is shown convincingly, through an analysis of tails applied to a series of representative data of the traces of the real traffic registered in the Ethernet network of the Department of Computer Science of the University of California at Los Angeles, CA, USA, that Hurst's parameter does not provide a precise prediction of the performance of the tails for a given LRD traffic, and that its behavior is not monotonic with respect to the presence or absence of bursts if the original series is disaggregated into smaller ones, which also implies that H does not serve to characterize the relative importance of groups within a whole. It is clearly seen that both results are opposed to the conventionalism of [6] since H cannot be used to size the traffic bursts in Ethernet. Questions on where, when, and under what set of circumstances the use of self-similar processes in modeling communications systems and related applications is completely valid, as well as the uncertainty on the existence of a scenario that brings together criteria and its influence on the basic characteristics of the processes remain unanswered. However, they seem to have an answer derived from the analysis of the research of Ryu and Lowen, [66], [67], on the use of fractal point processes (FPP) for modeling and analyzing self-similar traffic in networks [68]. Concretely, this research proposes to make a distinction between self-similarity at the application level and self-similarity at the network level for the purpose of the design and administration of wideband networks in terms of a correct provision of the QoS required by the applications, and their best results are the proofs that self-similarity of VBR traffic can frequently be ignored in the face of the sizing of the buffers in ATM, and that self-similar traffic at the applications level can be managed effectively in the context of the admission control for assigning resources with service quality guarantees, because it is independent of the network conditions under which it is sent.

3 Conclusions

A detailed discussion has been presented of the theoretical bases that support the position of considering that high speed computer network traffic shows self-similar behavior and long-range dependence. With respect to the above, and considering the arguments in favor and against this position, it is believed that traffic in present day computer network settings is of a statistically self-similar nature and presents a pronounced long range dependence (LRD). Accepting the above singularities as inherent to the traffic flow of present day high speed network settings, it is proposed that their behaviors are amenable to being modeled by limiting their applicability to the network layer level, estimating that the most relevant properties of self-similar stochastic processes are consistent to be used in the formulation of traffic models when that distinction is made, since the concepts of self-similarity and long-range dependence are justified by the need to describe faithfully the real traffic processes in present day computer network settings. To depict the concepts with which this research deals, the following work hypothesis is proposed:

"It is completely feasible to restrict the evolution of a statistically self-similar process to a well defined application setting without altering its nature and its more important properties, in that

way highlighting the validity of its postulates and giving greater plausibility to its physical interpretation."

Considering all the arguments given above, it is stated that it is theoretically feasible to prove it, since all its arguments are correctly founded and supported, and only their proof at the analytic and experimental levels remains as a future task.

Bibliography

[1] W. Stallings, *Redes e Internet de Alta Velocidad. Rendimiento y Calidad de Servicio,* 2nd ed., Madrid, Pearson - Prentice Hall, 2004, pp. 224-237.

[2] Kleinrock, *Information Flow in Large Communication Nets,* Ph.D. Thesis, Research Laboratory of Electronics, Massachusetts Institute of Technology, Cambridge, MA, 1961.

[3] L. Kleinrock, *Communication Nets: Stochastic Message Flow and Delay,* New York, McGraw-Hill, 1964.

[4] W.E. Leland, M.S. Taqqu, W. Willinger, and D.V. Wilson, *On the Self-Similar Nature of Ethernet Traffic*, in Proc. ACM SIGCOMM ?93, San Francisco, CA, pp. 183-193.

[5] W.E. Leland, M.S. Taqqu, W. Willinger, and D.V. Wilson, *On the Self-Similar Nature of Ethernet Traffic (Extended Version),* IEEE/ACM Trans. Netw., Vol. 2, No. 1, pp. 1-15, February 1994.

[6] A. Erramilli, R.P. Singh, and P. Pruthi, *Application of Deterministic Chaotic Maps to Model Packet Traffic in Broadband Networks,* in Proc. 7th ITC Specialist Seminar, Morristown, NJ, 1990.

[7] W.E. Leland and D.V. Wilson, *High Time-Resolution Measurement and Analysis of LAN Traffic: Implications for LAN Interconnections,* in Proc. IEEE INFOCOM ?91, Bal Harbour, FL, pp. 1360-1361.

[8] J. Beran, R. Sherman, M.S. Taqqu, and W. Willinger, *Long-Range Dependence in Variable-Bit-Rate Video Traffic,* IEEE Trans. Commun., Vol. 43, No 2/3/4, pp. 1566-1579, Feb/-Mar/Apr 1995.

[9] B. Mandelbrot, *Self-Similar Error Cluster in Communication Systems and the Concept of Conditional Stationarity*, IEEE Trans. Commun. Technol., Vol. 13, No. 1, pp. 71-90, Mar. 1965.

[10] http://ita.ee.lbl.gov/html/contrib/BC.html.

[11] W. Willinger, M.S. Taqqu, R. Sherman, and D.V. Wilson, *Self-Similarity Through High-Variability: Statistical Analysis of Ethernet LAN Traffic at the Source Level,* in Proc. ACM SIGCOMM '95, Cambridge, MA, pp. 100-113.

[12] W. Willinger, M.S. Taqqu, R. Sherman, and D.V. Wilson, *Self-Similarity Through High-Variability: Statistical Analysis of Ethernet LAN Traffic at the Source Level,* IEEE/ACM Trans. Netw., Vol. 5, No. 1, pp. 71-86, Feb. 1997.

[13] M.S. Taqqu, W. Willinger, and R. Sherman, *Proof of a Fundamental Result in Self-Similar Traffic Modeling,* ACM SIGCOMM Computer Communication Review, Vol. 27, No. 2, pp. 5-23, Apr. 1997.

[14] A. Adas and A. Mukherjee. (1994, Dec.).*On Resource Management and QoS Guarantees for Long Range Dependent Traffic.* Georgia Inst. Tech., GA. [Online]. Available: http://hdl.handle.net/1853/6797.

[15] S.M. Klivansky and A. Mukherjee. (1995, Aug.). The NFSNET. Georgia Inst. Tech., GA. [Online]. Available: ftp://ftp.cc.gatech.edu/pub/coc/tech_reports/95/GIT-CC-95-07.ps.Z.

[16] V. Paxson and S. Floyd, *Wide-Area Traffic: The Failure of Poisson Modeling,* IEEE/ACM Trans. Netw., Vol. 3, No. 3, pp. 226-244, Jun. 1995.

[17] M.E. Crovella and A. Bestavros. (1995, Oct.). Explaining World Wide Web Traffic Self-Similarity. Boston Univ., MA. [Online]. Available: http://www.cs.bu.edu/techreports.

[18] M.E. Crovella and A. Bestavros, *Self-Similarity in World Wide Web Traffic: Evidence and Possible Causes,* IEEE/ACM Trans. Netw., Vol. 5, No. 6, pp. 835-846. Dec. 1997.

[19] M. Arlitt, R. Friedrich, and T. Jin, *Workload Characterization of a Web Proxy in a Cable Model Environments,* Performance Evaluation Review, Vol. 27, No. 2, pp. 25-36, Sep. 1999.

[20] D.E. Duffy, A.A. Mc Intosh, M. Rosenstein, and W. Willinger, *Statistical Analysis of CCSN/SS7 Traffic Data from Working CCS Subnetworks,* IEEE J. Sel. Areas Commun., Vol. 12, No. 3, pp. 544-551, Apr. 1994.

[21] P. Pruthi and A. Erramilli, *Heavy-Tailed ON/OFF Source Behavior and Self-Similar Traffic,* in Proc. 1995 IEEE International Conference on Communications, Seattle, WA, Vol. 1, pp. 445-450.

[22] G. Rufa, *Developments in Telecommunications. Whit a Focus on SS7 Network Reliability,* Berlin, Germany: Springer-Verlag, 2008.

[23] M.W. Garrett and W. Willinger, *Analysis, Modeling and Generation of Self-Similar VBR Video Traffic,* Computer Communication Review, Vol. 24, No. 4, pp. 269-280, Oct. 1994.

[24] B. Tsybakov and N.D. Georganas, *On Self-Similar Traffic in ATM Queues: Definitions, Overflow Probability Bound and Cell Delay Distribution,* IEEE/ACM Trans. Netw., Vol. 5, No. 3, pp. 397-409, Jun. 1997.

[25] L. Yellanki, *Performance Evaluation of VBR Video Traffic Models,* M.Sc. Thesis, Dept. Comput. Sci., Univ. Saskatchewan, Saskatoon, SK, Canada, 1999.

[26] M. Zhonghua, *Analysis of Wireless Data Network Traffic,* M.Sc. Thesis, School of Engineering Science, Simon Fraser Univ., Burnaby, BC, Canada, 2000.

[27] D.R. Basgeet, J. Irvine, A. Munro, P. Dugenie, D. Kaleshi, and O. Lazaro, *Impact of Mobility on Aggregate Traffic in Mobile Multimedia System,* in the 5th International Symposium on Wireless Personal Multimedia Communications, Honolulu, HI, 2002, Vol. 2, pp. 333-337.

[28] Q. Liang, *Ad Hoc Wireless Network Traffic-Self-Similar and Forecasting,* IEEE Commun. Lett., Vol. 6, No. 7, pp. 297-299, Jul. 2002.

[29] T. Janevski, *Characterization and classification of IP traffic, in Traffic Analysis and Design of Wireless IP Networks,* Norwood, MA: Artech House, Inc., 2002, ch. 5, pp. 135-165.

[30] J. Yu, *Modeling of High-Speed Wireline and Wireless Network Traffic,* Ph.D. Dissertation, Elect. Comput. Eng. Dept., Drexel Univ., Philadelphia, PA, 2005.

[31] X. Yang, *Impulsive Self-Similar Processes, with Applications in Broadband Communication System Modeling,* Ph.D. Dissertation, Elect. Comput. Eng. Dept., Drexel Univ., Philadelphia, PA, 2001.

[32] J. Ridoux, A. Nucci, and D. Veitch, *Seeing the Difference in IP Traffic: Wireless Versus Wireline,* in Proc. 25th IEEE International Conference on Computer Communications, Barcelona, Spain, 2006, pp. 1-12.

[33] N. Hohn, D. Veitch, and P. Abry, *Does Fractal Scaling at the IP Level Depend on TCP Flow Arrival Processes?,* in Proc. 2nd ACM SIGCOMM Workshop on Internet Measurement, Marseille, 2002, pp. 63-68.

[34] N. Hohn, D. Veitch, and P. Abry, *Cluster Processes, a Natural Language for Network Traffic,* IEEE Trans. Signal Process., Vol. 51, No. 8, pp. 2229-2244, Aug. 2003.

[35] N. Hohn, D. Veitch, and P. Abry, *The Impact of the Flow Arrival Process in Internet Traffic*, in Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing, Hong Kong, 2003, Vol. 6, pp. VI-37-40.

[36] R. Fantacci and D. Tarchi, *Bridging Solutions for a Heterogeneous WiMAX-WiFi Scenario,* Journal of Communications and Networks, Vol. 8, No. 4, pp. 369-377, Dec. 2006.

[37] A. Erramilli, M. Roughan, D. Veitch, and W. Willinger, *Self-Similar Traffic and Network Dynamics,* in Proc. of the IEEE, Vol. 90, No. 5, pp. 800-819, May. 2002.

[38] J.P. Gianvittorio and Y. Rahmat-Samil, *Fractal Antennas: A Novel Antenna Miniaturization Technique, and Applications,* IEEE Antennas Propagat. Mag., Vol. 44, No. 1, pp. 20-36, Feb. 2002.

[39] R. Azaro, E. Zeni, M. Donelli, and A. Massa, *Fractal-based methodologies for WiMAX antenna synthesis,* in WiMAX: Technologies, Performance Analysis, and QoS, S. Ahson and M. Ilyas, Eds. Boca Raton, FL: CRC Press, 2008, ch. 2, pp. 21-39.

[40] G. Kramer, *Ethernet Passive Optical Networks,* USA, McGraw-Hill, 2005.

[41] A. Huang, B. Mukherjee, L. Xie, and Z. Li, *Time-Space Label Switching Protocol (TSL-SP),* in High-Performance Packet Switching Architectures, I. Elhanany and M. Hamdi, Eds., Germany: Springer-Verlag, 2007, ch. 9, pp. 197-210.

[42] M. Maier, *Optical Switching Networks,* New York: Cambridge University Press, 2008.

[43] M. Grossglauser and J-C Bolot, *On the Relevance of Long-Range Dependence in Network Traffic,* IEEE/ACM Trans. Netw., vol. 7, no. 5, pp. 629-640, Oct. 1999.

[44] J.-Y. Le Boudec and P. Thiran, *Network Calculus. A Theory of Deterministic Queuing Systems for the Internet,* Germany: Springer-Verlag, 2004, pp. 3-6.

[45] A. Adas, *Traffic Models in Broadband Network,* IEEE Commun. Mag., Vol. 35, No. 7, pp. 82-89, Jul. 1997.

[46] D.P Heyman and T.V. Lakshman, *What are the Implications of Long-Range Dependence for VBR-Video Traffic Engineering?,* IEEE/ACM Trans. Netw., Vol. 4, No. 3, pp. 301-317, Jun. 1996.

[47] K. Sriram and W. Whitt, *Characterizing Superposition Arrival Processes in Packet Multiplexers for Voice and Data,* IEEE J. Sel. Areas Commun., Vol. 4, No. 6, pp. 833-846, Sep. 1986.

[48] B.K. Ryu and A. Elwaid, *The Importance of Long-Range Dependence of VBR Video Traffic in ATM Traffic Engineering: Myths and Realities,* Computer Communication Review, vol. 26, no. 4, pp. 3-14, Oct. 1996.

[49] B.K Ryu and A. Elwaid, *The Relevance of Short Range and Long-Range Dependence of VBR Video Traffic to Real-Time Traffic Engineering,* unpublished.

[50] K. Chandra and A.R. Reibman, *Modeling One-and Two-Layer Variable Bit Rate Video,* IEEE/ACM Trans. Netw., Vol. 7, No. 3, pp. 398-417, Jun. 1999.

[51] J. Yu and A. Petropulu, *Is High-Speed Wireless Network Traffic Self-Similar?,* in Proc. of IEEE Int. Conf. Acoustics, Speech, and Signal Processing, Montreal, Canada, 2004, Vol. 2, pp. II-425-428.

[52] S. Bregni, P. Giacomazzi, and G. Saddemi, *Transport of Long-Range Dependent Traffic in Single-Hop and Multi-Hop IEEE 802.3e Networks*, in Proc. IEEE Global Telecommunications Conference, New Orleans, LA, 2008, pp. 1-6.

[53] A. Ge, F. Callegati, and L.S. Tamil, *On Optical Burst Switching and Self-Similar Traffic,* IEEE Commun. Lett., Vol. 4, No. 3, pp. 98-100, Mar. 2000.

[54] D.I. Sheluhin, S.M. Smolskiy, and A.V. Osin, *Self-Similar Processes in Telecommunications,* Chichester, UK: John Wiley & Sons, Ltd., 2007, ch.1, pp. 8-9.

[55] X. Yang, *Impulsive Self-Similar Processes, with Applications in Broadband Communication System Modeling,* Ph.D. Dissertation, Elec. Comput. Eng. Dept., Drexel Univ., Philadelphia, PA, 2001.

[56] F. M. Dekking, C. Kraaikamp H. P. Lopuhaä, and L. E. Meester, *The central limit theorem,* in A Modern Introduction to Probability and Statistics. Understanding Why and Wow. Springer-Verlag, 2005, ch. 14, pp. 195-202.

[57] M. Parulekar and A.M. Makowski, *Tail Probabilities for a Multiplexer with Self-Similar Traffic,* in Fifteenth Annual Joint Conf. IEEE Computer Societies. Networking the Next generation, San Francisco, CA, 1996, vol. 3, pp. 1452-1459.

[58] M. Parulekar and A.M. Makowski. (1996). $\mathsf{M}/\mathsf{G}/\infty$ Input Process: A Versatile Class of Models for Network Traffic, Univ. Maryland, College Park, MD, [Online]. Available: http://hdl.handle.net/1903/5778.

[59] B. Mandelbrot and W. Van Ness, *Fractional Brownian Motions, Fractional Noises and Applications*, SIAM Review, Vol. 10, No. 4, pp. 422-437, Oct. 1968.

[60] J. Beran, *Statistical Methods for Data with Long-Range Dependence,* Statistical Science, Vol. 7, No. 4, pp. 404-416, Nov. 1992.

[61] H. Rinne, *The Weibull Distribution. A Handbook,* Boca Raton, FL: Chapman & Hall/CRC, 2009.

[62] M. Parulekar and A.M. Makowski. (1995). Buffer Overflow Probabilities for a Multiplexer with Self-Similar Traffic. Univ. Maryland. College Park, MD, [Online]. Available: http://hdl.handle.net/1903/5727.

[63] M. Parulekar and A.M. Makowski (1996). Tail Probabilities for M|G|∞ Input Processes (I): Preliminary Asymtotics. Univ. Maryland. College Park, MD, [Online]. Available: http://hdl.handle.net/1903/5760.

[64] A. Patel and C. Williamson, *Statistical Multiplexing of Self-Similar Traffic: Theoretical and Simulation Results,* unpublished.

[65] R. Ritke, X. Hong, and M. Gerla, *Contradictory Relationship Between Hurst Parameter and Queuing Performance (extended version),* Telecommunication Systems, Vol. 16, No. 1-2, pp. 159-175, Jan. 2001.

[66] B.K. Ryu and S.B. Lowen, *Point Process Approaches for Modeling and Analysis of Self-Similar Traffic: Part I: Model Construction,* in Proc. Fifteenth Annual Joint Conference of the IEEE Computer Societies. Networking the Next Generation, San Francisco, CA, 1996, vol. 3, pp. 1468-1475.

[67] B.K. Ryu and S.B. Lowen, *Point Process Approaches for Modeling and Analysis of Self-Similar Traffic: Part II: Applications,* in Proc. 5th International Conference on Telecommunication Systems, Modeling, and Analysis, Nashville, TN, 1997, pp. 62-70.

[68] S.B. Lowen and M.C. Teich, *Estimation and Simulation of Fractal Stochastic Point Processes,* Fractals. Complex Geometry Patterns and Scaling in Nature and Society, vol. 3, no. 1, pp. 183-210, Mar. 1995.

Int. J. of Computers, Communications & Control, ISSN 1841-9836, E-ISSN 1841-9844
Vol. V (2010), No. 5, pp. 813-818

Software Solution for Monitoring Street Traffic and Generating Optimum Routes using Graph Theory Algorithms

M. Moise, M. Zingale, A.I. Condea

Maria Moise, Marilena Zingale, Alexandru Ioan Condea
Romanian American University, Romania
E-mail: maria.moise@rau.ro, zingale@un.org, zander.aq@gmail.com

Abstract: Nowadays, big cities are facing traffic jams, generated by the great number of automobiles in regard to the limited infrastructure capacity. Drivers are being presented with these problems: increased time spent between areas of interest, a higher risk of having an accident and of course stress suffering. In order to solve the urban traffic-jam problem a number of solutions have been developed. One of these is TomTom , which offers, free of charge, the possibility to generate navigation indications for a route. Unfortunately, the traffic monitoring service is limited to a few countries, but some countries are not on their coverage area. At this time there isn't a complete method to calculate the optimum route from a destination to another, taking into account street traffic. The only way to get relevant information is represented by the drivers personal expetraffic jams a series of solutiontraffic jams a series of solutionrience and the news on TV/radio. Thus, the choice for generating an optimum route is up to the driver/client and, as a consequence, this method is not a scientific one, being certified only empirically.
In this context, the paper presents a software solution, which determines the optimum route, taking street traffic into regard, thus contributing to a substantial reduction of time spent in traffic by drivers. The information needed for the application regarding the state of the street traffic can be supplied by the agents that check all available information sources(news bulletins, radio, police announcements) and the mobile agents that patrol the streets. The aim of this paper is to present a solution for determining the optimum route choice for cars. The solution is composed of two applications: First is MapMaker, which designs a street map. Second is BestRoute which can add traffic coefficients to streets and calculate the optimum route. In order to choose the best route two criteria are used: the minimum distance and the street traffic coefficients. The data regarding the streets map and the traffic situation is taken from a MySQL database; the optimum route from destination A to the destination B is calculated using a modified Dijkstra algorithm.
Keywords: optimum routes, graph theory, Dijkstra algorithm.

1 Introduction

Today most big cities are facing traffic jams that are generated by the exponential increase of automobiles versus the original limited traffic flow design. In order to solve the urban traffic jams a series of solutions have been proposed and used worldwide. One of these is TomTom , which offers free of charge the possibility to generate a navigation path on their website. Unfortunately, the traffic monitoring service is limited to a few countries, and Romania is not on their supported list.

In Romania such a solution has not been developed yet; existing solutions in Bucharest are limited to the synchronization of the traffic lights and the existence of a reduced number of intelligent traffic lights.

At this time a complete solution to calculate the optimum route from a destination to another, with the traffic taken into account, does not exist. The only information available represents the personal experience and/or the news on TV or radio. Thus calculating the optimum route is up to the driver(client) and as a consequence it is certified only empirically.

In this context, the paper presents a software solution, which determines the optimum route, taking street traffic into regard, thus contributing to a substantial reduction of time drivers spend in traffic. The information needed regarding the state of street traffic can be acquired either by agents that check all the available information sources or by agents that patrol the streets.

2 Solutions description

The solution for determining the optimum route is composed of two applications. First is MapMaker a map designer with the ability to use a background satellite image for easier plotting. For example, it could use a selection of satellite images from Bucharest, covering the Doamna Ghica area and the adjacent neighborhoods. After the background image is selected the map is designed node by node with the use of the mouse. The streets are made by connecting the desired nodes. The resulting map can be saved both on the local hard-disk and on an external MySQL database. The second application is BestRoute, used to set traffic coefficients and calculate optimal routes. The necessary data is taken from a MySQL database.

The application has two modes of usage:

• the client mode, which is used to calculate the optimal route based on the selection of two nodes. The result is shown both graphically and in text mode;

• the admin mode, which is used for altering the traffic coefficients of a street. In this mode, you can add, modify and delete user accounts. This mode also provides a set of reports about customers or streets.

The BestRoute application uses elements and algorithms from the graph theory. A modified Dijkstra algorithm has been used in order to calculate the minimum distance path between two nodes of the graph. The features of the algorithm's Dijkstra are the following:

• there are established a list of distances, a list of previous nodes, a list of visited nodes and a current node;

• the values from the list of distances are initialized with an infinite value, excluding the home node, which is set with value "0";

• all values in the list of visited nodes are set with value "false";

• all values in the list of previous nodes are initialized with the value "-1";

• the home node is set as the current node;

• the current node is marked as visited;

• the distances are updated based on nodes that can be viewed immediately from the current node;

• the current node is updated to the unvisited node, which may be visited by means of the shortest path from the home node;

• to be repeated (from point f) until all nodes are visited.

Formalization of the Dijkstra algorithm consists of:

Step 1. Initialization of the initial peak having a value of 0: w(X0) = 0;

Step 2. Setting up the lot A comprised of the initial node: A = X1;

Step 3. Analysis of nodes outside the lot A, namely:

i. If the nodes can be reached by direct arcs from nodes in A, then for these nodes, we calculate:

$$w(X_i) = \min(W(X_j + V(x_j, x_i)) \text{ for problems of minim} \quad (2.1)$$
$$X_j \in A$$
$$\exists(X_j, X_i)$$

And it is added to the lot A only that node for which the minimum value is obtained, then step 4 is initialized.

Step 4. Analysis of A multitude:

i. If $X_n \in A$, then its value represents the value of the optimum value path from X_1 to X_n; in order to find this path we start backwards from the final node x_n and we find the node $X_{k_1}, X_{k_2}, ..., X_{k_r}$ q which form the path searched for, where $X_{k_1} = X_n$,

$X_{k_r} = X_1$, and each other index k_{i+1} is the one for which:

$w(\ ) + v(X_{k_{i+1}}, X_{k_i}) = w(X_{k_i})$;

STOP.

ii. If$X_n \notin A$, the algorithm is resumed from step 3.

$w(\ ) + v(X_{k_{i+1}}, X_{k_i}) = w(X_{k_i})$;

STOP.

iii. If $X_n \notin A$, the algorithm is resumed from step 3.

The application BestRoute uses data from public sources and its own agents. On its execution the program queries the MySQL database and retrieves the map and its traffic coefficients, then it shows a color coded visual representation of the traffic map. Upon requesting a route from address A to address B a modified Dijkstra algorithm is used. It is also possible to select whether one wants the fastest route (traffic wise) or the shortest route (geographically). This application also has the possibility to generate reports on customers (useful for the administrator) and on the state of the streets. The reports generated by the application can be exported in Excel worksheets and/or PDF document format. For generating reports Microsoft Report Viewer component within the NET framework was used and the reports were created using the Report Wizard plus a manual tweak for better presentation of data.

3 Solution output presentation

3.1 MapMaker

In fig. 1 we have the main screen.

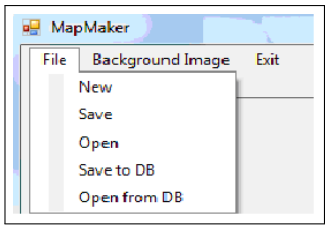


Figure 1: Main screen of module MapMaker.

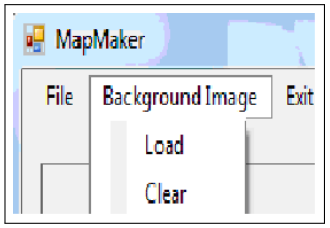


Figure 2: The menu Background Image.

By clicking on File we can:

- Create a new map - New;

- Save the existing map - Save;
- Open a saved map - Open;
- Save the map to the database - Save to DB;
- Open a saved map from the database - Open from DB.

Selecting a background image from the hard-disk is done by clicking on Background Image (fig. 2).

To design the nodes, we use the buttons from fig.3, which enable:

- Add a new node to the map - Nod nou;
- Show details about a selected street - Detalii muchie;
- Delete a selected node - Sterge;
- Clear the map scale - Clear Scara.

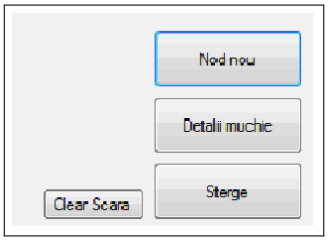


Figure 3: Buttons for creating, modifying and deleting nodes and clearing the maps scale.

Also MapMaker uses a system which automatically determines the scale for the map. It is necessary to introduce the distance between nodes only the first time. The calculation is done by the ratio between the initially introduced distance and the actual pixel distance between the two points.

3.2 BestRoute

The applications main screen is illustrated in fig. 4.

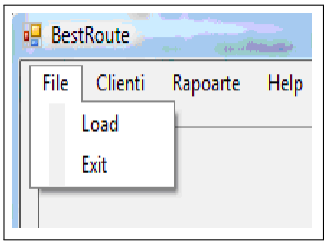


Figure 4: Options of the File menu.

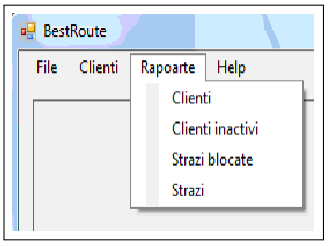


Figure 5: Options of the Reports menu.

By clicking on File we can download map and user data from the server or quit.

By clicking on Clienti while in administrator mode one can add new users, modify old ones, delete them or set whether a specific user is restricted or not.

By clicking on Reports (fig.5) one can request reports on clients, restricted clients and street traffic.

On opening the application, the system displays the map in img. 6, and selecting the start node is illustrated in fig. 7.

Generating a non-optimized route in fig.8;

After selecting the destination node the application calculates the shortest path available.

Optimized route provided by the application is illustrated in fig.9.

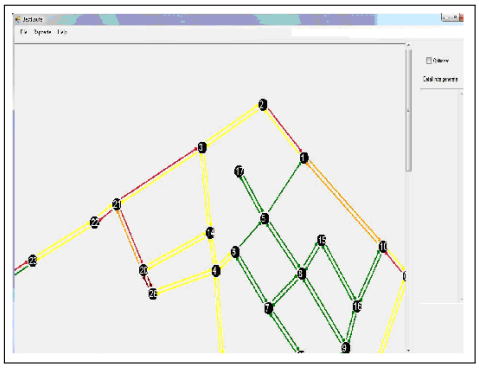


Figure 6: Initial state of the Map on opening the application.

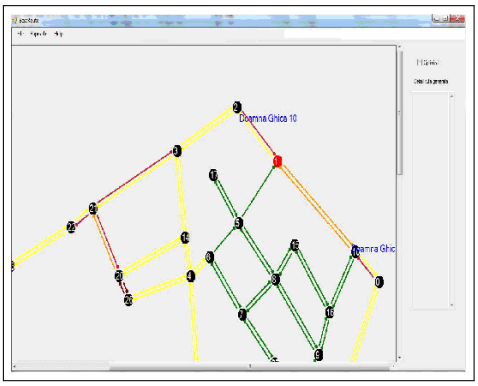


Figure 7: Selecting the start node.

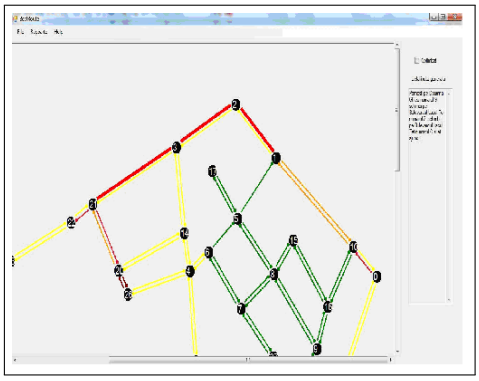


Figure 8: Image (red one) of non-optimized route.

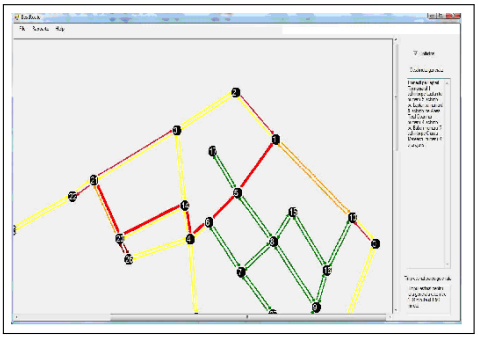


Figure 9: Image of optimized route, enabling avoiding crowded streets.

If the optimize option is checked the application will generate an optimum route to avoid heavy traffic areas. Also a time estimate is offered.

For calculating the optimal route, the following distance modifiers based on traffic coefficients were used: 1, 1.25, 1.66, 2.50, 5 and 999 for closed streets.

To calculate the estimated time, the following maximum speeds were assumed: for excellent traffic conditions a speed of 50kph, reducing speed by 10kph to the minimum of 10kph for very bad traffic. Thus, a relatively accurate time can be calculated for the generated route. For the best estimation possible, a interval of plus/minus 20% is assumed.

On both generating options a street by street text solution is also supplied.

4 Conclusions

Using such an application by drivers leads to avoiding stress and fatigue generated by traffic jam, thus reducing car crash risks. Also, the application provides information on routes unknown to drivers thus leading to a better awareness of the city. Implementing the application for a taxi or currier company offers an edge on competitors, generating a shorter delivery or reply time. On the whole, large-scale use of the application in partnership with the town hall and police may lead to general fluidization of traffic, equally improving the environment.

Bibliography

[1] Moise, M., Zingale, M., Condea, A., *Informatics application which determines the optimum routes for the cars*, in Proc. of the E-COMM-LINE 2009 Conference, Section V 35, pp. 5.

[2] Moise, M. *Data base informatics systems*, Prouniversitaria Publishing House, 2008, Bucharest

[3] *** http://msdn.microsoft.com - Microsoft Developer Network

[4] *** http://stackoverflow.com/ - StackOverflow

[5] *** http://dev.mysql.com/usingmysql/dotnet/ - Documentation MySQL

Int. J. of Computers, Communications & Control, ISSN 1841-9836, E-ISSN 1841-9844
Vol. V (2010), No. 5, pp. 819-824

Adaptive Web Applications for Citizens' Education. Case Study: Teaching Children the Value of Electrical Energy.

I. Moisil, S. Dzitac, L. Popper, A. Pitic

Ioana Moisil, Alina Pitic
Lucian Blaga Univeristy of Sibiu, Romania
E-mail: im25sibiu@gmail.com

Simona Dzitac, Laurentiu Popper
University of Oradea, Romania
E-mail: simona.dzitac@gmail.com, director@perfect-service.ro

Abstract: "The foundation of every state is the education of its youth."
Diogenes Laertius

Long-term energy saving and reduction of environmental consequences of energy consuming are among the most challenging objectives of our time. People are prone to routine and habit. To change these habits is almost a Sisif's work. In spite of continuous efforts from environmental specialists, we are witnessing an increase in electricity (and gas) consumption, at least at the level of households. Studies carried on have shown that consumers have in many cases an irrational behaviour. To correct that, researchers are studying consumers' decision making behaviour and try several intervention measures. In our paper we are presenting the design and development of a web based adaptive system aimed to educate citizens for an electrical energy saving behaviour. The system is composed of three subsystems: adaptation system, user's profile and knowledge base. We have used a user-centered design approach. For adults, users' profiles are build taking into account age group, educational level, gender, income, professional aspects, consuming behaviour. A set of questionnaires have been designed in order to collect users' data. For children, the standard profiles are more complicated and, in function of the age group, can be obtained off line through interviews or/and through online activities (games, quizzes etc.). The knowledge base is build for the electrical energy domain. The adaptation sub-system will present information to the user based on s/he profile. The system is populated with data for users of 6 to 10 years of age. For this users group a social and affective interaction design approach was used.
Keywords: adaptive web, electrical energy saving, citizens education, interaction design, user-centred design.

1 Introduction

The European Union (EU) has established that reducing energy consumption and eliminating energy wastage are among the main goals to be achieved in the near future. *"At the end of 2006, the EU pledged to cut its annual consumption of primary energy by 20% by 2020. To achieve this goal, it is working to mobilise public opinion, decision-makers and market operators and to set minimum energy efficiency standards and rules on labelling for products, services and infrastructure"*. [1]

It seems more than natural, in the information society, to use the web as a mean to make people aware of the energy saving problem. In fact there are many web sites informing about the ways of reducing energy consumption [1–6].

Analysing the content of these web presentations we have found out that though very informative and scientifically correct, they are not always answering to some users' expectations, being too serious, and too boring. This was the opinion of 56% of 200 users of age 6 to 22 investigated in 2009. It is clear that different actors that have a word to say in the process of reducing energy consumption must be approached in different ways.

It is also important to note that it is a problem of changing the behaviour, the mentality and this is a very sensitive process. Having this in mind we have tried to design and develop of a web based adaptive system aimed to educate citizens for an electrical energy saving behaviour. The system is composed of three subsystems: adaptation system, user's profile or model and knowledge base (domain model). We have used a user-centered design approach. For adults, users' profiles are built taking into account age group, educational level, gender, income, professional aspects, consuming behaviour.

A set of questionnaires have been designed in order to collect users' data. For children, the standard profiles are more complicated and, in function of the age group, can be obtained off line through interviews or/and through online activities (games, quizzes etc.). The knowledge base is build for the electrical energy domain. The adaptation sub-system will present information to the user based on s/he profile. At present the system is populated with data for users of 6 to 10 years of age.

2 Design Considerations

In our paper we are presenting a first case study with the educational software for electrical energy saving. The case study is considering users of 6 to 10 years of age. For this category of users we had use a hybrid user-centred design methodology, blending different kinds of user-centred designs with interaction (social interaction, affective interaction) and participatory design and taking into account learning objectives and learners age and preferences.

The design research phase had two steps. Firstly we identified the need of computerbased educational products, the interest of the users towards computers, the level of satisfaction concerning other computer-based educational products, what characteristics of human-computer interaction they prefer (sounds, colours, mediating agents) and also the general level of computer literacy. We have also investigated users' behaviour towards the educational/formative software market (are they buying educational software, if so, on what subjects, who is the buyer, how often, etc.). The second step aimed to investigate users' behaviour towards electrical energy consumption aspects (how interested they are, did they know what is electricity, how they perceive other educational software on this topic, are their parents concerned by energy saving, etc). 130 young pupils from both rural and urban locations participated in the design. 63% were girls and 37% boys, the age interval being from 6 to 10 years old. 84% had a computer at home. 82% have Internet connexion.

• Children in the first and forth grade (6 years old and 10 years old) are at the beginning of an educational cycle and therefore more exposed to change, with distributed interests and less interested than the others (50% are satisfied by the use of computers in class).

• 75% of the subjects that have a computer at home are very interested in using an educational software, but only 40% of those that do not have a computer at home showed a maximum interest.

• 86% of the subjects prefer computer games and 80% like also cartoons, movies, etc.; 66% of the girls choose stories and only 43% of the boys; attraction to music is growing with the age, from 60% at the age of 6 to 85% at the age of 10.

• Most parents are restricting children access to Internet.

• Interest to something new is decreasing with age, from 93% at the age of 7 to 33% at the age of 11.

• The influence of the family environment is very important.

• 27% of the subjects voted for a software based mediator agent.

• 94% of the subjects prefer software with sound incorporated.

• For 57% of the subjects software products are provided by parents, 27% are buying themselves these products.

• 87% of the children do not know the most important sources of electrical energy ("electricity is coming from the cable").

• 73% of the subjects appreciate the idea of quizzes on how to save electrical energy.

• Children with intellectual parents are more aware of the aspects concerning energy saving especially linked to global warming. Children from rural area are more concerned by the cost of energy than by global warming. Some of them are thinking that this is not their problem.

• 76% of the subjects are attracted by computers.

Based on these results it has been decided how to design the software interface and how to structure the content. The software has to use the children' preferred colours, to have sound facilities, to enable the use of the mouse but also of the keyboard, the concepts introduced have to be transparent, the children being able to navigate freely through the software, games have to be included in order to make more attractive the product and less boring, if the case, and the mediator has to be included in the software.

All these data have been obtained during working meetings with the children. For example, we have discussed the video-clip Energy, let's save it! [1]. The video-clip has been found attractive by children 8 to 10 years of age. For the others it was not interesting. One observation was that "it goes too quickly". In fact the video clip is presenting in less than 4 minutes a global view of energy waste in the day by day activities. The presentation is user-friendly, in an attractive way and it raises worries, but no more.

Another observation, concerning a different web application was that "it sounds like papa!" (Citing: "Don't leave lights on when no one is in the room. If you are going to be out of the room for more than five minutes, turn off the light." [2]). So it was very clear that the site must be designed taking into account the children preferences.

As a result we have structured the web application in four main chapters:

1. What is Electricity
    a. Short history
    b. Experiments
    c. Sources of electrical energy
        i. Classical
        ii. Alternative
2. Using and saving Electrical Energy
3. Games, activities, quizzes, puzzles, enigmas
4. Resources for teachers, parents and educators

The system architecture is composed of three subsystems: adaptation system, user's profile or model and knowledge base (domain model).

The domain model for electrical energy is a set of concepts, each concept consisting of a set of topics. Topics are linked to each other thus forming a semantic network. A topic is presented usually on a web page. A page is divided in sub-pages (chunks) that will differ in function of the user type [7–9]. The adaptation module is providing different interfaces and contents in function of the user profile (model). For the moment each page has implemented two variants: user in the age group 6 to 10, and children older than 10 years.

We have used plants and trees images ant the beginning because these were interesting for children of all ages (fig.2)

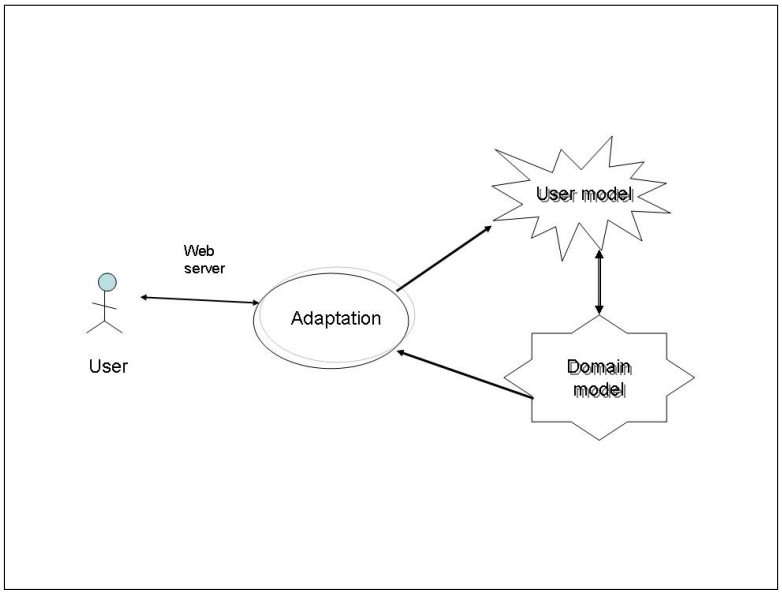


Figure 1: System architecture.

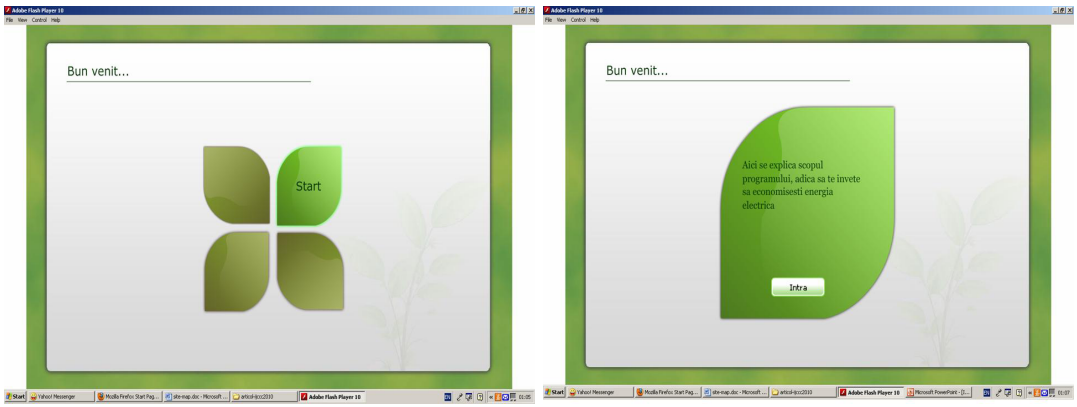


Figure 2: and 2b . Accessing the web application

The user is identified by a user name and a password. This is in fact the link to the user profile.

The user can freely navigate in the site but also s/he can choose a certain topic looking at the icons on the screen. For example in fig.4 there several icons (buttons) giving access to information on electrical energy sources. Each button is also explained by voice.

The children can experience the effect of using too much light (fig.5 and 6).

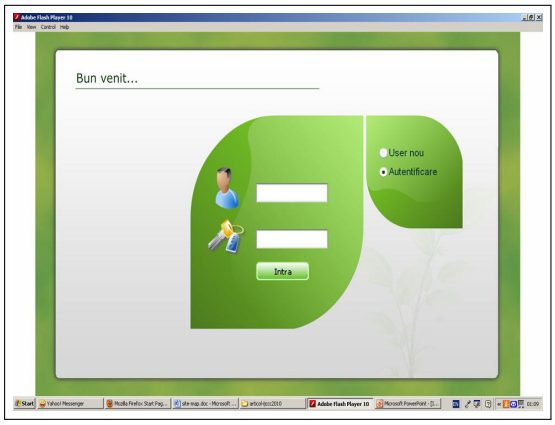


Figure 3: Access to the web application.

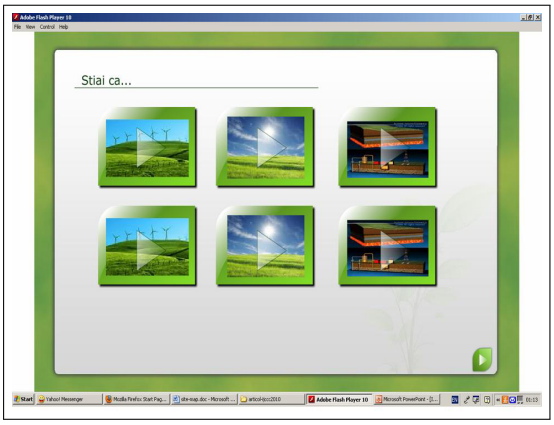


Figure 4: Buttons giving access to information on electrical energy sources.

Figure 5: All electrical bulbs are off (the tree has all leaves)

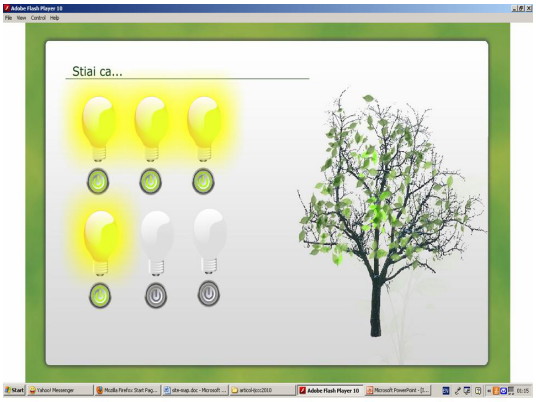


Figure 6: Four electrical bulbs are on (the tree has lost most of the leaves)

The client is implemented as Web pages. There is a number of linked frames. We have used Flash, PHP, MySQL and JavaScript language that made possible to overcome HTML limitations.

Evaluation

The current interface has been evaluated from the point of view of presentation aspects and functionalities by the children participating in the design and also by other 34 children (6 to 11 years of age). 92% were very satisfied with the web application. 7% would have liked more games and 1% declared that he is not interested in the subject. Two teachers and two parents participated also in the evaluation process. The first appreciation was positive.

3 Conclusion and Further Research

In this paper we have described the design and development of an Adaptive Webbased System for citizens' education in respect to electrical energy consumption reduction. The system is a complex one. The core system is implemented and also several modules.

We had presented a first case study with the educational software for electrical energy saving. The case study is considering users of 6 to 10 years of age. For this category of users we had use a hybrid user-centred design methodology, blending different kinds of user-centred designs with interaction (social interaction, affective interaction) and participatory design and taking into

account learning objectives and learners age and preferences. The results of the first evaluations were positive.

Future research will continue to develop the system with other users' profiles (other age category, in the first stage) and to evaluate the educational effectiveness of the system's adaptation.

Bibliography

[1] http://ec.europa.eu/energy/efficiency/index_en.htm.

[2] http://www.energyquest.ca.gov/saving_energy/index.html.

[3] http://tonto.eia.doe.gov/kids/energy.cfm?page=3.

[4] http://www.kyotoinhome.info/.

[5] www.eais.info.

[6] Thttp://www.oppapers.com/subjects/energy-saving-page1.html.

[7] Ayersman, D.J. & Minden, A.V. (1995). Individual differences, computers, and instruction. Computers in Human Behavior, 11(3-4), 371-390.

[8] Brusilovsky, P.(1996). Methods and Techniques of Adaptive Hypermedia. User Modeling and User-adapted Interaction. 6, 87-129.

[9] Jonassen, D. & Wang, S. (1993). Acquiring structural knowledge from semantically structured hypertext. Journal of Computer-based Instruction, 20(1), 1-8.

[10] Liu, Y., Ginther, D. (1999). Cognitive Styles and Distance Education. On-line Journal of Distance Learning Administration, 2,3.

[11] C.Y. Huang, T.T. Yang, W.L Chen, S. Y. Nof, Reference Architecture for Collaborative Design, Int. J. of Computers, Communications & Control, ISSN 1841- 9836, E-ISSN 1841-9844 Vol. V (2010), No. 1, pp. 71-90.

[12] Antonios Andreatos, Virtual Communities and their Importance for Informal Learning International Journal of Computers, Communications & Control, Vol. II (2007), No. 1, pp. 39-47.

Int. J. of Computers, Communications & Control, ISSN 1841-9836, E-ISSN 1841-9844
Vol. V (2010), No. 5, pp. 825-836

A Genetic Algorithm for Multiobjective Hard Scheduling Optimization

E. Niño, C. Ardila, A. Perez, Y. Donoso

Elías Niño, Carlos Ardila
Department of Computer Science
Universidad del Norte
Km 5, Via Pto Colombia. Barranquilla, Colombia
E-mail: {enino,cardila}@uninorte.edu.co

Alfredo Perez
Department of Computer Science and Engineering
University South Florida
4202 E. Fowler Ave. Tampa, Florida
E-mail: ajperez4@cse.usf.edu

Yezid Donoso
Department of Computing and Systems Engineering
Universidad de los Andes
Cra 1 No. 18A-12. Bogotá, Colombia
E-mail: ydonoso@uniandes.edu.co

Abstract: This paper proposes a genetic algorithm for multiobjective scheduling optimization based in the object oriented design with constrains on delivery times, process precedence and resource availability.

Initially, the programming algorithm (PA) was designed and implemented, taking into account all constraints mentioned. This algorithm's main objective is, given a sequence of production orders, products and processes, calculate its total programming cost and time.

Once the programming algorithm was defined, the genetic algorithm (GA) was developed for minimizing two objectives: delivery times and total programming cost. The stages defined for this algorithm were: selection, crossover and mutation. During the first stage, the individuals composing the next generation are selected using a strong dominance test. Given the strong restrictions on the model, the crossover stage utilizes a process level structure (PLS) where processes are grouped by its levels in the product tree. Finally during the mutation stage, the solutions are modified in two different ways (selected in a random fashion): changing the selection of the resources of one process and organizing the processes by its execution time by level.

In order to obtain more variability in the found solutions, the production orders and the products are organized with activity planning rules such as EDD, SPT and LPT. For each level of processes, the processes are organized by its processing time from lower to higher (PLU), from higher to lower (PUL), randomly (PR), and by local search (LS). As strategies for local search, three algorithms were implemented: Tabu Search (TS), Simulated Annealing (SA) and Exchange Deterministic Algorithm (EDA). The purpose of the local search is to organize the processes in such a way that minimizes the total execution time of the level.

Finally, Pareto fronts are used to show the obtained results of applying each of the specified strategies. Results are analyzed and compared.

Keywords: Scheduling, Process, Genetic Algorithm, Local search, Pareto Front.

1 Introduction

The Genetics Algorithms (GA) are a powerful tool for solving combinatorial problems. Nowadays, it exits a lot of algorithms inspired in GA for solving real problems such as design of vehicle suspensions [1], product deployment in telecom services [2], design of the flexible multi-body model vehicle suspensions based on skeletons implementing [3], job-shop scheduling [4], economic dispatch of generators with prohibited operating zones [5], multi-project scheduling [6], inversion analysis of permeability coefficients [7], path planning in unstructured mobile robot environments [8], rough mill component scheduling [9] and power plant control system design [10].

In productive systems is very critical the assignments of resources, for instance, a product has process and the process requires resources. The programming of the execution of the processes affects the overall cost and time of the products. Due to this, it is very important try to do the planning and scheduling in the best way. It can be accomplished with a Genetic Algorithm.

2 Preliminaries

2.1 Local Search

Local search are techniques that allows finding solutions in a set of solutions. It always tries to improve the actual solution through perturbations. A perturbation is a simple way for changing a solution. The perturbation depends of the way for representing the solutions, for instance in figure 1 can be seen a binary representation of a solution, in this case the perturbation can be done changing ones (1) by zeros (0). On the other hand, in figure 2 can be seen a no-binary representation of a solution (tour of the Traveling Salesman Problem [11] for example), in this case the perturbation can be done swapping two elements of the tour.

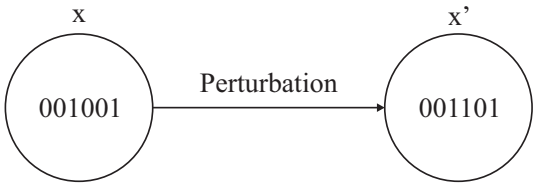


Figure 1: A binary representation of solutions. In this case x is the representation of the decimal number 9. The perturbation was done changing the 4th 0 to 1. Due to this, it creates a new solution x' that is the representation of the decimal number 13.

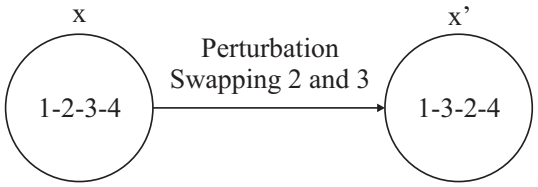


Figure 2: A no-binary representation of solutions. In this case x is the representation of a tour in TSP. The perturbation was done changing the 2 and 3 in the string. Due to this, it creates a new solution x'.

There exist a many local search algorithms such as Tabu Search (TS) [12], Simulated Annealing (SA) [13] and Exchange Deterministic Algorithm (EDA) [14]. Due to this, it is necessary to find a good representation of the solutions. Obviously, it depends of the problem to solve.

2.2 Genetic Algorithm

Genetic Algorithm (GA) is a search technique used for solving optimizations problem. The most important in GA is the design of the chromosome. It is the representation of the feasible solutions. Consequently, the behavior of the GA depends of the chromosome. Due to this, a bad chromosome implies a bad behavior of the GA. On the other hand, a good chromosome may imply a good behavior of the GA. The framework of GA can be seen in figure 3.

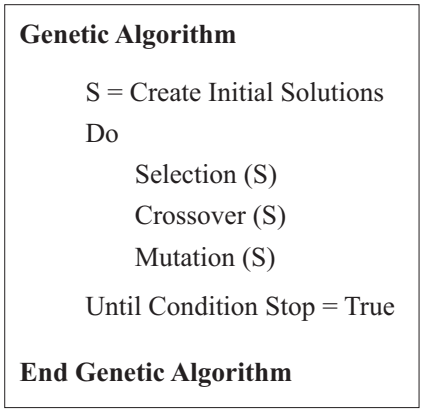


Figure 3: The Framework of a Genetic Algorithm.

GA has three important steps. First, it selects the solutions for the crossover and mutation step. This selection can be done using a metric, for example the Inverse Generational Distance (IGD) [15]. Second, it takes pairs of solutions for crossing. It can be done at random. Crossover consists in creates new solutions with parts of two solutions. The two original solutions are named parents (father and mother) and the two new solutions sons. It is created with half from father and half from mother. Lastly, it takes some sons for mutating it. The mutation is a step that allows creating new solutions. It can be done using a perturbation or a local search. The three steps of GA can be seen in figure 4.

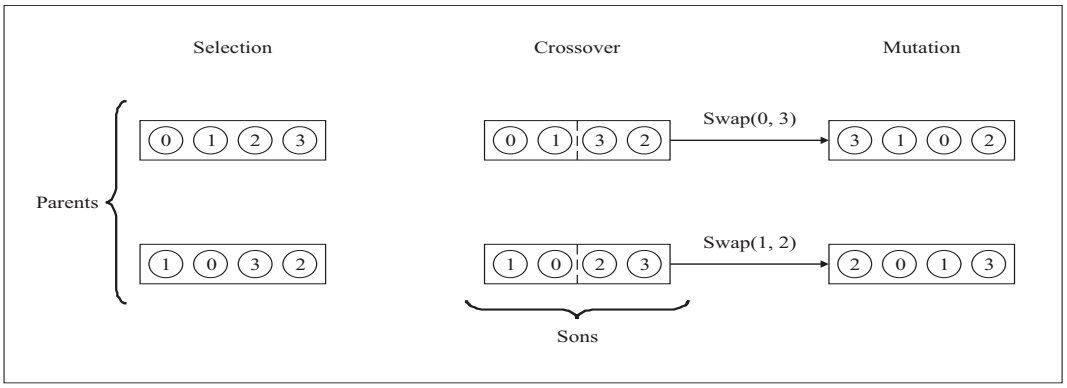


Figure 4: Steps of a Genetic Algorithm.

3 A Genetic Algorithm for Multiobjective Hard Scheduling Optimization

We state a new Genetic Algorithm for Multiobjective Hard Scheduling Optimization (GAMHSO). This algorithm works in scenarios with the following characteristics: there are production orders that are composed by a set of products. Each of the products is described by a product tree that contains all the processes needed to build such product.

For a product to be ready, it is required the execution of all processes that belongs to its tree. A process may need the execution of another process (precedence) or other subpart before it can be executed. The execution of some processes can be only done in certain times (schedules). To execute a process, a group of resources is required. The defined resources are: machinery, employees, and vehicles. It is not necessary for a group of resources to contain all types of resources.

Finally, if a subcomponent is required, it has to be constructed by a set of processes or it can be modeled by a process that indicates idle state (the subcomponent has not arrived yet to the system). This scenario can be seen in a domain model in figure 5.

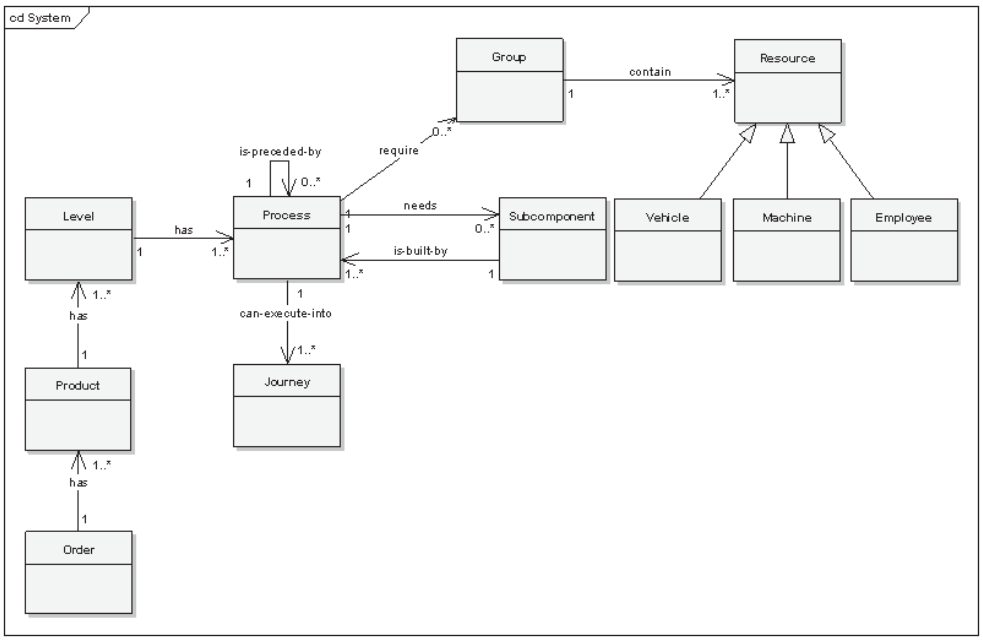


Figure 5: Domain model for scenario of GAMHSO

Formally GAHMSO is defined in figure 6. The two objectives of GAMHSO is found a set of solutions nondominated of the programming of the processes minimizing the overall time and cost.

The initial solutions are created with some heuristics. The heuristics organizes the production orders with rules. GAHMSO use three rules for creating the initial solutions. It selects the heuristic in at random. The available heuristics for creating the initial solutions are Early Due Date (EDD), Long Process Time (LPT) and Short Process Time (SPT). EDD organizes the production orders from lower to upper respect to the due date. SPT organizes the production orders from lower to upper respect to the summation of the processes time of the products. LPT is the contrary to SPT.

Once the initial solutions are created, the selection step selects solutions from the overall set of solutions. The selected solutions are named candidates. The candidates are the solutions that can be crossed for creating new solutions. The amount of candidates is defined by the candidate rate C_R.

It is very important the definition of the chromosome for the crossover step. For instance, consider the product of the figure 7. The representation of two feasible solutions could be to program the execution of processes in the order 8 - 11 - 9 - 10 - 6 - 7 - 5 - 4 - 1 - 2 - 3 and 11 - 10 - 6 - 5 - 4 - 2 - 8 - 9 - 7 - 1 - 3. The program indicates a feasible solution for programming the processes. Some processes can be executed parallel, so the order indicates the priority in the utilization of the resources. For instance, processes 8 and 9 can be executed at the same time, but in the first solution, if 8 uses a machine that 9 requires, 9 could not executed until 8 free the resource. The problem with this representation is that the crossover step could create unfeasible solutions. For example, if we split the two solutions mentioned in the middle and later we cross those solutions, we will obtain the solutions 8 - 11 - 9 - 10 - 6 - 2 - 8 - 9 - 7 - 1 - 3 and 11 - 10 - 6 - 5 - 4- 7 - 5 - 4 - 1 - 2 - 3. Obviously, those are unfeasible solutions.

```
Inputs: C_R, ρ_R
Output: A set (S) of nondominated solutions.

GAMHSO
    S = Create Initial Solutions(EDD, LPT, SPT)
    Do
        S' = Selection of Candidates (S, C_R)
        C = Crossover of Candidates (S', ρ_R)
        x = generated a integer in [0,2] with a uniform distribution.
        switch(x)
            case 0: C = TabuSearch(C)
            case 1: C = SimulatedAnnealing(C)
            case 2: C = Exchange Deterministic Algorithm(C)
        end
        S = S U C
        S = Remove Domiated Solutions(S)
    Until Condition Stop = True
End GAMHSO
```

Figure 6: The framework of GAMHSO

4 Definition of the Chromosome

The objectives of GAMSHO are the optimization of overall cost and time. It plays with the programming order of the execution of the processes. So it is very important to provide a good representation of the solutions.

Consider again the product of the figure 7. It has subparts because it is modeling the reality. But, żwhat does a subpart mean? It means that the process that contains the subpart cannot be executed until the processes that build it have been executed. In other words, there exists a precedence constrain between the process that contain the subpart and the processes

that build the subpart. If we applied this to the figure 7, we are going to obtain the tree of the figure 8. In this tree does not exist subpart, we replace the subpart for the precedence between the processes. Once the tree is ready, we need to create a chromosome that allows the representation of the programming.

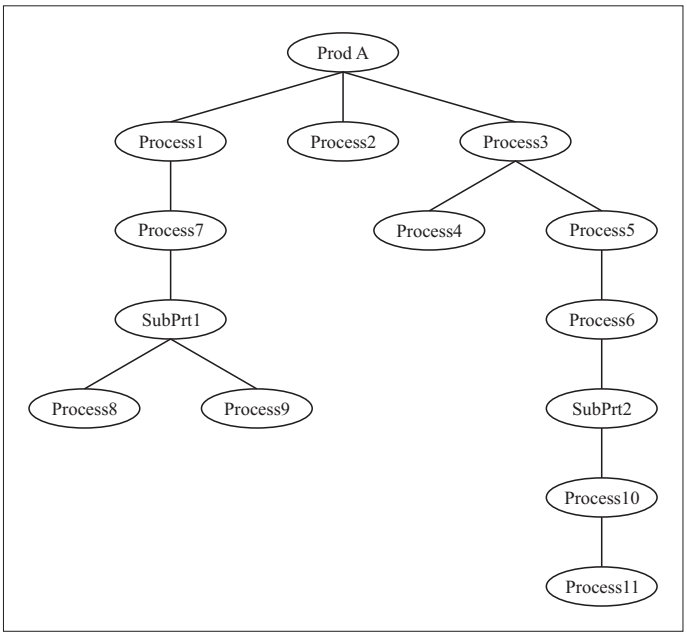


Figure 7: Tree Product A. Product A requires of the execution of the processes 1, 2 and 3. Process 1 requires the execution of the process 7. Process 7 requires the subpart1. SubPart1 requires the execution of the processes 8 and 9. Process 2 does not require processes. Process 3 require the execution of the processes 4 and 5.Process 5 requires the execution of the process 6. Process 6 requires the SubPart 2. SubPart 2 requires the execution of process 10 and process 10 requires the execution of process 11.

First we group the processes by level. It means that the time execution of a process in a superior level depends on the finalization execution time of the parents in a low level. Formally:

$$p.start\_time = max(parent_i.finalization\_time) \tag{4.1}$$

For instance, if process 5 finishes its execution in time 20 and the process 4 finishes its execution in time 40, process 3 will start its execution in time 40. On the other hand, the process 2 can be executed since time 0.

Once the levels of the processes have been identified, those are grouped in a level structure. It can be seen in figure 9. Each process knows its children in the superior level. Due to this, the chromosome for GAMHSO can be seen in figure 10. The production orders are executed from left to the right (from up to down). The products are built from left to the right (from up to down). The processes are executed from right to the left (from down to up). The processes are processed from left to the right (from up to down).

The crossover step consists in select two solutions, split in the middle and cross. It can be done by levels. An example can be seen in figure 11. The number of solution that can be crossed

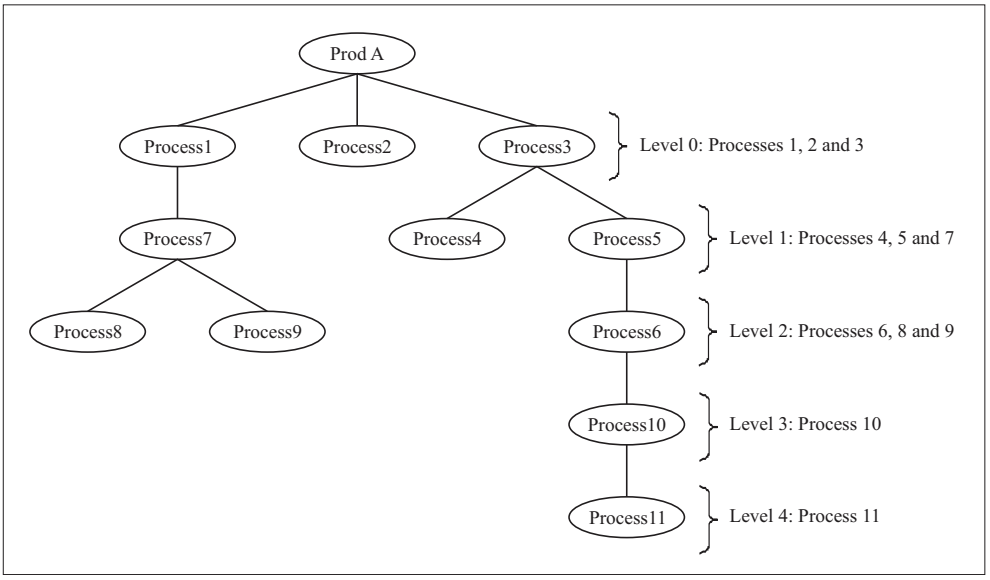


Figure 8: A view of the precedence tree of Product A group by level.

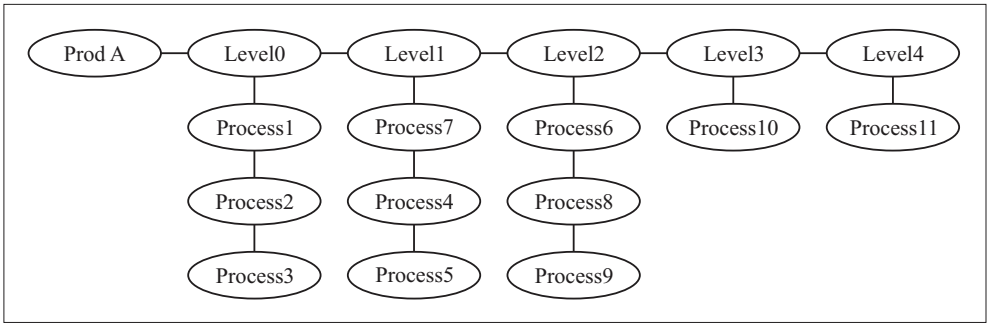


Figure 9: A representation of the Product A in a level structure.

is specify by the crossover rate (ρ_R). Once the solution is created, it is necessary to schedule the processes of the solution for obtaining the time and cost of the solution. The Programming Algorithm (PA) is an algorithm that requires a solution for programming all the processes for all the products of the production orders. It verifies is a process can be executed with three validations: First, it verifies the process precedence. Second, it verifies if the resource are available for the execution of the process. Lastly, it verifies is the process can be executed in the journey. Once a process completes its execution, it set the initial time to his children to his finalization time. PA can be seen in figure 12.

The mutation step of GAMHSO consists in the improvement of the solutions through Local Search (LS). GAMHSO works with three LS: Tabu Search (TS), Simulated Annealing (SA) and Exchange Deterministic Algorithm (EDA).

5 Experimental Settings

We tested GAMHSO in a computer AMD Turion 64, 2 GB of RAM and a Hard Disk of 120 GB.

The test consisted in build 1000 products of type A (Figure 7). The parameters were $C_R = 0.4$,

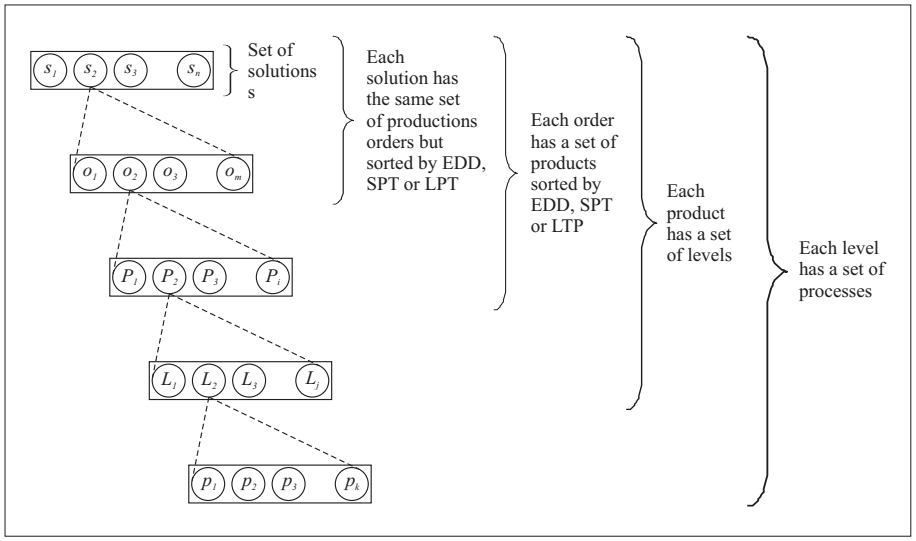


Figure 10: The Chromosome for GAMHSO.

$\rho_R = 0.5$. We tested the performance of GOMHSA with EDA, TS and SA. For making a real comparison, we use the Inverted Generational Distance (IGD) metric [15]. It is defined as follows:

Given a reference set A^*, the IGD value of a set $A \subset R^m$ is defined as:

$$IGD(A, A^*) = \frac{1}{|A^*|} \sum_{v \in A^*} \{\min_{u \in A} d(u, v)\} \tag{5.1}$$

6 Results of GAMHSO

The Pareto Fronts for each LS can be seen in figure 12. The results of IGD metric between the LSs can be seen in table 1. The running times for GAMHSO for each LS can be seen table 2.

	EDA	SA	TS
EDA	0	1137.0605	7485.2654
SA	1134.2321	0	6157.3669
TS	7367.1584	6388.229	0

Table 1: The IGD-metrics values for each LS against the rest of LS.

	Running Time in seconds
EDA	32
TS	57
SA	128

Table 2: Running times of GAMHSO with each LS.

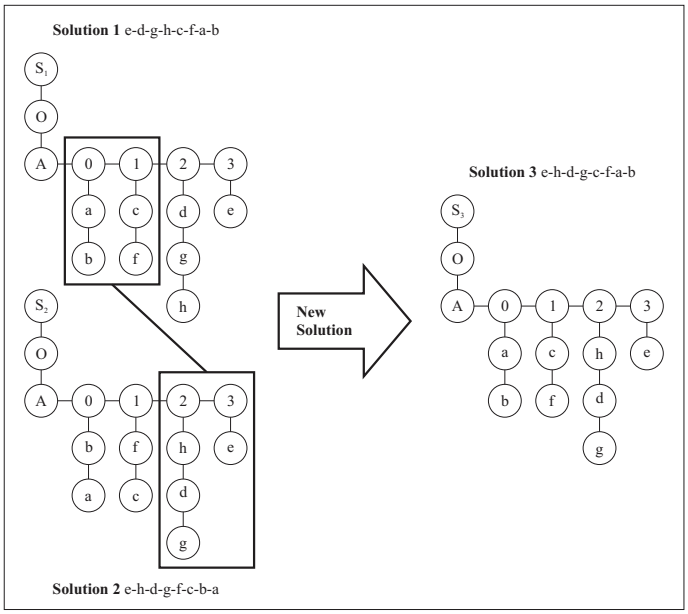


Figure 11: An example of the crossover step of GAMHSO.

7 Conclusions and Future Works

We designed a new Genetic Algorithm for Hard Scheduling Optimization (GAMHSO). It works with very difficult scenarios of productive systems. Also, we define a chromosome for GAMHSO that avoids the creation of unfeasible solutions. Due to this, it is not necessary to verify if a solution (for the crossover step) is a feasible solution. Consequently, the performance of the algorithm is satisfactory in comparison with the size of the feasible solutions space. On the other hand, we state a new Programming Algorithm (PA) for scheduling of a set of production orders. PA is a flexible algorithm that allows the incorporations of new restrictions to the processes. It allows calculate the overall time and cost of a set of production orders. We made a real comparison of the GAMHSO behavior with some local search strategies such as Exchange Deterministic Algorithm (EDA), Tabu Search (TS) and Simulated Annealing (SA). The best performance of GAMHSO was using EDA and SA. Lastly, we will investigate a new chromosome that allows the crossover between production orders.

```
Programming Algoritm(O as a set of production orders)
        Cost = 0
        Time = 0
        For Each Order o in O
                For Each Product P in o
                        For Each Level l in P
                                For Each Process p in P
                                t = p.initialTime
                                        Do
                                                While p cannot be executed in time t
                                                        t = t+1
                                                End While
                                                it = t
                                                While p can be executed in time t
                                                        t = t+1
                                                End While
                                                ft = t
                                                For all Resources r used for p. r.programTask(it,ft). Cost = Cost+r.getCost
                                        Until t = 0
                                        For all Processes Children pc of p. pc.initialTime = max(pc.initialTime,t)
                                        If  t > Time Then
                                                Time = t
                                        End If
                                End For
                        End For
                End For
        End For
        Return Time,Cost
End Programming Algorithm
```

Figure 12: The Framework of the Programming Algorithm (PA) for getting the overall cost and time of the programming of a set of production orders.

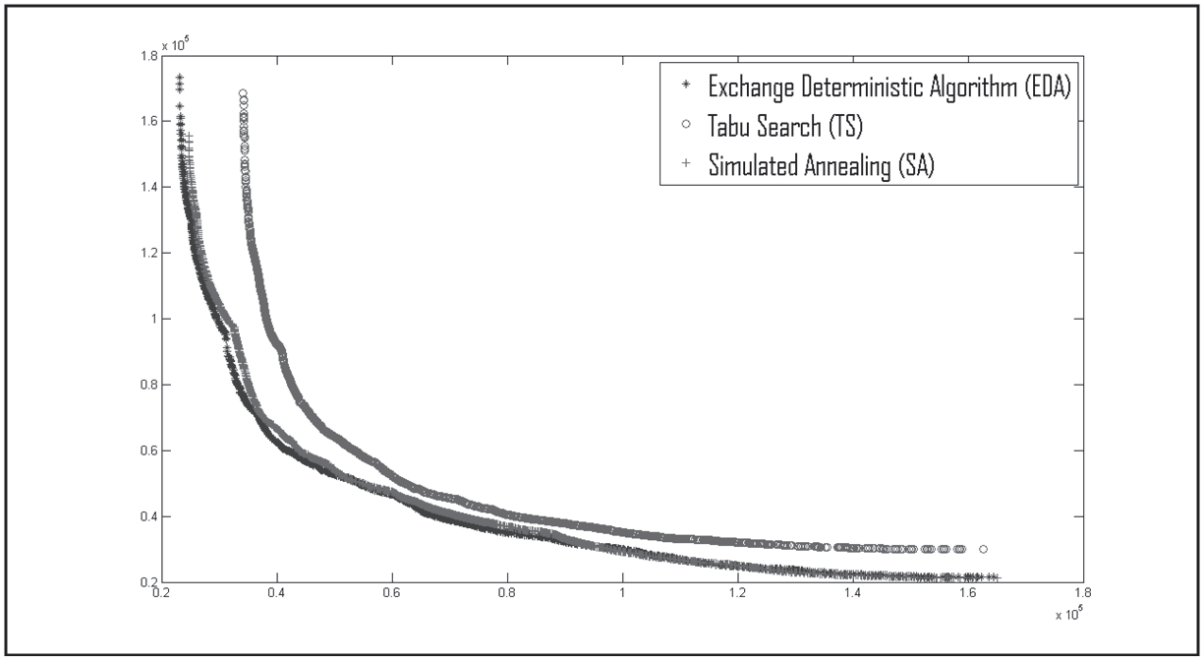


Figure 13: Pareto Fronts for GAHMSO for each LS. Notice that TS is dominated by SA and EDA. GAHMSO a similar behavior for SA and EDA

Bibliography

[1] Jingjun Zhang; Yanhong Zhang; Ruizhen Gao, "Genetic Algorithms for Optimal Design of Vehicle Suspensions", Engineering of Intelligent Systems, 2006 IEEE International Conference on , vol., no., pp.1-6, 0-0 0.
URL: http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=1703182&isnumber=35938

[2] Murphy, L.; Abdel-Aty-Zohdy, H.S.; Hashem-Sherif, M., "A genetic algorithm tracking model for product deployment in telecom services", Circuits and Systems, 2005. 48th Midwest Symposium on , vol., no., pp.1729-1732 Vol. 2, 7-10 Aug. 2005.
URL: http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=1594454&isnumber=33557

[3] Guangyuan Liu; Jingjun Zhang; Ruizhen Gao; Yang Sun, "A Coarse-Grained Genetic Algorithm for the Optimal Design of the Flexible Multi-Body Model Vehicle Suspensions Based on Skeletons Implementing", Intelligent Networks and Intelligent Systems, 2008. ICINIS '08. First International Conference on , vol., no., pp.139-142, 1-3 Nov. 2008.
URL: http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=4683187&isnumber=4683146

[4] Wu Ying; Li Bin, "Job-shop scheduling using genetic algorithm", Systems, Man, and Cybernetics, 1996., IEEE International Conference on , vol.3, no., pp.1994-1999 vol.3, 14-17 Oct 1996.
URL: http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=565434&isnumber=12283

[5] Orero, S.O.; Irving, M.R., "Economic dispatch of generators with prohibited operating zones: a genetic algorithm approach", Generation, Transmission and Distribution, IEE Proceedings-, vol.143, no.6, pp.529-534, Nov 1996.
URL: http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=556730&isnumber=12146

[6] Zhao Man; Tan Wei; Li Xiang; Kang Lishan, "Research on Multi-project Scheduling Problem Based on Hybrid Genetic Algorithm", Computer Science and Software Engineering, 2008 International Conference on , vol.1, no., pp.390-394, 12-14 Dec. 2008.
URL: http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=4721769&isnumber=4721668

[7] Xianghui Deng, "Application of Adaptive Genetic Algorithm in Inversion Analysis of Permeability Coefficients", Genetic and Evolutionary Computing, 2008. WGEC '08. Second International Conference on , vol., no., pp.61-65, 25-26 Sept. 2008.
URL: http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=4637395&isnumber=4637374

[8] Yanrong Hu; Yang, S.X.; Li-Zhong Xu; Meng, Q.-H., "A Knowledge Based Genetic Algorithm for Path Planning in Unstructured Mobile Robot Environments", Robotics and Biomimetics, 2004. ROBIO 2004. IEEE International Conference on , vol., no., pp.767-772, 22-26 Aug. 2004.
URL: http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=1521879&isnumber=32545

[9] Siu, N.; Elghoneimy, E.; Yunli Wang; Gruver, W.A.; Fleetwood, M.; Kotak, D.B., "Rough mill component scheduling: heuristic search versus genetic algorithms" Systems, Man and Cybernetics, 2004 IEEE International Conference on , vol.5, no., pp. 4226-4231 vol.5, 10-13 Oct. 2004.
URL: http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=1401194&isnumber=30426

[10] Lee, K.Y.; Mohamed, P.S., "A real-coded genetic algorithm involving a hybrid crossover method for power plant control system design", Evolutionary Computation, 2002. CEC '02.

Proceedings of the 2002 Congress on , vol.2, no., pp.1069-1074, 2002.
URL: http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=10043914&isnumber=21687

[11] Pepper, J.W.; Golden, B.L.; Wasil, E.A., "Solving the traveling salesman problem with annealing-based heuristics: a computational study", Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on , vol.32, no.1, pp.72-77, Jan 2002.
URL: http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=995530&isnumber=21479

[12] Blesa, M.J.; Hernandez, L.; Xhafa, F., "Parallel skeletons for tabu search method", Parallel and Distributed Systems, 2001. ICPADS 2001. Proceedings. Eighth International Conference on , vol., no., pp.23-28, 2001.
URL: http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=934797&isnumber=20222

[13] Rose, J.; Klebsch, W.; Wolf, J., "Temperature measurement and equilibrium dynamics of simulated annealing placements", Computer-Aided Design of Integrated Circuits and Systems, IEEE Transactions on , vol.9, no.3, pp.253-259, Mar 1990.
URL: http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=46801&isnumber=1771

[14] Niño, E.;Ardila, J. , Algoritmo Basado en Automatas Finitos Deterministas para la obtención de óptimos globales en problemas de naturaleza combinatoria. Revista de Ingeniería y Desarrollo. No 25. pp 100 - 114. ISSN 0122 - 3461.

[15] Minzhong Liu; Xiufen Zou; Yu Chen; Zhijian Wu, "Performance assessment of DMOEA-DD with CEC 2009 MOEA competition test instances", Evolutionary Computation, 2009. CEC '09. IEEE Congress on , vol., no., pp.2913-2918, 18-21 May 2009.
URL: http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=4983309&isnumber=4982922

[16] H. Li and J.D. Landa-Silva, Evolutionary Multi-objective Simulated Annealing with Adaptive and Competitive Search Direction, Proceedings of the 2008 IEEE Congress on Evolutionary Computation (CEC 2008), IEEE Press, pp. 3310-3317, 01-06 June, 2008, Hong Kong.

Int. J. of Computers, Communications & Control, ISSN 1841-9836, E-ISSN 1841-9844
Vol. V (2010), No. 5, pp. 837-843

Modeling Gilliland Correlation using Genetic Programming

M. Olteanu, N. Paraschiv, O. Cangea

Marius Olteanu, Nicolae Paraschiv, Otilia Cangea
"Petroleum-Gas" University of Ploiesti
Romania, Ploiesti, 100680, Bucuresti blvd., no.39
E-mail: {molteanu,nparaschiv,ocangea}@upg-ploiesti.ro

Abstract: The distillation process is one of the most important processes in industry, especially petroleum refining. Designing a distillation column assesses numerous challenges to the engineer, being a complex process that is approached in various studies. An important component, directly affecting the efficient operation of the column, is the reflux ratio that is correlated with the number of the theoretical stages, a correlation developed and studied by Gililland. The correlation is used in the case of simplified control models of distillation columns and it is a graphical method. However, in many situations, there is the need for an analytical form that adequately approximates the experimental data. There are in the literature different analytical forms which are used taking into account the desired precision. The present article attempts to address this problem by using the technique of Genetic Programming, a branch of Evolutionary Algorithms that belongs to Artificial Intelligence, a recently developed technique that has recorded successful applications especially in process modeling. Using an evolutionary paradigm and by evolving a population of solutions or subprograms composed of carefully chosen functions and operators, the Genetic Programming technique is capable of finding the program or relation that fits best the available data.
Keywords: Gilliland correlation, artificial intelligence, genetic programming.

1 Introduction

The early pioneers of computer science, like Alan Turing, John von Neumann, Norbert Wiener studied natural systems as guiding metaphors for their desire to understand nature and create intelligent computer programs capable to learn and adapt to their environment. In the 1950s and 1960s, several scientists from Germany and United States independently studied evolutionary systems with the aim to use evolution as an optimization tool for engineering problems. Several techniques have been created in this period by different research groups: Evolution Strategies, Evolutionary Programming and Genetic Algorithms (see [5]). The basic idea behind all this techniques was to start with a random population of candidate solutions to the specific problem and by applying a set of genetic operators, inspired from the field of genetics, to modify this candidate solutions in such a way to achieve a better fitness or adequacy of the solution for the engineering problem in an iterative process.

In this article we apply a technique of Evolutionary Algorithms, that of Genetic Programming, with the aim at finding an analytic expression for a well studied and used correlation, the Gililland correlation, applied to the design of control models for distillation columns. The algorithms for implementing Genetic Programming are characterized by many heuristic tuning parameters, this paper underlines the most important ones as a result of the simulations.

2 Genetic programming based symbolic regression

Genetic programming was introduced by John Koza (see [4]) and it can be seen as an extension of the Genetic Algorithms by the increase of the complexity of the structures used to represent the potential solution to the problem. In his 1992 book ([1]), John Koza suggested that these potential solutions should be represented as trees of functions and operators, dynamic structures of varying size and shape. The classes of problems that can be best approached using this technique are symbolic regression, in which an analytic expression for a function has to be discovered such that a set of experimental data is fitted and machine learning, domain that uses a set of possible computer programs that produce the desired behavior in the case of some particular input data.

In genetic programming, the set of possible structures is determined by the set of N_f functions from $F = \{f_1, f_2, ..., f_{N_f}\}$, each function can take a specified number of arguments denoted $a(f_i)$ called its *arity* and the set of N_t terminals from $T = \{t_1, t_2, ..., t_{N_t}\}$. Some examples of functions are: arithmetic operations: plus (+), minus (-) , multiply (*), divide (/); mathematical functions: logarithm (log), trigonometric functions - sin, cos, etc; logical functions: AND, OR, NOT. The terminals are variable atoms that represent input variables, signals form sensors / detectors or constant atoms, for example the number 11.25 or the boolean constant *true*. An example of a simple function represented by such a tree is given in figure 1, the corresponding analytic expression being: $f(x, y) = x + \sqrt{y} - 2$.

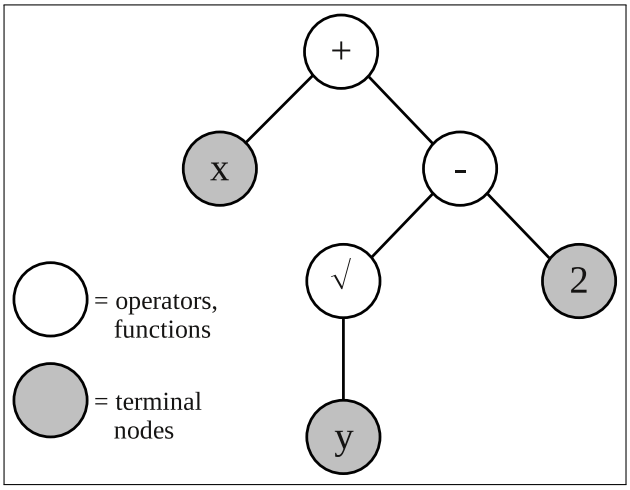


Figure 1: Example of a tree representing a possible solution

As a particular aspect of functions and terminals representation is that of closure which assumes that each of the function from the set must be capable to accept as its input arguments any value and data type that could possibly be returned by any function in the set and any value accepted by any terminal. Another property is that of sufficiency which states that the set of functions and terminals must be capable of defining a solution for the actual problem, the designer of the algorithm is the one that decides what are the most probable functions and constants that could express best a solution. The adequacy or fitness of a particular member of the population has to be measured for a set of fitness cases, in the case of symbolic regression these are the experimental data. The algorithm for implementing Genetic Programming has a structure that resembles that of a Genetic Algorithm:

1. The designers establishes the set of functions / operators and the set of terminals for the specific problem

2. t=0, t - being the generation counter
3. A random initial population of solutions P(0) is generated
4. Fitness evaluation for all the population
5. A new population is created by applying the genetic operators of cloning, mutation and crossover.
6. If the stop condition is met, than the algorithm stops, else t=t+1 and go to step 4

3 Genetic programming applied to the Gilliland correlation

The Gilliland correlation or Gilliland plot (figure 2) correlates the reflux ratio and the number of theoretical stages for a distillation column (see [3]):

$$x = \frac{R - R_{min}}{R + 1}, y = \frac{N - N_{min}}{N + 1} \tag{3.1}$$

given that the minimum reflux ratio R_{min} is calculated from Underwood's equation and the minimum number of stages N_{min} by Fenske's method.

The resulting curve stretches form (0,1) coordinate at minimum reflux to (1,0) at total reflux. In the literature (see [3]) there are a series of numerical equations derived for this correlation, but because there is some scatter in the fit of data to the Gilliland plot, the expressions that best fits the plot are not always the best reflux-stages correlations.

One of the most used expressions is that of Molokanov:

$$y = 1 - e^{\frac{1-54.4x}{11+117.2x} \cdot \frac{x-1}{\sqrt{x}}} \tag{3.2}$$

used for higher precision, alternative relations being that of Eduljee:

$$y = 0.75(1 - x^{0.5668}) \tag{3.3}$$

and Rusche (see [2]):

$$y = 0.1256 \cdot x - 0.8744 \cdot x^{0.291} \tag{3.4}$$

Also, other correlations have been obtained as a result of research in the domain of optimal control (see [6], [7] - polynomial analytical expression).

For implementing the Genetic Programming algorithm it was used a free MatLab toolbox created by S. Silva, a toolbox well documented and highly modular having many configurable parameters (see [8]).

A usual running of the algorithm is started by entering the following command at the Mathlab prompt: >> [vars, b]=gplab(g,n);
where:

g=maximum number of generations as stop condition
n=the number of individuals in the population
vars=a structure containing all the algorithm variables
b=best fitted individual

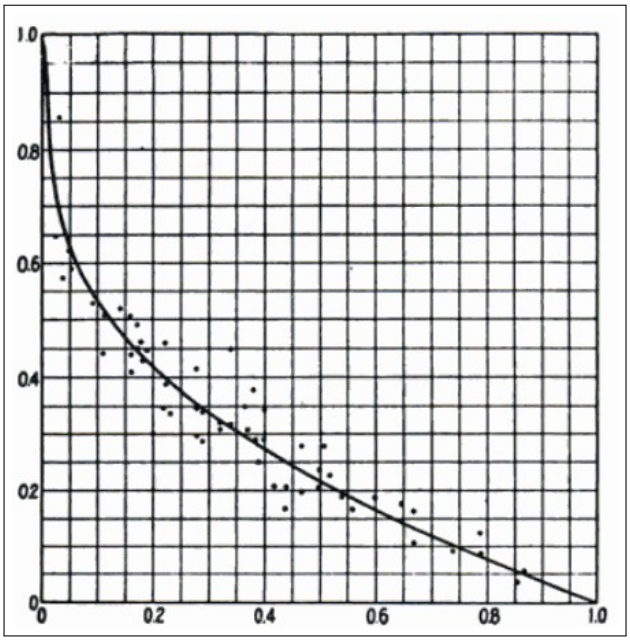


Figure 2: Plot of the original Gilliland correlation

The main modules that compose the toolbox have the following functions: *Variables initializing*, *Initial population* and *Generation creating*. Among the many important features of the toolbox, the parameters that were especially important in the symbolic regression applied for the Gililland correlation:

- Population initialization has three possible methods: *fullinit*, *growinit* and *rampedinit* which was used in the algorithm, and that produces an initial population having very diverse trees (a combination of *fullinit* and *growinit* methods, see [4], [8]);
- With the purpose of avoiding function bloating, the toolbox uses a parameter called *dynamiclevel* that specifies if the trees depth or the trees number of nodes has a fixed limit or not. Another experimental property, called *veryheavy* specifies if this dynamic limit can be decreased under the initial value during the running in the case that the best individual has a smaller depth or number of nodes. By using this option, a much simpler expression for the function has been obtained, and also the running time of the algorithm and the memory resources implied were substantially reduced;
- The methods available for selecting the most adequate individuals are the classical *roulette* and *tournament* methods, in addition there are implemented other methods like *lexictour* or *doubletour* that chooses taking into account the shortest (having the smaller depth or number of nodes) individual. The best results were achieved using the *lexictour* method.

The *crossover* operator (figure 3) randomly choose nodes from both parents and swaps the respective branches creating one or two offspring, in the case of the *mutation* a random node is chosen from the parent individual and substituted by a new random tree, taking into account the imposed restrictions on the depth and number of nodes (see [1]).

A set of six functions have been chosen and two random constants with values between 0 and 1. The functions were: plus, minus, times, custom divide (having protection to divide-by-zero error) also custom square root and custom natural logarithm.

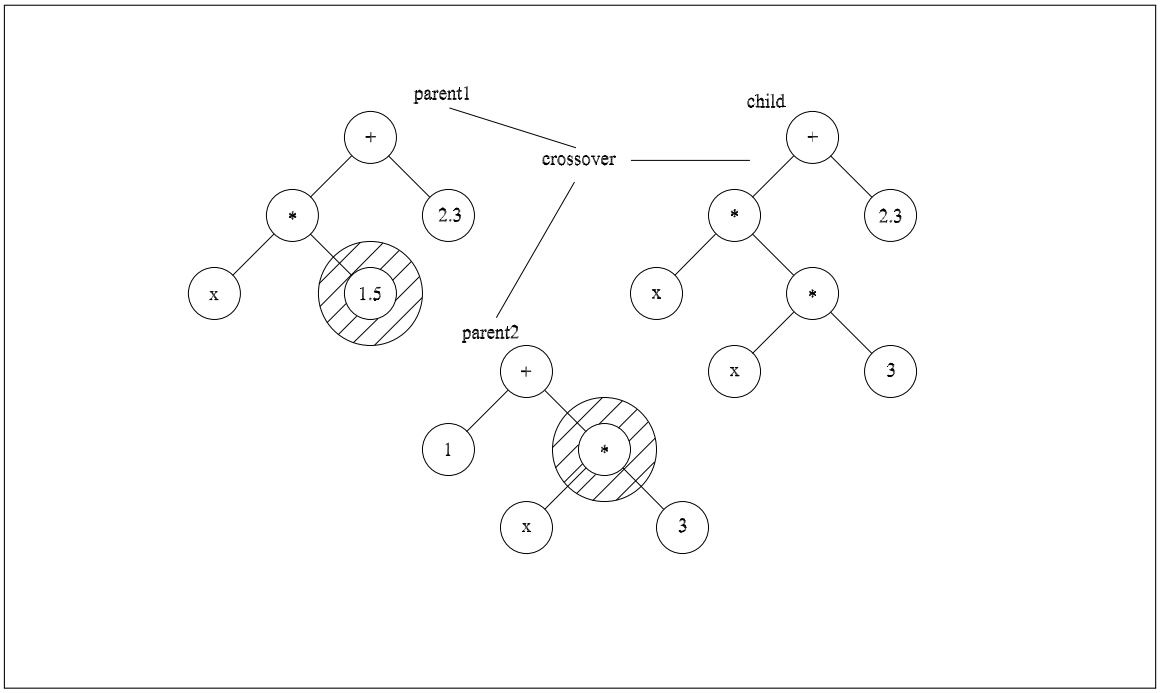


Figure 3: Example of applying the crossover operator

The algorithm ran on a computer with an Intel Core 2 Duo processor, having 2GB of memory and Matlab 7.4 for Windows XP. For a population with 1000 members and a number of 20 generations the running time was in the range of 4 minutes to 4.5 minutes. If the number of generations is increased too much it always results a very big expression for the final function, making it very hard for implementing and studying, a very slow increase in performance (fitness) being obtained. Adding the two final points (0,1) and (1,0) in the data set conducts to a function that does not approximate well the middle data points having a poor general performance because of the relative big scatter of the terminal points from the plot, so for most of the simulations, the two ending points were not included. The fitness function used calculates the sum of the absolute difference between the desired output values and the value computed by the individual on all fitness cases.

$$\mathrm{fitness} = \sum_{i=1}^{N} |y_i - f(x_i)| \tag{3.5}$$

where N is the number of fitness cases, y_i the desired output and $f(x_i)$ is the value returned by the individual.

With a generation number of 40 and a population of 500 individuals for a running time of 169 seconds, the expression obtained is presented in the following tree plot (figure 4):

Another common representation is the string representation, used in Matlab to represent the function:

f=plus(times(minus(mysqrt(mydivide(0.96486,0.56835)),times(plus(0.96486,0.33765),mysqrt(X1))),mylog(mysqrt(mydivide(0.8073,0.22837)))),mysqrt(plus(times(0.33765,times(X1,times(plus(mysqrt(0.8073),0.56835),mysqrt(X1)))),mylog(minus(mysqrt(mydivide(0.9393,0.56835)),times(plus(mysqrt(0.8073),0.56835),mysqrt(X1)))))),

from which we can write the following simplified relation:

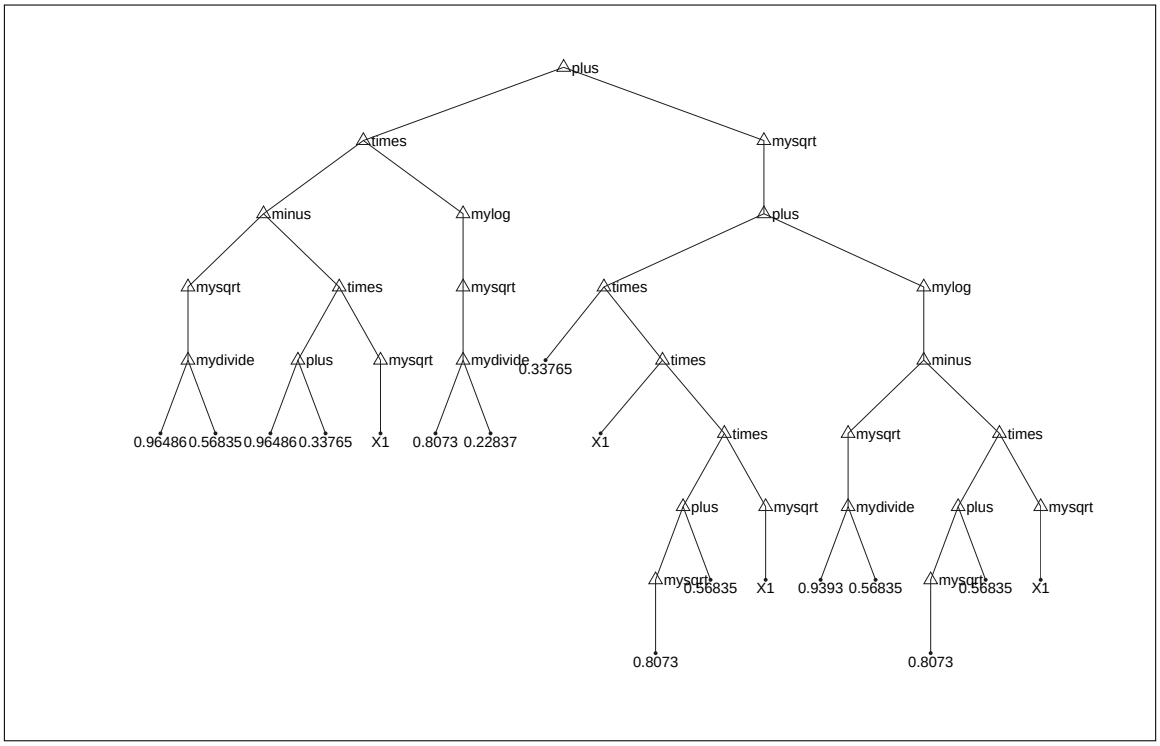


Figure 4: The tree representation of the final solution

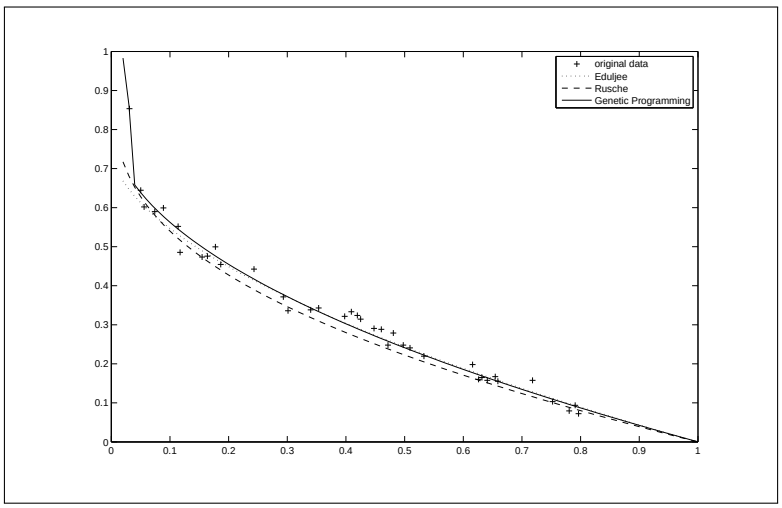


Figure 5: Plot of the final solution and other correlations together with the original data

$$f = 0.631(1.30294 - 1.30251\sqrt{x}) + \sqrt{|0.495 \cdot x\sqrt{x} + \ln(|1.285 - 1.467\sqrt{x}|)|} \quad (3.6)$$

From the plot of the classic correlations (figure 5) and that of the Genetic Programming function we can conclude that a very good approximation is obtained, challenging the methods used in the domain of identification.

4 Conclusions and Future Works

The study presented in this article aims at applying a recently and not too well studied technique of Artificial Intelligence, that of Genetic Programming to the problem of finding the best function that fits some data. The well known Gililland correlation, applied in the domain of process control of the distillation process has been studied. Using the classic representation of the potential solutions as trees, many interesting results have been obtained. After running the algorithm with varying parameters for initial population for the method of selection the best individuals, the method of creating the new offspring, the genetic operators of crossover and mutation and ending with the number of generations and of individuals in the population, it can be stated that the technique of Genetic Programming has a promising future potential, proved by the good estimation of the experimental data. One aspect that proved to be important is the careful chosen of the running parameters which directly influence the quality of the solution.

Bibliography

[1] M. Affenzeller, S. Winkler, S. Wagner, A. Beham, *Genetic Algorithms and Genetic programming - Modern Concepts and Practical Applications,* CRC Press, 2009.

[2] J.R. Couper, W.R. Penney, J.R. Fair, S.M. Wallas, *Chemical Process Equipment - Second Edition,* Elsevier - Gulf Professional Publishing, 2005.

[3] H.Z. Kister, *Distillation Design,* McGraw-Hill, 1992.

[4] J.R. Koza, *Genetic Programming - On the Programming of Computer by Means of Natural Selection,* MIT Press, 1992.

[5] M. Mitchell, *An Introduction To Genetic Algorithms,* MIT Press, 1996.

[6] N. Paraschiv, *Equipment and Programs for Optimal Control of Fractionation Processes*, PhD Thesis, "Petroleum-Gas" University of Ploiesti, Ploiesti, 1987.

[7] N. Paraschiv, *An Analytical Form of the Gilliland Graphical Correlation for the Advanced Control of Fractionation Processes,* Chemistry Magazine no.7-8, 1990.

[8] S. Silva, *GPLAB - A Genetic Programming Toolbox for Matlab User Manual,* ECOS - Evolutionary and Complex Systems Group, University of Coimbra, Portugal, 2007 - accessible from http://gplab.sourceforge.net/index.html.

Int. J. of Computers, Communications & Control, ISSN 1841-9836, E-ISSN 1841-9844
Vol. V (2010), No. 5, pp. 844-851

A Microcontroller-based Intelligent System for Real-time Flood Alerting

M. Oprea, V. Buruiana, A. Matei

Mihaela Oprea, Vasile Buruiana, Alexandra Matei
University Petroleum-Gas of Ploiesti, Department of Informatics
Bdul Bucuresti No 39, 100680, Ploiesti, Romania
E-mail: mihaela@upg-ploiesti.ro, bvasea@gmail.com, matei.alexandra@hotmail.com

Abstract: The development of efficient flood alerting systems became more demanding in the last years. In this paper it is presented the first version of a prototype intelligent system for flood forecasting and real-time alerting. The system is implemented by using a microcontroller from the ARM family, Marvell 88F6281, and has a user interface realized under the Free UnixBSD operating system. Also, a knowledge base is integrated in the system. The real-time alert is sent to the decision making factors via a communication channel (such as, the internet, a mobile phone, or radio communication). Some experimental results obtained so far are also discussed in the paper.
Keywords: microcontroller, flood, artificial intelligence.

1 Introduction

The development of an efficient hydrologic monitoring system requires the use of an automated data acquisition system, the analysis of several parameters that are monitored (water level, water flow, rainfall fell etc), and a real-time alerting system in case of flood production. As hydrologic monitoring is very important in flood prevention, in the last years, different modern techniques were applied for flood forecasting, including some artificial intelligence based approaches, such as expert systems, artificial neural networks, intelligent agents and multiagent systems (see [8]). The systems that were reported in the literature are specific to a given hydrographic basin or to a dam or river, and provide particular solutions, which hardly can be adapted and used in other situations. Our long term research purpose is to design and implement a prototype intelligent system for flood forecasting that provides real time alert in case of flood production, and can be adapted to any hydrographic area (basin, river or dam) with minimal infrastructure changes. The objective is to develop a flexible and robust intelligent system by keeping it as simple as possible. Some related work reported by the research community in the area of real time alert systems provides solutions such as the embedded controller systems (see [2]), the microprocessor systems (see [3]), and the mobile sensor nodes for alert systems applications (see [9]). A preliminary report on the use of the microcontroller-based system, described in this paper, was given in (see [5]), where the application was in the area of real time seismic alert, and no artificial intelligence technique was used. We have started the development of an intelligent prototype system from our previous work described in (see [5]) (data acquisition and alert transmition), (see [6]) (implementing a low power, high performance BSD Unix microcontroller-based system), (see [7]) (intelligent algorithms) and (see [8]) (agent-based modelling).

In this paper it is presented the first version of a prototype intelligent system for flood forecasting and real-time alerting. The system is implemented by using a microcontroller from the ARM-family, and has a user interface realized under the UnixBSD operating system, and a knowledge base integrated in the system. The real-time alert is sent to the decision making factors

via Internet or mobile phone or by radio communication. The experimental results obtained so far are also discussed in the paper.

2 The architecture of the microcontroller-based intelligent system

The microcontroller-based intelligent system was designed arround the Marvell 88F6281 microcontroller by using the Marvell Sheevaplug development platform (see [11]) , and the Unix Free BSD 8.0 Beta 2 operating system (OS) (see [12]) . The architecture of the system is presented in figure 1.

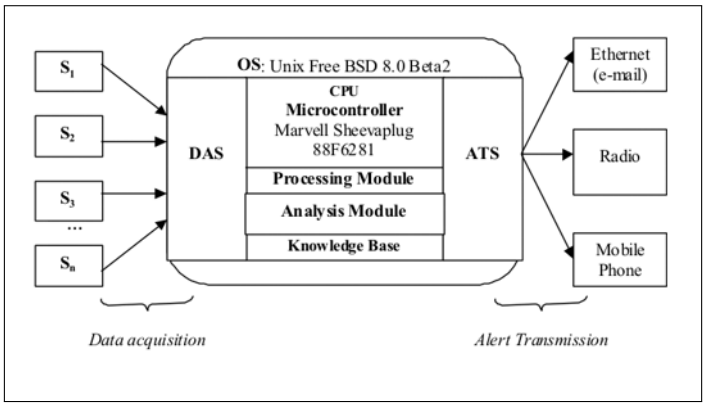


Figure 1: The architecture of the system

As interfaces with the environment, the intelligent system has a data acquisition system (DAS), and an alert transmission system (ATS), both systems being developed within the microcontroller Marvel 88F6281. The data acquisition system takes the measurements of different parameters that are monitored from a variety of sensors (S_i) and measurement devices (e.g. rain gauges). The alert transmission system will send the hydrologic alert code through a communication channel (internet, radio or mobile phone) to the decision factors. We have included also in the microcontroller a data processing module, a knowledge base, and an analysis module, those based on the knowledge of the intelligent system will provide the hydrologic alert code to ATS.

The microcontroller 88F6281 is built under the SoC form, i.e. system on chip, and has a large variety of interfaces as shown in figure 2. Its main destination is the mobile telephone market due to its multimedia facilities.

We have compiled on this system the Unix FreeBSD 8.0 operating system, available in the beta version, due to its stability and simple programming (see [4]). This OS provides support for the ARM9 architecture, and it was compiled with the specifications of the 88F6281 microcontroller according to (see [6]). Moreover, the OS that was chosen is a good option for embedded real time applications (see [1]).

The hydrologic alert transmission could be sent by email, SMS (on the mobile phone) or by radio transmission (e.g. to a PDA - Personal Digital Assistant).

3 Hydrographic monitoring and analysis system

The monitoring and analysis system of a catchment basin structure consists by a complex of sensors that acquire information about the monitored river parameters, as well as rainfalls in the catchment area. Another important set of information is given by the catchment physiographic

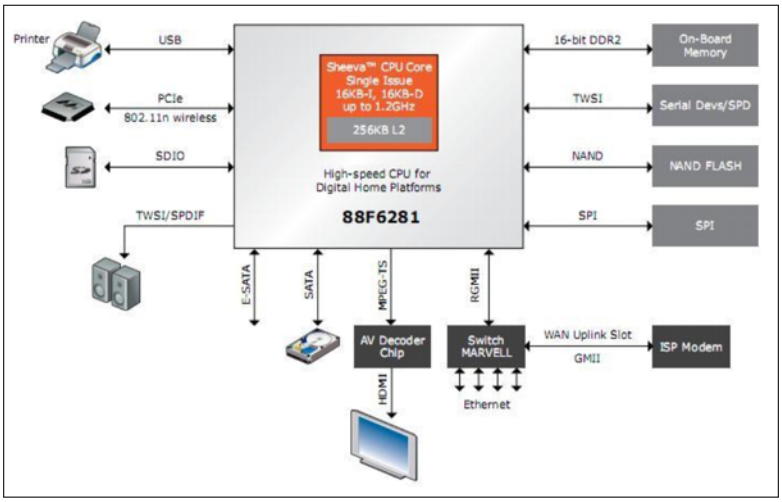


Figure 2: The Marvell Sheeva 88F6281 interfacing options

characteristics that influence the flow regime during the flood (see [10]). Our hydrographic monitoring and analysis system was designed for the monitoring of the evolution process of rainfall transformations into collecting streams by the hydrographic network from one river basin. The system is able to acquire and store the database information on river water level (H), current river speed (v), water flow rates (Q), the rainfalls level and frequency (X) etc and transmit them to higher level of the analysis tasks. After the information processing is done by using the analysis algorithms (see [7]), the system has the ability to make decisions and to transmit alerts to the Committee of Emergency Situations in case of flood risks.

The input data used by the system can be obtained from parameters measurements time-series. The analyzed parameters are as follows:

- the river water level (H - cm), measured by the gauging stations that are located in different regions of the catchment;
- the rainfalls quantity (X - l/m^2);
- the river water velocity (v - m/s);
- the air temperature (T - ^oC);
- the depth (a), width (L) and slope (p) of the river;
- the thickness of the ice/snow (G - cm).

The output of the system is given by the numerical expressions that characterize the interdependence of the hydrological process and the factors that determine it: the maximum flow rate (Q_{max}) and the average precipitation. The main steps followed by the hydrographic monitoring and analysis system are given bellow.

1. parameters_measurement (H, Q, X, v, T, a, l, p, G);
2. parameters_processing (H, X);
3. parameters_transmission ($X_{average}$);
4. parameters_analysis (); // use of the knowledge base

5. IF flood_risk = yes THEN

   (a) hydrologic alert code transmission.

The parameters measurement step requires periodic measurements of specific parameters such as water flow, water level, river flow rate, temperatures. The equipment used for performance measurements and processing of hydrological parameters of the water courses are the gauging stations. The instrumentation used for parameters measuring vary from groom gauging, pluviometric devices to measure the rainfall, river ratchets gauging for speed measurements, to automatic instrumentation that can measure most of the hydrological parameters that are monitored. The parameters processing step involves the form factor computing for each sub-catchment coefficient of flood form, and the determination of the coefficient of leakage in the warning sub-basins. The parameters transmission step is realized through a communication channel, such as internet, radio or mobile phone. The parameters analysis step is the most important step because the subsequent correct decisions are taken in case of flood waves. It consists in the following three steps:

- determination of the average level of precipitations;
- calculation of the maximum flow in each sub-basin of the downstream warning sub-basins;
- based on the calculations made in the previous two steps, the system analyses the risk of flood production by using the rules from the knowledge base that is incorporated in the intelligent system; the rules are given the hydrological alert code (yellow, orange and red).

As an example, the main characteristics of the knowledge based system are presented in Table 1.

Table 1: The rules of the knowledge base (selection)

RULE 1	IF flow < attention AND precipitations < $40\text{l}/\text{m}^2$ THEN hydrologic_code = green;
RULE 2	IF flow ≥ attention AND flow < alert AND precipitations < $40\text{l}/\text{m}^2$ THEN flood_risk = YES AND hydrologic_code = yellow;
RULE 3	IF flow ≥ attention AND flow < alert AND precipitations > $40\text{l}/\text{m}^2$ THEN flood_risk = YES AND hydrologic_code = orange;
RULE 4	IF flow ≥ alert AND flow < danger AND precipitations < $40\text{l}/\text{m}^2$ THEN flood_risk = YES AND hydrologic_code = orange;
RULE 5	IF flow ≥ alert AND flow < danger AND precipitations > $40\text{l}/\text{m}^2$ THEN flood_risk = YES AND hydrologic_code = red;
RULE 6	IF flow ≥ danger THEN flood_risk = YES AND hydrologic_code = red;

The precipitations are occurring predominantly as rain or snow. The fall of rain is the largest factor contributing to increased river flows. Snow versus rain is the second source of precipitation. Melting snow during spring has a considerable impact on river basin management. Tracking the level and the periodicity of precipitation involves measuring them by using existing sensors and transducers gauging stations. Warning about the risk of flood is used to take appropriate decisions from the analysis of all information related to the maximum river flow. If the analysis step indicates a possible flood production then the prevention system will send signals to the members of the Emergency Situations Committee. The hydrological alert (containing the alarm severity code of flooding) can be transmitted online to PDAs, smart terminals with a GPS and

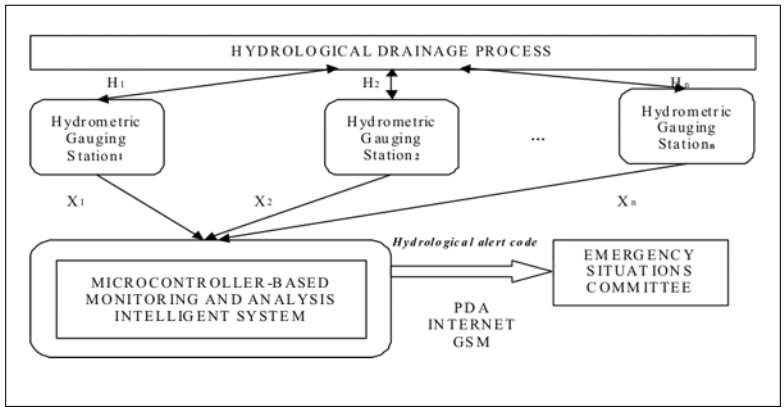


Figure 3: The structure of the monitoring and analysis system

GPRS (General Packet Radio Service) or Internet. In figure 3 it is shown the structure of the monitoring and analysis system for a catchment river basin.

$X_1, X_2, \ldots, X_n$ are the rainfalls in each sub-catchment;
$H_1, H_2, \ldots, H_n$ are the rivers level.

The system meets three important functions:

- the function of monitoring the evolution of the hydrographic basin - real time measuring of the river level and rainfall level and periodicity;
- the analysis function - parameters processing and comparing them with the references parameters of the catchment;
- the decision and warning function.

4 Experimental results of flood alerting case study

As a case study we have experimented the microcontroller-based intelligent system for flood alerting in the Prahova catchment basin presented in figure 4. The hydrographic network forms a highly developed basin in a palm form flowing NW-SE. The main rivers that compose the Prahova sub-basin are the Prahova river and its main tributaries: Azuga, Doftana, Teleajen, and Cricovul Sarat. The water recources in the Prahova country have increased significantly due to two large lakes, Paltinu (Doftanei Valley) and Maneciu-Streams (Teleajen Valley). The Prahova river is the largest collection of water in the Prahova county, with a length of 193 km, of which the first 6 km and the last 16 km are located in the counties Brasov and Ilfov.

The Prahova river basin is characterized by three types of climates: mountain, hill, and plain. The annual quantity of precipitation is 1000-1400 mm in the mountains, 500-1000 mm in the hills, and 550-600 mm in the plain. Summer rainfall is more abundant, where the flood may occur in the gauging stations deepened Moara Domneasca and Adancata. For the visualization of the hydrological process evolution in time we have used the program Multi Router Trafic Grapher (mrtg) that generates graphics as a function of time. An example of Prahova river level evolution and flow in 24 hours during 7 days is given in figure 5.

For the Prahova hydrometric station the attention (ATC), alert (AC) and danger (DC) cotes for the rivers level are the following: ATC = 250 cm, AC = 350 cm and DC = 400 cm and for water flow are: ATC = 105 m^3/s, AC = 230 m^3/s and DC = 320 m^3/s. Our research goal was to use the microcontroller-based intelligent system, that was presented in the previous sections,

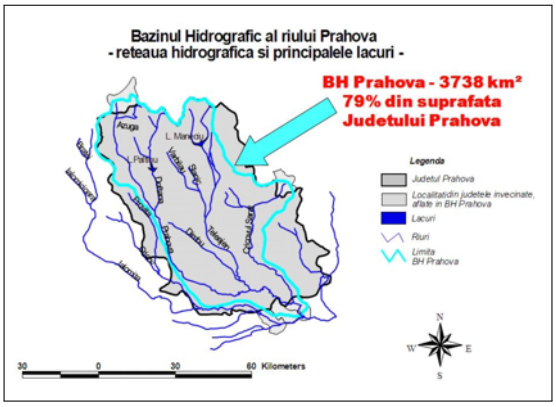


Figure 4: The Prahova catchment basin

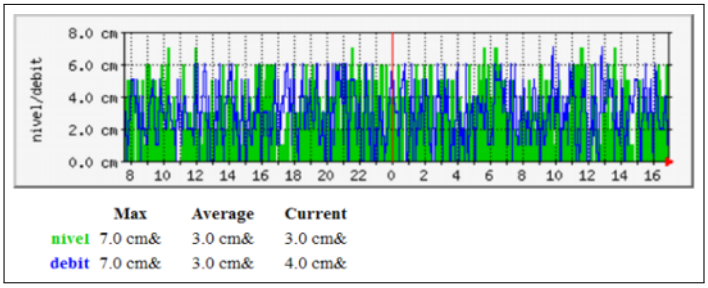


Figure 5: The evolution of Prahova river level and flow in 24 hours during 7 days

for the Prahova river basin with the purpose of sending warning in real time to the Emergency Situations Committee, and to the possible affected population in case of flood waves.

In the experiments that were run so far, the data acquisition was simulated. The experiments made so far used for alert transmission two Motorola MX300 radio transmission stations, those were modified to work in the 146 MHz bandwidth. Figure 6 shows the Motorola MX300 radio stations that were used.

Figure 6: The two Motorola MX300 radio stations

The main characteristics of the radio stations are that they can be automatically activated when they receive a radio signal through the incorporated microphone or through the microphone input. Therefore, the sound output of the microcontroller-based system was connected to the microphone input of the transmitter and according to the hydrologic alert code provided by the analysis module will run audio alert records according to following script:

The evolution of the simulated hydrological process is shown on the microcontroller-based

```
♯ !/bin/bash
♯ script receives argument via parameter {$1} which is received at startup time.
♯ {$1} may be as following: 'green', 'yellow', 'orange', 'red' and are sent on via the
main program
♯ alert on system console echo "Alerting code - {$1}. Activating radio transmitter"
♯record alert via syslog (unix system logger) echo " 'date |cut -c1-19': Receiving
alerting code {$1} |logger -s -
♯playing one of the audio files: 'green.wav', 'yellow.wav', 'orange.wav', 'red.wav'
/usr/local/bin/esdplay -s localhost {$1}.wav
```

intelligent system screen as shown in figure 7. Three hydrometric gauging stations were included in the simulation, Buşteni, Cîmpina and Prahova, all for the Prahova river basin.

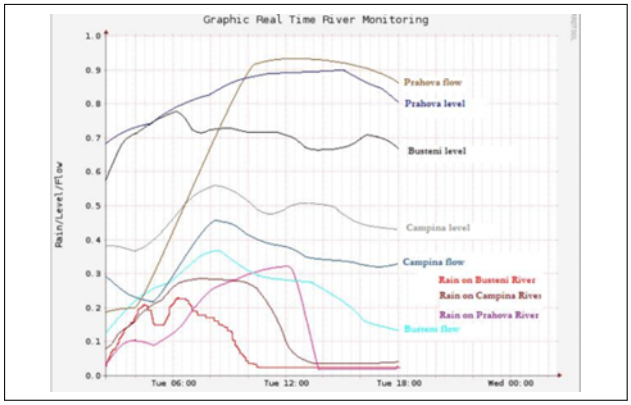


Figure 7: Rain/Level/Flow evolution for the case study

Figure 8: The experimental system

Figure 8 shows the experimental system. In the experiments made, for each hydrological alert code it was sent via radio communication the corresponding audio alert signal (attention, alert and danger), according to the results of the analysis step.

5 Conclusions

As most of the hydrological processes require real time monitoring and online alerting, we have developed a microcontroller-based intelligent system that can be connected to various sensors and measuring devices for hydrological parameters measurements, and can transmit online hydrological alert codes to the decision making factor (such as Emergency Situations Committee that exist in each county from Romania). The system can be adapted to any hydrographic basin with minimal changes in its configuration. At the time the system was developed, all the data coming in was simulated, but at any time its knowledge base system can be fed with data coming from real sensors. During long time hard condition tests under simulated air humidity, simulated power fluctuations, while running a CPU stress testing program and also our simulation software, the intelligent system kept its main characteristics of flexibility and robustness, as no indicators of system failure have occurred at Unix FreeBSD 8.0 operating system kernel level. As a future work we shall extend our experiments by using simultaneously with the radio hydrological alert transmission other communication channels such as the internet (by sending an e-mail) and the

mobile phone (by sending a SMS). Also, we shall analyze some complex hydrological situations by taking into account the interdependencies between different rivers flows and levels from a hydrographic basin and the meteorological forecasts, improving flood prevention due to an early hydrological alert sent to the decision making factors in order to take more efficient prevention measurements.

Bibliography

[1] D. Abbot, *Linux for embedded real-time applications*, Newnes, USA, 2003.

[2] K. Arnold, *Embedded controller hardware design*, LLH Technology Publishing, USA, 2001.

[3] S. R. Ball, *Embedded microprocessor systems - real world design*, 3rd edition, Newnes, USA, 2002.

[4] BSDi, *The FreeBSD Handbook*, Berkeley University of California, USA, 2000.

[5] V. Buruiana, Wireless Seismic Sensor, *Proceedings of Process Control Symposium SPC-09*, Ploiesti, 2009.

[6] V. Buruiana, *Experimental Research regarding the Development of a Microcontroller-Based Intelligent Monitoring and Alert System*, Research Report, University Petroleum-Gas of Ploiesti, Department of Informatics, 2009.

[7] A. Matei, *Intelligent Algorithms for Hydrographic Monitoring, Analysis and Prediction*, Research Report, University Petroleum-Gas of Ploiesti, Department of Informatics, 2009.

[8] M. Oprea, A. Matei, E. Petre, Agent-based Modeling of a Dam Monitoring System, *Proceedings of 17th International Conference on Control Systems and Computer Science - CSCS 2009*, Politehnica Press, 509-514, 2009.

[9] D. Popescu, R. Varbanescu, A. Iordan, S. Arghir, Interconnection of mobile sensor nodes for alert system applications, *Proceedings of CSCS 2009*, Politehnica Press, Bucharest, 249-254, 2009.

[10] I. Watson, A. D. Burnett, *Hydrology - An Environmental Approach*, Taylor & Francis, CRC Press, 1995.

[11] http://www.tgdaily.com/hardware-opinion/41525-marvells-plug-computer-a-tiny-discrete-fully-functional-5-watt-linux-server

[12] ftp://ftp.freebsd.org/pub/FreeBSD/ISO-IMAGES-i386/8.0/8.0-RC2-i386-disc1.iso

Int. J. of Computers, Communications & Control, ISSN 1841-9836, E-ISSN 1841-9844
Vol. V (2010), No. 5, pp. 852-861

Agent Technology in Monitoring Systems

B. Pătruţ, C. Tomozei

Bogdan Pătruţ, Cosmin Tomozei
"Vasile Alecsandri" University of Bacău,
Romania, 600115, Calea Mărăşeşti, 157
E-mail: bogdan@edusoft.ro, cosmin.tomozei@ub.ro

Abstract: The aim of this paper is to make a brief presentation of the results obtained by the authors regarding agent technology application in distributed monitoring systems development. Consequently, we would like to present MAgeLan and ContTest as monitoring systems based on intelligent agents technology. Formal aspects regarding intelligent agents will be mentioned. Quantitative issues concerning efficiency, maintenance and reengineering are also to be taken into account.
Keywords: intelligent agents, distributed systems; monitoring systems; hyper-encyclopedia, reengineering.

1 Intelligent agents in distributed computing

Software applications are nowadays in position and necessity to solve some problematical situations by their own, in order to save time and reduce cost. Intelligent agent integration in distributed computing proved to be an important way of achieving this objective in the last years. Furthermore, the union distributed computing with intelligent agents theoretical perspective offers a good practical basis to distributed artificial intelligence [2]. Intelligent agents have to be reliable, robust in order to offer accuracy in the results, in dissimilar, open or unpredictable environments [7], [8]. Software agents are situated in particular environments and capable of autonomous actions, in order to fulfill their objectives. The concept of autonomy and the fact that intelligent agents are enriched with it presume that human action is minimized. In [8] it is affirmed that the agent is an entity which can perceive the environment by sensors and acts in order to realize its objectives by effectors. On the other hand, distribution of hardware, software and data offers the possibility for the agents to be replicated on diverse nodes on the computer network.

2 Developing Students' Metacognitive Competences

Metacognitive skills development is an important formative intellectual object in education of the students, as reaching this level involves a route through effective education, appropriate to each one in particular [7]. Metacognitive skills suppose that students are aware of their own cognitive activity, i.e. learning activity, and self-adjustment mechanisms consisting in cognitive controls (rules, procedures, strategies).

Next we will introduce the notions of environment, agent, s-agent, necessary to develop some intelligent systems as those described in chapter 5 of this article. The following theoretical concepts are different from the FIPA standard [9] ones or from the concepts defined by authors like Wooldridge in [8], and are necessary for the implementation of the systems MAgeLan and ContTest.

The following statements are to introduce theoretical concepts regarding intelligent agents and distributed technology.

Definition 1. We use the name of environment for a set of elements $E = \{e_0, e_1, e_2, e_3, ..., e_n\}$ among which there is a relation of partial order marked with "$<$". We use the notation $e \leqslant f$ for the fact that $e < f$ or $e = f$, respectively $f > e$ if $e < f$.

The environment can be, at a certain point, in a certain state e, which we will express by $st(E) = e$. At first, the environment leaves an initial state e_0, for which $e_0, e_1, \forall i \in \{1, 2, ...n\}$. The state e_n is called final state for which it is considered that $e_i, e_n, \forall i \in \{1, 2, ...n\}$.

Definition 2. We use the name of **agent** for a triple of the type (3.1) $A = (S, s_0, R)$ where S is a finite set of states, s_0 in S is called the agent's initial state, and R is a set of evolution rules.

If agent A is in state s, then we express this by $st(A) = s$. Among the states of S there is a special state marked with λ. At first $st(A) = s_0$, and when $st(A) = \lambda$ we say that the agent is inactive. For the rest, the agent is active. The rules in R are of the type (1) or (2):

$$r_1 = (A, s, e) \rightarrow (A, t, f) \tag{2.1}$$

$$r_2 = (B, s, e) \rightarrow (B, t, f) \tag{2.2}$$

Rule (1) states that if $st(A) = s$, $st(E) = e$, then $st(A)$ becomes t and $st(E)$ becomes *f*, if $e \leqslant f$. The second rule (2) states that agent A ceases its implementation ($st(A)$ becomes λ), transferring the control to agent B, for which $st(B)$ becomes t, and the environment remains in the same state.

If $st(E) = e$ and there are two agents A and B with $st(A) = a \neq \lambda$ and $st(B) = b \neq \lambda$ and $(A, s, e) \rightarrow (C, z, f)$ and $(B, t, e) \rightarrow (D, z', f')$, then we will consider $st(E) = max(f, f')$ if f and f' are comparable, one of them respectively, if f and f' are not comparable, and $st(A) = 0$ and $st(C) = z$ if $f < f'$, respectively $st(B) = 0$ and $st(D) = z'$, if $f < f'$.

This can be generalized for more active agents.

Definition 3. We use the name of **s-agent** for an n-uple of the type (3)

$$S = (C, A_1, A_2, ..., A_n) \tag{2.3}$$

where C is an agent called **coordinating agent**, and $A_1, A_2, ..., A_n$ are agents corresponding to the definition above that are called **effectors** or **atomic agents**. The coordinating agent will interact directly with the user and the architecture and its functionality depends on the concrete implementation (the examples will be offered in the following chapters).

Because inside an s-agent, the agents transfer the control form one to another, an s-agent behaves like a communicative multi-agent system.

Definition 4. Let there be $S = (C, A_1, A_2, ..., A_n)$ an s-agent. If $s_1, s_2, ..., s_n$ are the states of the atomic agents that make up S (without the coordinator C), we then say that $(s_1, s_2, ..., s_n)$ is the state of S (at a given moment).

Definition 5. Let there be $S = (C, A_1, A_2, ..., A_n)$ an s-agent in the environment E, that cannot be modified by the user. If the initial state of S is of the type $(0, \lambda, \lambda, ..., \lambda)$ we say that S is a normal s-agent (s_{01} marks the initial state of the agent A_1).

Directly, we obtain the following lemma:

Lemma 1. In a normal s-agent, $\{s_1, s_2, ..., s_n\} = \{s, \lambda, \lambda, ..., \lambda\}$ is enacted, with $s \neq \lambda$, at any point of time t.

Proof. The s-agent being normal, it follows that its initial state (at the moment t_0) is $(0, \lambda, \lambda, ..., \lambda)$, therefore $\{s_{01}, s_{02}, ..., s_{0n}\} = \{s_{01}, \lambda, \lambda, ..., \lambda\}$, and $s_{01} \neq \lambda$.

Let us suppose that the state in the moment t_i is $\{s, \lambda, \lambda, ..., \lambda\}$, with $s \neq \lambda$. This means that an active A_i agent exists and that it is unique, with and $st(A_i) = s \neq \lambda$ and $st(A_j) = \lambda, \forall j \neq i$. If there is a relation of evolution of the type $(A_i, s, e) \rightarrow (A_j t, f)$, then by applying this relation of evolution, we will obtain in the moment $t_{i+1} : st(A_i) = \lambda$ and $st(A_j) = \lambda$, with $t \neq \lambda$. Therefore the state of the s-agent will become (no matter the situation) $(\lambda, \lambda, ..., \lambda, t, \lambda, ...\lambda)$, with $t \neq \lambda$, on the position j, therefore. If there is no relation of evolution with (A_i, s, e) on the left side, then the s-agent will remain in the state $\{s, \lambda, \lambda, ..., \lambda\}$, with $s \neq \lambda$.

The fact that the agent will remain in that certain state will be called **blockage**, a notion that we will eventually formally define.

Definition 6. We use the name of **multi-agent monitoring system (MAMS) of the environment** E, based on s-agents (or on groups) for a triple of the type (4)

$$MAMS = (Sa, L, E) \tag{2.4}$$

where Sa is a set of s-agents, having the same structure, E is the environment within which these exist and are implemented, and L are communication or linkage rules of the type $C_i \rightarrow C_j$, where C_i and C_j are coordinating agents of some s-agents from Sa.

The communication relations among the s-agents form an oriented graph depending on the architecture and the concrete implementation of the multi-agent system, as we will see in the following chapters. Our interest lies in some evolution rules of the environment within an s-agent and the structure and graphic representation of an s-agent.

A MAMS containing only normal s-agents is called a **normal MAMS**.

Definition 7. The purpose of an s-agent is to take the environment E to a state as close as possible to its final e_n state. If the relation (5) can be obtained, we then say that the aim of the s-agent can be carried out.

$$St(E) = e \wedge \neg \exists f \in E, f \neq e_n : e < f \tag{2.5}$$

By extension, we can say that the aim of MAMS is reached if all the targets of the component s-agents are achieved.

Definition 8. Two rules of evolution of the type $r_1 = a \rightarrow b$ and $r_2 = b \rightarrow c$, where a, b, and c are triples of the type of those in (1) or (2), they are called adjacent.

Definition 9. Let there be $r_1, r_2, ..., r_k$ a line of adjacent rules of evolution, two by two. We will use the notation (6) and we will call this relation as the derivation relation (see (6)).

$$r_1 \Rightarrow r_k \tag{2.6}$$

Within an s-agent, the following results are obvious:

Proposition 1. If st(E)=e, there is an agent A with st(A)=$s \neq \lambda$, (A, s, e) $\Rightarrow$ (A', s', e') and there is no $f \neq e_n$, so that $e' < f$, then the aim of the s-agent can be reached (with the environment in state e'). (From the very beginning, we have noted the final state of the environment with e_n). Most of the times, set E is not fully ordered. If, however, a fully ordered relation "<" is found, then the following sentence is enacted.

Proposition 2. If the relation "<" is fully ordered, then the purpose of the s-agent can be reached with the environment in the final state (e_n) if there is a derivation of the type $(A, s_0^A, e_0) \Rightarrow (B, t, e_n)$. Proposition 2 affirms that the purpose of the s-agent can be reached if there is a derivation which leads the environment to the state en, starting from the environment's initial state e_0 and the initial state of any agent (A), without the user's intervention (in the case of a fully ordered relation "<").

If $|\{A; st(A) \neq \lambda\}| = 1$, for any $e = st(E)$ (therefore a single agent is active at a given moment, as in lemma 1), then the MAMS operates sequentially. This is what happens most of the times.

We consider that the environment E modifies its current state in two cases:

• as a result of applying an evolution rule, by one of the system's agents;

• as a result of the direct intervention of a human user. We can also consider, in some exceptional cases, that the state a certain agent is in can be modified by the evolution rules as by the user as well. Therefore, we cannot hold control over what is going to happen, or how the state of the environment is going to evolve within a certain interval of time. If, ideally, the human user cannot randomly modify neither the environment E nor the current state of the agents, then the system is entirely deterministic.

3 Blockages and Infinite Cycles

Our interest does not lie simply in building absolutely sequential systems or deterministic systems, but in specifying in the best possible way the evolution rules so that blockages cannot occur.

Definition 10. If relation (7) is certified, then we say that the s-agent is under **blockage**.

$$\exists A \in S : st(A) = s, st(E) = e, e \neq e_n \wedge \neg\exists(A, s, e') \rightarrow (B, s', f), f, e' \neq e_n \tag{3.1}$$

Therefore, we say that an s-agent is under blockage when an atomic agent of the system gets into a state s, and the environment is in a non-final state e, and there is no evolution relation that can allow for the agent's passing out from the state s, although the environment is suddenly modified by the user, into another non-final state e'. In (7) f and e' are certain non-final states of the environment (e' and f may also be even e), and s' a certain state of a certain agent B from the s-agent, being even possible for B to be A.

In other words, there is no evolution rule, with s on the left side, which can lead to another state of A or within another agent B, no matter the evolution of the environment E.

Definition 11. In an s-agent a derivation of the type $(A, s, e) \Rightarrow (A, s, e)$ is called **infinite cycle**.

This occurs when, if the user does not intervene through modifying the environment, the execution of the s-agent's agents cycles infinitely, without the environment reaching the final state. In this case, the user may intervene to take the MAMS out of the cycle.

Blockages and infinite cycles can be identified easier if we represent the s-agents and the MMS. Graphically, a multi-agent monitoring system (MAMS) can be represented in the shape of a graph oriented thus (figure 1):

• optionally, more s-agents are represented in the shape of some polygons or other geometrical figures, containing more s-agents, the internal structure being represented only for one of them;

• optionally, the links among the s-agents will have the shape of some curves; graphically only one s-agent will be represented, because all s-agents are considered to have the same internal structure;

• the generic s-agent will be represented through a geometrical figure where all atomic agents are represented;

• the atomic agents are represented through some rectangles labeled with their names; inside those squares there will be circles representing the different states of the respective agent;

• each state will represent a point in an oriented graph, where the edges are the evolution relations, labeled with a pair of clues for the states of the environment: the starting state and the state that is finally reached.

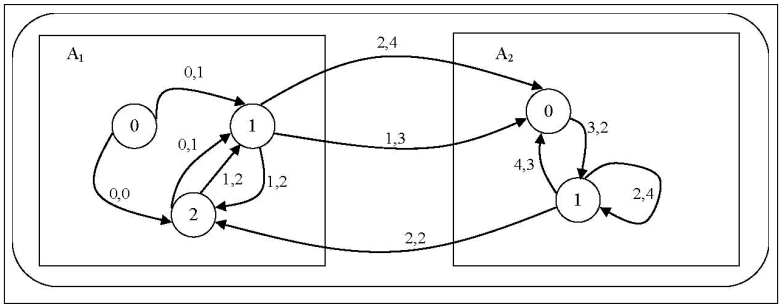


Figure 1: A simple s-agent containing a blockage in A_1 and an infinite cycle in A_2

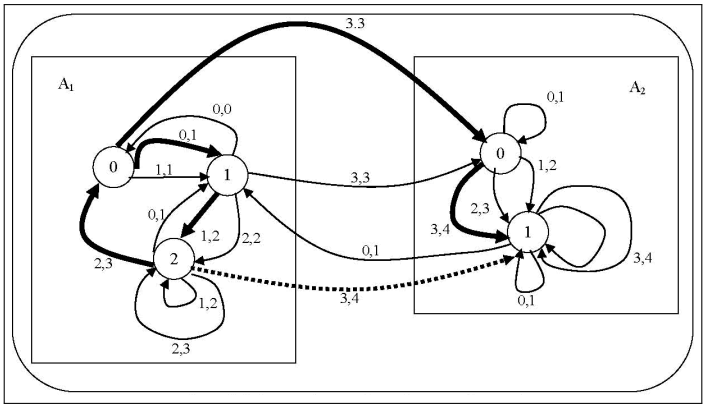


Figure 2: A normal s-agent and its reduced s-agent

This is an ideal example of MAMS functioning, illustrated by the bold arcs in figure 2. Obviously, the other arcs except the bold ones are useless in this graph, because they will never be passed through. If, however, the user interferes, for example the moment the MAMS is in state 2 from the atomic agent A_1, modifying the state of the environment from e_2 to e_3, then the evolution $(A_1, 2, e_3) \rightarrow (A_2, 1, e_4)$ will occur, expressed in figure 3 through the dotted arc. We can obviously "endow" an s-agent with many evolution rules. Inside a co-operating working environment, different users will introduce different rules of evolution. The question is, if under the circumstances of some normal agents, some of those rules will ever be applicable, will ever come into action. It is as if we had a program with functions or pieces of codes which are not resorted to by anywhere, cannot be touched, or an expert system with production rules whose part of the premises will never be fulfilled.

Thus, we will show that normal agents can be reduced to other normal agents, on the basis of an algorithm, so that certain evolution rules could become useless and could be eliminated from the system, the later behavior being not affected. On the contrary, the systems become more simple. By an s-agent' behavior we mean the applicability of the rules of its component agents. Therefore, if a certain evolution rule could ever be applied, it will be a part of the s-agent's

behavior, and if not, then it will not be a part of this behavior.

Definition 12. Let there be S_1 and S_2 two normal s-agents, functioning simultaneously within the same environment E. If the behavior of S_1 is identical with the one of S_2 at all moments, meaning that the same evolution rule is applied both in S_1 and S_2, we say that the two agents are **equivalents**. Equivalence also refers to the situations of blockages or infinite cycles.

4 Reducing the Normal S-agents

We will further consider a normal s-agent. We will show that the following theorem is enacted.

Theorem 1. A normal s-agent S can be reduced to a normal s-agent T that is equivalent with S, by eliminating some evolution relations, according to the following algorithm.

Reduction algorithm(s-agent S, s-agent T);
{Local Variables: A, s, e, A_i, blockage, obj_realized, M, $T = \emptyset$;
STEP 1: for each A_i *from S*;
if $st(A_i) \neq \lambda$ *then* $(A, s, e) \leftarrow (Ai, st(A_i), st(E))$;
end for;
STEP 2: $M = \emptyset$; *cycle = False; blockage = False; obj_ realized = False*;
while (not cycle) and (not obj_ realized) and (not blockage);
{*M = M* $\cup$ *(A, s, e)*;
if in S exists edge (A, s)$\rightarrow$ *(B, t) labeled with (e, f) then* {;
Add T in Vertex s in agent A; Vertex t in agent B;
Add edge (A, s) $\rightarrow$ *(B, t) between s and t; Label edge (A, s)* $\rightarrow$ *(B, t) with (e, f)*;
if f = e_n *then obj_ realized = True else*;
if $\neg \ni f' > f$ *and* $f' \neq e_n$ *then obj_ realized = True*;
else if (B, t, f) $\in$ *M then cycle=True* };
else blockage=True } }.

Proof. According to lemma 1, the s-agent S has always a state made up of n-1 inactive states (λ) and a single active state belonging to one single agent. This means that the I^{st} part in an algorithm will determine a unique agent A with a singular active state s, and e will be the current state of the environment. The cycle "while" from the second part ends because inside it all possible cases are brought into discussion on the "if-then-else" branches. Building the s-agent T is done using the instructions in the framed part. We have used the graph representation of S and T. The instructions inside the dotted frame realize the inclusion in T of the agents A and B (unless they are already there), of their states s and t (unless they are already there) and of the relation of evolution $(A, s, e) \rightarrow (B, t, f)$.

The algorithm will extract out of the graph of S a chain T, until the final state has been reached, the "largest" final state, a blockage will occur or an infinite cycle will be entered into. T will thus be a normal s-agent, because if, at a given moment, according to lemma 1, $\{s_1, s_2, ..., s_n\} = \{t_1, t_2, ..., t_n\}$ and $s_1, s_2, ..., s_n = \{s, \lambda, \lambda, ..., \lambda\}$ is enacted, it also results that $\{t_1, t_2, ..., t_n\} = \{s, \lambda, \lambda, ..., \lambda\}$.

Observation 1. The previous result is equivalent, in the theory of formal and automatic languages, with the elimination of inaccessible and unusable states from a finite deterministic automate [1].

Corollary 1. An s-agent S gets blocked if and only if its reduced agent T gets blocked.

Corollary 3.2. An s-agent S has an infinite cycle if and only if its reduced agent T has an infinite cycle.

According to the algorithm in theorem 1, the s-agent T is equivalent to the s-agent S, therefore the implementations of S and T are identical. If S gets blocked, then T gets blocked and reciprocally. If T has an infinite cycle, then S has the same cycle too (though it may have some others as well).

The algorithm in theorem 1 can be completed with a sequence of pseudo code, in order to display the configuration in which both S and T get blocked or enter an infinite cycle.

5 Practical Implementations

One first example of MAMS is that where the environment is given by an intelligent hyper-encyclopedia. We described in [4] the concept of intelligent hyper-encyclopedia, which, intuitively means an encyclopedia distributive developed by several users and which uses artificial intelligence tools. For the time being, we only mention the fact that the environment E is made up of the content of the hyper-encyclopedia. By the content of hyper-encyclopedia, we understand the totality of its entries. Each entry is defined by a word, a definition of this and the references to the other related entries. The states of the environment are states of the content of this hyper-encyclopedia at a certain moment, that is the totality of the entries introduced by the respective moment. We certainly have an initial state e_0 when the hyper-encyclopedia contains nothing, but also a final state e_n, when the hyper-encyclopedia is considered by the users to be definite, complete. The order relation "<" between the states of the hyper-encyclopedia is a relation of partial order, because we can say that one state is smaller than the other only if the contents of the encyclopedia at a certain moment is included in the contents of the encyclopedia at another (later) moment. However, we cannot compare two states in which the hyper-encyclopedia has completely different contents, or has only one common part. The aim of the MAMS is to develop the hyper-encyclopedia as much as possible.

The s-agents will be local multi-agent systems. Each user who contributes to the enlargement of the hyper-encyclopedia has his/her own s-agent. In [3], [4] we presented a real architecture of such an s-agent. Within each s-agent there is a series of atomic agents bearing different tasks. We presented the agents responsible for monitoring the Microsoft Word editing activity of the content of the hyper-encyclopedia, while we have dealt with the agents that monitor the web search by using the Internet Explorer browser.

The encyclopedia behaves intelligently in these directions: it raises certain questions in order to generate new information, based on the existing ones; it answers certain questions from the users in order to provide them with the information necessary to generate new knowledge; it adapts itself to the users' language and to their way of working (consults and modifies the databases or interacting with the others; it provides an intelligent mediation in the communication among users, so that these may be able to learn from each other, and the encyclopedia from them. The intelligent hyper-encyclopedia - having so many "human" traits, even in its limited alternative - will have a network structure where each knot will detain a local data basis, a human user and a system of intelligent agents, thus: an agent which observes the user's behavior while consulting the encyclopedia and adapts itself to it; an agent which observes the user's behavior while updating the local data basis (modifying, deleting, adding data or generating rules) and acts accordingly; an agent which alone modifies the local data basis in order to adapt it to the user from that knot (in order to ease his/her work). All the above mentioned agents communicate among themselves and cooperate to learn from each other and to modify their own local data bases, but also the way in which they interact with the users from the knots. Communication among agents and human users will often be achieved by natural language, through certain

conversational agents. They are also capable of "metamorphosis' in order to adapt themselves to the users, or, in case this cannot be done, agents can move from one knot to another. The MAgeLan [9] operates to solve tasks characteristic of artificial trainers which assist (monitor, help and manage) several students (users) in developing an intelligent hyper-encyclopedia.

One rule of evolution of the type $(A, s, e) \rightarrow (A, t, f)$ can be thus interpreted: if the hyper-encyclopedia is in state e, and the agent A notices that the user searches the web for the term s, then the hyper-encyclopedia passes to the state f (its content is improved by creating some new links between the articles or by some other means) and the user is suggested to search for the term t. Obviously, the user may or may not follow the system's piece of advice.

A second example is given by an instructive system, ContTest [5]. The environment is represented by the knowledge of the learner. In this case, first of all we assume that the student knows nothing, therefore the environment lies in the state e_0. When the student has achieved what we had had in mind the environment is in its final state. The aim of the MAMS is to make the student know everything, or at least as much as possible.

If we assume that the s-agents are multi-agent systems, each of them responsible for one lesson or one learning unit, then each s-agent will contain atomic agents which will have specific pedagogical tasks. For example, an agent will be responsible for teaching some content, another will be responsible for offering examples, another one for generating tests. The states of an s-agent are stages in teaching, illustration, testing and so on. Let us consider the agents as being didactic actions. For example, agent A can mark the teaching, and agent B, the testing.

One rule of evolution of the type $(A, s, e) \rightarrow (A, t, f)$ will communicate to us something like this: if we have taught (A) the notion s, and the student's level is e, then we are to teach the notion t next and the student will reach level f. One rule of evolution of the type $(A, s, e) \rightarrow (B, t, f)$ will communicate to us that if we have taught the notion s and the student's level is e, we can then pass on to testing (agent B) the student through exercise t and we expect the student to reach the (superior) level f.

6 Reengineering MMS Systems

Multi-agent monitoring systems are being submitted to the process of reengineering in order to achieve new objectives. Software reengineering proves to be [6] valuable in large-scale application development. By applying this procedure, we will keep a valuable software system, such as MAgeLan hyper-encyclopedia in working condition for a longer period of time. The cost is significantly reduced and the new tasks are efficiently accomplished.

If we take into consideration the application's functionalities, the following formula (8) will describe the elements regarding the connection of the modules that remain unchanged. Elements which are to be modified or deleted will be identified as well, for the entire application to become a powerfully connected system.

$$Ir = \bigcup_{i=1}^{n} Funct_{i0} + \bigcup_{i=1}^{n} Funct_{i1} - \bigcup_{i=1}^{n} DelFunct_i \tag{6.1}$$

Ir represents the indicator of reengineering, $Funct_{i0}$ is the functionality in the initial moment, $Funct_{i1}$ is the functionality in the present moment added due to reengineering, consequently $DelFunct_i$ is the functionality eliminated by reengineering, and n represents the number of functionalities.

ContTest (figure 3) significantly evolved due to reengineering, the agents being enriched with new teaching and learning methods and procedures. Therefore, evolution in objectives brought about evolution in the functionalities of agents. Pedagogical tasks of agents have been

considerably improved. Consequently, higher educational objectives and elevated results in the evaluation process have been obtained.

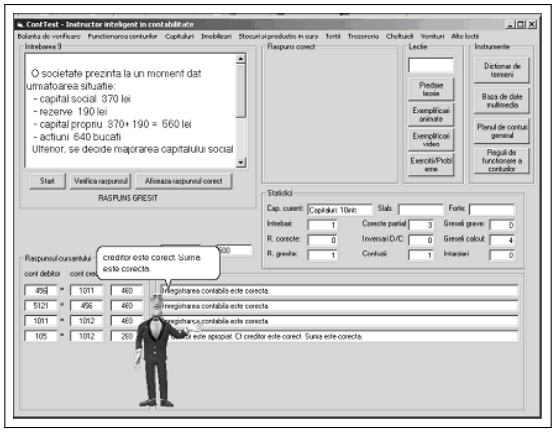


Figure 3: Validation of answers

We may state that each software project has been enhanced with a higher level of quality, evolving correspondingly with the evolution in demands, by integrating reengineering.

7 Conclusions

Our aim was to summarize the results obtained in development and maintenance of distributed applications and multi-agent monitoring systems. Practical examples, such as MAgeLan and ContTest have been given, in order to implement and to test the theoretical approaches.

Multi-agent monitoring systems bring quality and efficiency in almost every area of activity, including scientific research, education, healthcare or defense.

By implementing the solid theoretical concepts of artificial intelligence and reuniting them to the infinite resources of distributed systems and the Internet, we get the possibility to increase the quality of the results and reduce the duration of reaching the objectives.

Reengineering maintains the multi-agent systems and expands their utility in time and functionality. Consequently, software which proved to be precious in the activity continues to be efficient in service for a longer period of time.

Bibliography

[1] Atanasiu, A., 2007, *Formal languages and automata*, InfoData Publishing House, Cluj-Napoca, Romania

[2] Dzitac, I., Bărbat, B.E., Artificial Intelligence + Distributed Systems = Agents, *Int. J. of Computers, Communications & Control*, ISSN 1841-9836, E-ISSN 1841-9844, 4(1):17-26, 2009

[3] Pătruţ, B., Pandele, I., 2008, How to Compute the References Emergencies in a Hyper-encyclopedya, in "Recent Advances in Systems Engineering and Applied Mathematics. Selected papers from the WSEAS conferences in Istanbul, May 27-30, 2008", ISBN 978-960-6766-91-6, ISSN 1790-2769, Istanbul, Turkey, pp.72-75

[4] Pătruţ, B., Socaciu, T., 2008, *Constructing a Hyper-enciclopedya: How to Transfer the Emergencies Between the Nodes*, in Proceedings of the Fourth International Bulgarian-Greek Conference (Computer Science' 2008), 18-19 September 2008, Kavala, Greece, pp. 468-473

[5] Pătruţ, B., Vârlan S. E., Socaciu, T., 2008, *ContTest - a Multiagent System for Accounting Education*, in Proceedings of the Third International Multi-Conference on Computing in the Global Information Technology, ICCGI 2008, July 27 - August 1, 2008, Athens, Greece

[6] Tomozei, C., 2009, *Security Engineering and Reengineering on Windows 2008 Server Based Distributed Systems*, Journal of Information Technology & Communication Security, SECITC 2009, Bucharest, pp. 63-73

[7] Vlassis, N., 2007, *A Concise Introduction to Multiagent Systems and Distributed Artificial Intelligence. Synthesis Lectures on Artificial Intelligence and Machine Learning*, Department of Production Engineering and Management, Technical University of Crete, Morgan and Claypool Publishers

[8] Weiss, G. (ed.), 1999, *Multiagent Systems - A Modern Approach to Distributed Artificial Intelligence*, MIT Press, ISBN 978-0-262-23203-6

[9] FIPA Standards Specifications: http://www.fipa.org/specifications/index.html

Int. J. of Computers, Communications & Control, ISSN 1841-9836, E-ISSN 1841-9844
Vol. V (2010), No. 5, pp. 862-870

A Novel QoS Framework Based on Admission Control and Self-Adaptive Bandwidth Reconfiguration

A. Peculea, B. Iancu, V. Dadarlat, I. Ignat

Adrian Peculea, Bogdan Iancu, Vasile Dadarlat, Iosif Ignat
Technical University of Cluj-Napoca
Romania, 400020 Cluj-Napoca, 15 Constantin Daicoviciu
E-mail: {Adrian.Peculea,Bogdan.Iancu,Vasile.Dadarlat,Iosif.Ignat}@cs.utcluj.ro

Abstract: This paper proposes a novel end-to-end QoS framework, called Self-Adaptive bandwidth Reconfiguration QoS framework (SAR). SAR provides end-to-end QoS guarantees on a per-flow basis through admission control and end-to-end bandwidth reservation. In order to adapt to short and long time traffic load changing, SAR performs dynamic bandwidth reconfiguration. Due to a new organization of the network physical lines, SAR allows for a better utilization of the links' capacity and a smaller number of rejected flows, increasing the network's availability.
Keywords: end-to-end QoS, admission control, bandwidth reconfiguration.

1 Introduction

Computer networks transport simultaneously several flows, fact that makes necessary a multiplexing mechanism. Transport procedures affect the traffic flows, reason for which the traffic has to be characterized and quality of service (QoS) requirements need to be established. Traffic types and their QoS requirements impose the implementation of QoS methods and architectures. This paper presents the design and implementation of a new end-to-end QoS framework with self-adaptive bandwidth reconfiguration.

Integrated Services (IntServ) [1] provide end-to-end quality of service (QoS) guarantees for individual flows by maintaining the state and by reserving bandwidth for each flow at routers on the path between source and destination. The additional loading introduced by the per-flow bandwidth reservation processing and by the per-flow state maintaining at each router is significant and is increasing along with the network. For this reason, Integrated Services presents scalability problems.

Differentiated Services (DiffServ) [1] group the flows in traffic classes at the edge of the network. Interior routers forward each packet function of the per-hop behavior associated to the traffic class of the packet. Because of the flow aggregation and the lack of admission control, Differentiated Services do not provide end-to-end QoS guarantees to individual flows.

On-Demand QoS Path (ODP) [2] provides end-to-end QoS guarantees to individual flows introducing an additional load much lower than in the case of Integrated Services and maintaining a similar scalability to the one of the Differentiated Services. ODP exercises per-flow admission control and end-to-end bandwidth reservation at the edge of the network. Inside the network ODP differentiates the traffic classes as in the Differentiated Services. The main disadvantage of ODP is that the bandwidth adjustment is only inside the traffic class and does not allow for bandwidth redistribution between classes. The free bandwidth of the Provisioned Links that are not used or present a low utilization can not be made available for other Provisioned Links, the free bandwidth remaining unused. Another disadvantage of this framework is the fact that it does not include a module for determination of the bandwidth necessary for each input flow.

In order to eliminate the disadvantages above mentioned, we elaborated, implemented and proposed a framework for end-to-end quality of service guaranteeing through admission control

and self-adaptive bandwidth reconfiguration, which allows for bandwidth redistribution between classes. In this approach, the Physical Line is divided into two main sections, a part being the Guaranteed Link (GL) necessary for guaranteeing a minimum bandwidth (where is the case) for traffic classes (TCs), and a common part named Common Link (CL), which can be used by any TC. Having two separated sections, the framework guarantees a minimum bandwidth for any trunk and offers a common bandwidth which can be used by every trunk, irrespective to their TC. This allows for better bandwidth utilization and for the decrease of the rejected flows number. This paper is organized in the following manner. Section II presents related work, Section III describes the architecture and the functioning of the proposed framework, Section IV and Section V present the admission control method, and respectively the self adaptive reconfiguration technique of the proposed framework and, finally, Section VI presents the experimental results and the concluding remarks.

2 Related Work

Integrated Services (IntServ) framework uses Resource Reservation Protocol (RSVP) to reserve bandwidth for each flow at every router along the path of the flows. Using per-flow based hop-by-hop signaling, consisting of PATH and RESV messages, Integrated Services provides end-to-end guarantees. These guarantees come with the overhead of processing per-flow bandwidth reservation and maintaining per-flow state at each router along the flow's path. Because this overhead is significant and is increasing along with the network size, IntServ presents scalability problems.

Differentiated Services (DiffServ) framework classifies packets into traffic classes at the boundary of the network. During the classification process each packet is marked according to its traffic class. The routers inside the network recognize the traffic class of the packets and, using a scheduling mechanism, forward each packet function of the per-hop behavior associated to the traffic class of the packet. In the case of this framework, the service is provided on a per-class basis instead of a per-flow basis as in IntServ framework. This approach removes the overhead specific to IntServ framework reason for which DiffServ framework is much more scalable. However, DiffServ framework does not exercise admission control at the edge of the network, so the network can be overloaded, reason for which this framework does not provide end-to-end guarantees.

On-Demand QoS Path (ODP) provides end-to-end QoS guarantees to individual flows with less overhead than in the case of IntServ, maintaining a similar scalability to the one of the DiffServ. Two types or routers are defined in this framework: edge and core. ODP exercises per-flow admission control and end-to-end bandwidth reservation at the edge of the network. Inside the network ODP differentiates the traffic classes as in the DiffServ. ODP organizes link bandwidth hierarchically. Each physical link is statically divided into several Provisioned Links (PLs), each PL being dedicated to a traffic class. Each PL is divided into several trunks, each trunk being dedicated to an edge router. An edge router keeps track of available bandwidth of its trunks and performs admission control locally without hop-by-hop signaling through network. The main disadvantage of ODP is that the bandwidth adjustment is only inside the traffic class and does not allow for bandwidth redistribution between classes. The free bandwidth of the Provisioned Links that are not used or present a low utilization can not be made available for other Provisioned Links, the free bandwidth remaining unused. Another disadvantage of this framework is the fact that it does not include a module for determination of the necessary bandwidth for each input flow.

3 The Architecture of the Framework

The proposed framework serves the user networks and defines two types of routers, edge and core, and entities for common bandwidth control.

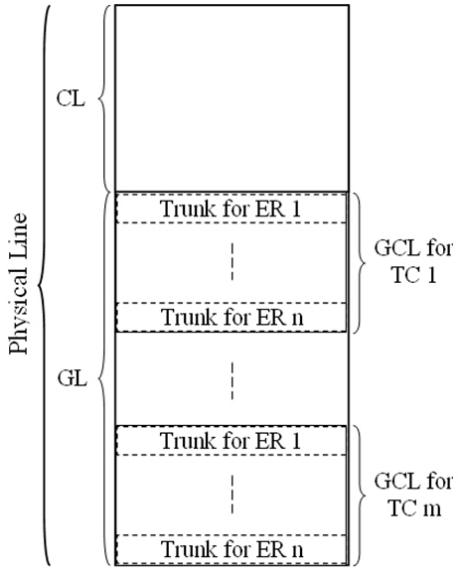


Figure 1: Bandwidth organization in the proposed framework

Edge routers (ERs), connected to served networks, determine the necessary bandwidth for each input flow, take admission or rejection decision for each input flow, dynamically reconfigure the bandwidth assigned to trunks, map flows to corresponding TCs and transmit the packets belonging to the admitted flows in the network. Core routers (CRs), connected to edge or core routers, recognize TCs and provide class based service differentiation. Entities for common bandwidth control monitor and update common bandwidths utilization and accept or reject the requests for additional bandwidth for trunks received from ERs.

The bandwidth is hierarchically organized. Each Physical Line is divided in two sections as it is presented in Figure 1. A first section guarantees the minimum bandwidth, which can be also 0, for each class and each trunk. The second section, CL, offers a common bandwidth which can be used by every trunk function of their bandwidth requirements, irrespective to their belonging TC or ER. So, trunks can acquire additional bandwidth without being conditioned by the available bandwidth of the belonging class. First section is statically divided in several Guaranteed Class Links (GCLs). Each GCL is reserved to a TC existing a one to one mapping between the TCs supported by the Physical Line and GCLs. Each GCL is divided in several trunks, each trunk being dedicated to an ER. A trunk belonging to a GCL supports the flows belonging to the TC that corresponds to the considered GCL, originating from the ER to which the trunk is dedicated, irrespective to their destination. An ER keeps track of available bandwidth of its assigned trunks and performs admission control locally, without hop-by-hop signaling through network. A Virtual IP Path (VIP) is a path from a source ER to a destination ER for a TC, being a concatenation of trunks belonging to the source ER over a source-destination path.

The bandwidth assigned to trunks has a minimum guaranteed value which can be also 0 and, by using CL, is dynamically adjusted function of the network traffic modifications.

Function of the entities for common bandwidth control there are three possible approaches: Central Control (CC), Router-Aided (RA) and Edge-to-Edge (EE).

The architecture of the framework is presented in Figure 2 and it is composed of two en-

tities: edge router and entity for common bandwidth control. The edge router determines the necessary bandwidth for each input flow, takes the admission or rejection decision for each input flow, reserves the necessary bandwidth for each admitted flow, dynamically reconfigures the bandwidth assigned to trunks and classifies the packets belonging to the admitted flows. The entity for common bandwidth control monitors and updates common bandwidths utilization and accepts or rejects the additional bandwidth requests for trunks, received from edge routers. The communication between the two entities is realized through a predefined message set.

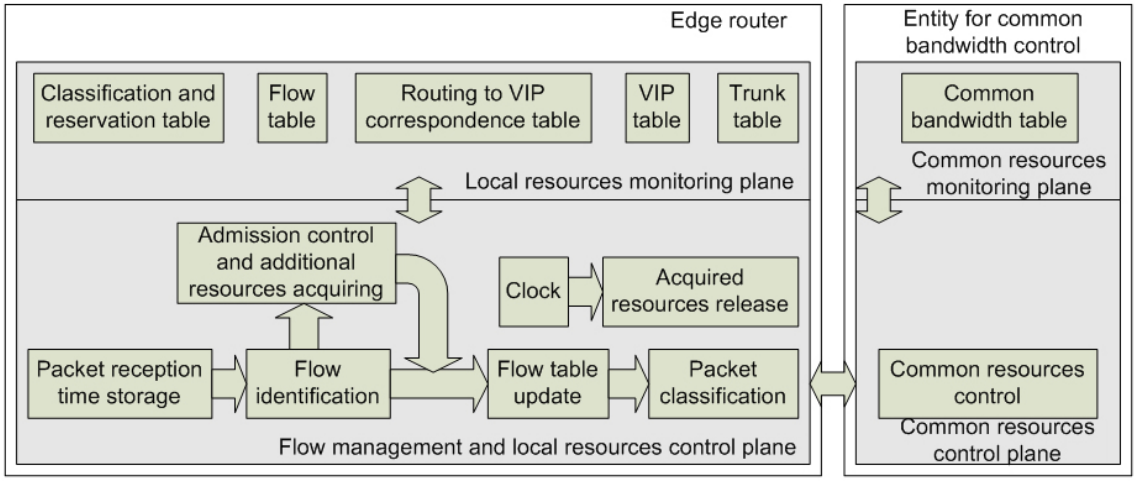


Figure 2: The architecture of the proposed framework

The edge router is composed of two planes: local resources monitoring plane and flow management and local resources control plane. The local resources monitoring plane is composed of the following tables: classification and reservation table which realizes a correspondence between flow types, the elements that identify them, corresponding traffic class, their necessary bandwidth and the maximum necessary bandwidth for any flow from the respective traffic class, flow table which stores the admitted flows and the time of the last packet from each flow, routing to VIP correspondence table which allows for VIPs determination, VIP table which stores the VIPs and trunk table which stores the reserved bandwidth, bandwidth being used and minimum reserved bandwidth for every trunk belonging to the ER. The flow management and local resources control plane takes the packets from the traffic policy module and delivers them to the routing process being composed of the following blocks: packet reception time storage which reads the receiving time of each packet, flow identification which determines and identifies the packets membership to admitted flows, flow table update which updates the reception time of the last packet from each flow from flow table, admission control and additional resources acquiring which admits the flows for which there are enough resources and rejects the flows when there is not enough bandwidth for them, acquires additional bandwidth for trunks, reserves the necessary bandwidth for the admitted flows and inserts the admitted flows in the flow table and packet classification which identifies the packets function of classification and reservation table criteria and marks them according to the identification criteria. The second task of this plane is to determine finished admitted flows and release the acquired resources used for these flows. The following blocks realize this task: clock generates the time period when acquired resources are released and acquired resources release which determines finished admitted flows and releases reserved acquired resources for these flows.

The entity for common bandwidth control is composed of two planes: common resources monitoring plane and common resources control plane. The common resources monitoring plane

contains common bandwidth table which stores the reservation and utilization for common bandwidths of CLs. The common resources control plane contains the common resources control block which updates the common bandwidth table and decides if additional bandwidth requests for trunks received from ERs can be accepted or not.

4 Admission Control

Admission control is performed at the arrival of the first packet from a new flow, by the source ER. Admission control and additional resources acquiring module stores the packet into the not admitted flows memory and determines if there are other packets belonging to this flow stored in the memory. If there are no more such packets, it determines from classification and reservation table the necessary bandwidth for the flow and TC, determines from routing to VIP correspondence table the flows corresponding VIP and extracts from VIP table the trunks that belong to the determined VIP. Then, for trunks which have enough available bandwidth, reserves the flows necessary bandwidth by updating the Bdw being used field. For a trunk, the condition to have enough available bandwidth is:

$$\mathrm{Reserved\_Bdw} \geq \mathrm{Bdw\_being\_used} + \mathrm{Necessary\_Bdw} \tag{4.1}$$

where Reserverd_Bdw and Bdw_being_used are the amounts of reserved and utilized bandwidth for the trunk and Necessary_Bdw is the flows necessary bandwidth.

The update of the Bdw_being_used field is done in the following manner:

$$\mathrm{Bdw\_being\_used} = \mathrm{Bdw\_being\_used} + \mathrm{Necessary\_Bdw} \tag{4.2}$$

If the VIP has enough bandwidth to support the input flow, the admission control accepts the flow. If there are trunks which do not have enough available bandwidth, admission control and additional resources acquiring module tries to increase the reserved bandwidth of those trunks sending in this sense a request to the entities for common bandwidth control. If the request is admitted, the reserved bandwidth of the trunks is increased by updating Reserved_Bdw field, so that these trunks too will have enough available bandwidth to support the input flow. For these trunks, admission control and additional resources acquiring module reserves the flows necessary bandwidth by updating the Bdw_being_used field. In this case too, the admission control accepts the flow. After a flow acceptance, the flow is inserted into the flow table and the packets belonging to this flow, stored into the not admitted flow memory, will be transmitted to the flow table update module for the rest of the processing and transmission. If the request is rejected, the flow is rejected, the reservations made on the trunks which had enough available bandwidth are canceled by updating the Bdw_being_used field and the packets belonging to the flow, stored into the not admitted flow memory, are discarded.

A flow is considered finished after an inactivity period that exceeds a predefined value. Each ER, using the acquired resources release module, periodically inspects its own flow table in order to identify the finished flows and, as a consequence of finished flow identification, releases the bandwidths correspondingly. If there are finished flows, the reserved bandwidth and TC for these flows are determined from the classification and reservation table and the flows are discarded from the flow table. Then, the acquired resources release module, determines from routing to VIP correspondence table the corresponding VIPs and extracts from the VIP table the trunks belonging to the determined VIPs. After this, releases the reserved bandwidth for the flows by updating the bandwidth being used field from the trunk table for each trunk belonging to the VIPs. The update of the Bdw_being_used field is done in the following manner:

$$\mathrm{Bdw\_being\_used} = \mathrm{Bdw\_being\_used} - \mathrm{Necessary\_Bdw} \tag{4.3}$$

Also, it extracts from the classification and reservation table the maximum amount of bandwidth for the corresponding TCs and verifies if the trunks utilization is under the predetermined lower threshold. For a trunk, the condition to have the utilization under a predetermined lower threshold is:

$$Reserved\_Bdw > Bdw\_being\_used + n * TC\_maximum\_necessary\_Bdw \quad (4.4)$$

where TC_maximum_necessary_Bdw is the maximum amount of bandwidth for the corresponding TC and n is a predefined parameter having a value larger than or equal to 1.

Also, it extracts from the trunk table the minimum reserved bandwidth for the trunks and verifies if the trunks have additional bandwidth acquired from CLs. For a trunk, the condition to have additional bandwidth acquired from CL is:

$$Reserved\_Bdw > Trunk\_minimum\_reserved\_Bdw \quad (4.5)$$

where Trunk_minimum_reserved_Bdw is the minimum reserved bandwidth for the trunk

If there are trunks whose bandwidth being used is under the predetermined lower threshold and the trunks have additional bandwidth acquired from CLs, the acquired resources release module, in the limit of the acquired bandwidth, computes de bandwidth that will be released from the reserved bandwidth of the trunks. Reduction of the reserved bandwidths is accompanied by appropriated resources release for the common bandwidths.

5 Self-Adaptive Bandwidth Reconfiguration

The proposed framework dynamically adjusts the bandwidth assigned to the trunks, in order to adapt to changes in network traffic. A source edge router has the option to request additional bandwidth for its trunks or it can release bandwidth not used by the trunks, depending on bandwidth usage of his trunks. Bandwidth adjustment is done using the CL's bandwidth. This adjustment allows all trunks, regardless of the class of traffic or the edge router where they belong, to share the bandwidth provided by LC. The trunk reconfiguration process of the proposed framework involves three main actions: (1) the control of the Common Bandwidth Table, (2) the release of bandwidth not used by the trunks, and (3) acquisition of additional bandwidth for trunks.

A Common Bandwidth Table stores the common bandwidth utilization of the network CLs. As shown in Figure 3, an entry in this table contains: CLs identifier, the reserved amount of shared bandwidth and the amount of shared bandwidth for the CL.

CL ID	*Reserved BdwC*	*BdwC being used*

Figure 3: Common Bandwidth Table

Depending on the share bandwidth entities, three approaches are being proposed: Central Control (CC), Router-Aided (RA) and Edge-to-Edge (EE). In the Central Control approach the Common Bandwidth Table is managed by a network management server (NMS) and the Common Bandwidth Table stores the bandwidth utilization of all CLs in the network. In the Router-Aided approach, each core router manages a Common Bandwidth Table, and each of these tables stores the bandwidth utilization of the LCs belonging to all physical links directly connected to that core router. In the Edge-to-Edge approach each edge router manages a Common Bandwidth Table, which will store the bandwidth utilization of all the CLs in the network.

1. If $\mathrm{Re}served\_Bdw > Bdw\_being\_used + n \times$ TC_maximum_necessary_Bdw and
   $\mathrm{Re}served\_Bdw >$ Trunk_minimum_reserved_Bdw
   1.1. $\mathrm{Re}leased\_Bdw = \mathrm{Re}served\_Bdw - Bdw\_being\_used - n \times$ TC_maximum_necessary_Bdw
   1.2. If Trunk_minimum_reserved_Bdww $> \mathrm{Re}served\_Bdw - \mathrm{Re}leased\_Bdw$
      1.2.1. $\mathrm{Re}leased\_Bdw = \mathrm{Re}served\_Bdw -$ Trunk_minimum_reserved_Bdw
   1.3. $\mathrm{Re}served\_Bdw = \mathrm{Re}served\_Bdw - \mathrm{Re}leased\_Bdw$

Figure 4: Reserved bandwidth update algorithm

Each edge router periodically examines its own Flow Table and determines which flows are finished. If there are any finished flows, the flow table and the trunk table will be updated. A next step for the edge router is to examine the trunk table and to obtain the bandwidth utilization of its own trunks. If the bandwidth utilization of any trunk is under a predetermined lower threshold and those trunks have additional bandwidth acquired from the Common Link, the source edge router computes the amount of bandwidth to be released from the reserved bandwidth, adjusts the released bandwidth of the trunks, in the limit of the additional acquired bandwidth, updates its own trunk table and sends a control message to the entities for common bandwidth control, in order to release the used shared bandwidth. Adjusting a trunk's bandwidth is done only in the limit of the additional acquired bandwidth. The algorithm that describes the reserved bandwidth update process for the trunks is presented in figure 4.

The trunk reconfiguration process is always initiated by a source edge router using a threshold and computed values driven mechanism.

6 Experimental Results and Conclusions

For the development and testing of the proposed QoS framework (SAR framework) and also for the developing of new ones, an experimental methodology was used, rather than simulation techniques, thus an integrated solution - a development tool, was created [3]. Also a benchmarking system for QoS parameters [4] was developed in order to allow the testing of the proposed SAR QoS framework. The benchmarking system generates traffic for the defined testbed and measures the following parameters: delay, IP delay variation (IPDV) or jitter and bandwidth, both on TCP and UDP. The benchmarking allows a user to define and store complex traffic patterns that can be recharged for making further measurements, to test various QoS techniques based on the same traffic characteristics.

For simulations purposes, the Self-Adaptive bandwidth Reconfiguration QoS framework (SAR) described in the previous section and the ODP framework were tested, in a comparative manner, using the development tool and the benchmarking system. The final testbed is a network of programmable routers, and consisted of three edge routers and three served networks. The tests were intended as performance comparison between ODP and SAR frameworks. Traffic classes and traffic patterns were defined similarly in both frameworks tested. Four classes of traffic were considered. Two test traffic patterns were defined. In the first traffic pattern considered flows are injected from classes 2 and 3 and in the case of the second traffic pattern flows belonging to class 2 are injected. For both traffic patterns a balanced distribution of traffic from and to the served networks is ensured.

After testing and analyzing the results (Figure 5) it was found that the number of flows

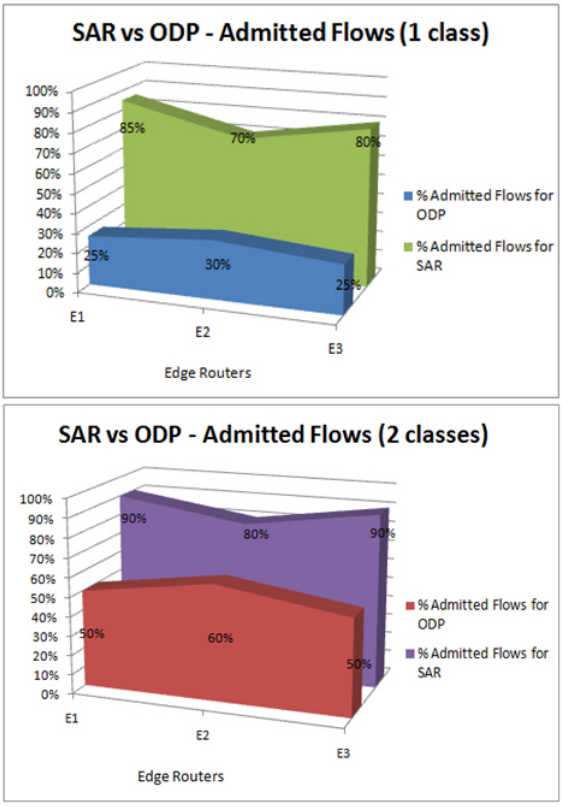


Figure 5: Test Results - Admitted Flows

admitted for SAR framework is higher than in the case of ODP framework, on both tested traffic patterns, which demonstrates a more efficient use of network resources. Also, the equal number of control messages transmitted by the two frameworks shows that SAR is a scalable framework. Finally, tests confirmed that admission control has eliminated network congestion.

This paper presents a new end-to-end QoS framework, called Self-Adaptive bandwidth Reconfiguration QoS framework (SAR). The proposed dynamic allocation method guaranties a minimum bandwidth available for each traffic class and trunk, and provides a common bandwidth section which can be used by every trunk, function of their bandwidth requirements, irrespective to their belonging TC or ER. Thus, trunks can acquire additional bandwidth without being conditioned by the available bandwidth of the belonging class. The new framework, SAR, uses the proposed bandwidth organization, allowing the increase of the traffic volume it handles, guaranteeing end-to-end quality of service through network resources monitoring, admission control and resource reservation for new flows. The end-to-end QoS framework with self-adaptive bandwidth reconfiguration overcomes the disadvantages of ODP by providing minimum service guaranties and bandwidth redistribution between classes.

Acknowledgments

This work was supported by the PNII-IDEI 328/2007 QAF - Quality of Service Aware Frameworks for Networks and Middleware research project within the framework - National Research, Development and Innovation Programme initiated by The National University Research Council - Romania (CNCSIS - UEFISCSU).

Bibliography

[1] Z. Wang, *Internet QoS: Architectures and Mechanisms for Quality of Service*, Morgan Kaufmann, San Francisco, 2001.

[2] M. Yang, Y. Huang, J. Kim, M. Lee, T. Suda, M. Daisuke, An End-to-End QoS Framework With On-Demand Bandwidth Reconfiguration, *Twenty-third Annual Joint Conference of the IEEE Computer and Communications Societies (INFOCOM)*, Hong-Kong, Vol. 3, pp. 2072 - 2083, 2004.

[3] A. Peculea, V. Dadarlat, I. Ignat, B.Iancu, L. Cobarzan, On Developing a Qos Framework With Self-Adaptive Bandwidth Reconfiguration, *Pollack Periodica An International Journal for Engineering and Information Sciences*, Vol.4, No.1, pp. 121-129, 2009.

[4] A. Peculea, B. Iancu, V. Dadarlat, I. Ignat, E. Cebuc, Z. Baruch, Benchmarking System for QoS Parameters, *Proceedings of the IEEE 3rd International Conference on Intelligent Computer Communication and Processing 2007 (ICCP 2007)*, Cluj-Napoca, Romania, p.255-p.258, 2007.

Int. J. of Computers, Communications & Control, ISSN 1841-9836, E-ISSN 1841-9844
Vol. V (2010), No. 5, pp. 871-883

Natural Language based On-demand Service Composition

F.-C. Pop, M. Cremene, J.-Y. Tigli, S. Lavirotte, M. Riveill, M. Vaida

Florin-Claudiu Pop, Marcel Cremene, Mircea Vaida
Technical University of Cluj-Napoca,
Faculty of Electronics, Telecommunications and IT,
Cluj-Napoca, Romania
E-mail: {florin.pop, marcel.cremene, mircea.vaida}@com.utcluj.ro

Michel Riveill, Jean-Yves Tigli, Stéphane Lavirotte
Université de Nice - Sophia Antipolis,
Sophia Antipolis CEDEX, France
E-mail:{riveill, tigli, lavirott}@unice.fr

Abstract: The widespread of Web services in the ubiquitous computing era and the impossibility to predict a priori all possible user needs generates the necessity for on-demand service composition. Natural language is one of the the easiest ways for a user to express what he expects regarding a service. Two main problems need to be solved in order to create a composite service to satisfy the user: a)retrieval of relevant services and b) orchestration/composition of the selected services in order to fulfill the user request. We solve the first problem by using semantic concepts associated with the services and we define a conceptual distance to measure the similarity between the user request and a service configuration. Retrieved services are composed, based on aspect oriented templates called Aspects of Assembly. We have tested our application in an environment for pervasive computing called Ubiquarium, where our system composes a service according to the user request described by a sentence. The implementation is based on the WComp middleware that enables us to use regular Web services but also Web services for devices.
Keywords: Natural Language, Service Composition, On-demand, Middleware, Templates

1 Introduction

Background. Since Web 2.0 marked it's appearance as a concept in the fall of 2004 and introduced the principle of the Internet as a platform [19], the complexity and diversity of this platform grew together with the more enhanced features it was providing to its users. Given the fact that the software in the Internet era is delivered as a service, not as a product and there is no release cycle for the services, it is the user who's in charge of finding a service and using it.

In a near future services will be more diverse and widespread as computers will become ubiquitous. In the same way the information across the Web is structured, classified and then presented to a user that sends a natural language request to a search engine, so a collection of applications should be assembled, classified and deployed using the services that are found within a given context based on a similar unrestricted language request coming from the user.

The problem. In a context where various services with different functionality are available, it's possible to compose new services on-demand, based on a user request, only when the right selection of components is used.

On-demand service composition involves two operations: *service retrieval* and *service orchestration. Service retrieval* refers to identifying those specific services that are addressed by the user

request or the closest functional match to the request. The transition from a natural language request to a list of services is a challenge that is even more difficult when no restrictions are added to control the request. *Service orchestration* or service assembly is the process of linking the retrieved services in a functional flow so that the user demand is fulfilled. Both mentioned problems make the object of this paper, but while service retrieval was the field of our research, for service orchestration we used an existing approach.

Scenario. We consider the following scenario to illustrate the purpose of dynamic service composition based on natural language requests. A handicapped person lives inside an intelligent house, surrounded by intelligent devices, sensors and actuators. Each device has its own inputs and outputs and is able to process different types of data, leading to a large number of possible functional combinations of those devices. The person in the intelligent house wants to use those devices by combining their functionality (e.g. link the output from a sensor to the input of an actuator), but his disability prevents him to physically interact with the devices or he simply lacks any technical knowledge. Therefore, he expresses his need using the natural language (either written or spoken): *"I want to use my remote control on the wheel chair to turn off the light, change the channel on TV and play some music on the media center"*. Each device the user addresses through his request (remote control, light, TV, media center) provides a different service with specific actions that can be used in various configurations. Finding a way to sort these configurations by relevance to the user's request is a key requirement for the imagined scenario. Also, when one of the devices the user wants to use is not present in the intelligent house or it was replaced with an updated version, the system should adapt and assemble a service that is the closest match to the user's need.

Approach. Dynamic service composition solves the problem of adaptation to different contexts and user preferences. Also, by composing services on demand, the learning curve required for the user to work with new configurations is reduced as the user *"gets what he wants"* from the application. Existing systems for dynamic service composition based on natural language requests either provide a restricted natural language interface or don't offer support for adaptation to structural and behavioral changes of the service configuration.

We start with an initial set of services that are discoverable across a network. The user requests a completely new service using an unrestricted natural language sentence. In order to find specific devices to satisfy the request, we use semantic concepts and define a conceptual distance between the request and a service configuration. Concepts are leafs on a lexical tree that is generated by deriving and generalizing a notion. Once the services that match the request are identified, some aspect oriented advices are used to connect the services so that when a service disappears from the context or a new service is made available, the service configuration adapts.

Outline. This paper is organized as follows: the next section examines the problems a dynamic service composition system should solve in order to be usable in the modern context of Web services. Section 3 describes the solution we propose including the principles that lead to this solution, while section 4 focuses on the design and implementation patterns we used, along with the test results. Section 5 examines some of the existing dynamic service composition approaches. The paper ends with conclusions and further research.

2 An overview of the problem

On-demand dynamic service composition based on natural language requests raises some challenges that need to be studied before a solution is to be proposed.

Service retrieval. The first problem we approach in this paper is finding and retrieving a particular service. The large variety of Web services useable for composition needs to be classified

in a way that would make it machine meaningful and semantic-rich for the search to provide the best results.

One Web extension, called Semantic Web [3] is focused on enabling better Web service interoperation. The Semantic Web's purpose is to bring structure to the meaningful content of Web pages, creating an environment where software agents roaming from page to page can readily carry out sophisticated tasks for users. One dialect of the DAML [14] family of Semantic Web markup languages was proposed in [17] for the markup of Web services. This so-called semantic markup of Web services creates a distributed knowledge base (KB) that provides a means for agents to populate their local KBs so that they can reason about Web services to perform automatic Web service discovery, execution and composition and interoperation. But the semantic mark-up uses a narrow, predefined vocabulary as identified in [16], which makes possible only the retrieval of those Web services for which the vocabulary is known. Queries or requests from Web services or user requests, using another vocabulary than the predetermined vocabulary are not suitable to find or retrieve such a Web service. Therefore, a narrow vocabulary for the semantic mark-up of Web services is not appropriate to be used in combination with natural language requests.

Service composition. The second problem that needs to be solved is the actual composition of the retrieved services. There are 3 types of systems for dynamic service composition according to [12].

Template-based systems [18,21] are using a service template to compose an application. They can handle complex interactions between components and allow some level of flexibility by choosing different sets of components. The drawback of this approach is that they cannot compose applications for which templates are not available.

Interface-based systems [7] allow the user to submit a set of inputs and outputs for the application he is requesting. These systems have a higher adaptability than the template-based systems, but certain applications cannot be represented as a set of inputs and outputs (e.g. an email sending service does not output any data).

Logic-based systems [20,23] extend the interface-based approach by adding extra information into interface information using first order logic or linear logic. A user requests an application by submitting a first order formula representing the logic that must be satisfied by the application. They are more adaptable than the template-based systems since they donÕt require service templates and offer support for more varieties of services than the interface-based systems. Their main disadvantage is given by the fact that they are not extensible and are not suitable for a distributed environment.

Adaptation. The last, but the most important problem that needs to be solved by a dynamic service composition system is adaptation. We distinguish two types of changes that require adaptation [2]: structure changes and behavior changes. Structural adaptation consists in modifying the retrieved services while preserving the global behavior of the application. The behavior describes the sequence of operations to be executed to fulfill the user request. Structural changes are triggered when a retrieved services disappears from the context or is replaced by another (for example the analog TV is replaced by a digital TV). Behavior changes are related to the user who may decide that the current service does no longer satisfy his need.

3 Proposed solution

Our dynamic service composition system was inspired out of the user's need to interact with the intelligent devices that surround him. This user-machine interaction should be as natural as human interaction, through unrestricted natural language. Intelligent devices are entities that provide services to the user and have networking capabilities. Roughly, there are two types of

intelligent devices: *basic devices* that are simple service producers (e.g. a light that offers the illuminating service, a TV that offers the tuning service that allows changing the TV channels) and *controller devices* that can consume services by other devices, acting as interfaces for a composite service (e.g. a mobile phone, an ultra-mobile PC, a netbook). Hybrid devices that implement both previously mentioned functionalities can also be imagined.

The intelligent devices are connected to form an Ad-Hoc network inside the intelligent house, which they use to exchange information. This means that the devices can appear and disappear from the network structure on-the-fly: a device can auto-configure when it joins the network and then leave the network without notice. Network and device management is a task for the middleware that runs the intelligent house.

3.1 Semantic descriptions for Web services

WSDL [10] is intended for the functional description of Web services and the Semantic Web mark-up is limited to a narrow vocabulary, which is not suited for natural language requests. A lexical tree [16] would add too much semantic information to a service and would not be suited for embedded devices.

To overcome these limitations, we propose the use of general notions, called *concepts*, to describe the utility of a service. A service is not entirely identified by a single concept, but by an infinite number of concepts that are determined through the generalization of a notion. This notion will serve as a semantic description for an intelligent device that offers a service. We use the *television* notion to describe a TV, for example. Through generalization we find that both the *television* and *electronic equipment* concepts refer to the same device. To increase the precision, a lexical analysis is also conducted for the service description and the user request by the composition system. This way, the service description suffers little or no modifications due to the extra semantic information.

3.2 Linguistic processing

We consider the scenario where the user interacts with appliances and he expresses his need through a sentence: *"I want to use my phone to turn off the light, turn on the TV and play some music on HiFi"*. In order to retrieve the services required to satisfy the user need, the request goes through a linguistic processing module, responsible for:

- Text segmentation required to separate the words in the phrase (e.g. *switch off the light* is transformed into *switch*, *off*, *the*, *light*);
- Removing stop words that are considered to be irrelevant (e.g. *the*, *to*, *and*);
- Stemming (e.g. *lights* is transformed into *light*, *using* becomes *use*);
- Spell-checking to correct the misspelled words and the words "damaged" during stemming.

The output text segments for the user request in the considered scenario are: *want*, *use*, *phone*, *switch*, *light*, *turn*, *tv*, *play*, *music*, *hifi*.

3.3 The graph of concepts

The *text segments* together with the *service descriptions* are nodes in a graph, called the *graph of concepts*. The arcs in this graph connect each text segment to each service description. The weight of each arc represents the *conceptual distance* between the text segment and the

service description. We introduced the conceptual distance to measure the relationship between two notions.

Measures of semantic similarity or relatedness are found in linguistic processing literature. According to the study published in [6], the Jiang and Conrath's measure proved to be best for practical usability. All compared distances are based on the information content of the lexical terms (the probability of encountering an instance). While this information may be valuable for other applications, it is of less importance for service composition and it adds to the complexity of the implementation, therefore making it slower. Our approach, the conceptual distance, is faster and less complex than Jiang and Conrath's measure, while the last is more accurate.

The conceptual distance is a numerical evaluation of how accurate two notions refer to the same concept. For example the words *phone* and *telephone* describe the same concept - a communication device, therefore the conceptual distance is null. On the other hand, the words *phone* and *electronic equipment* can describe the same concept - a communication device, but one of them is more general, therefore it can address more concepts, which leads to a non-null distance between these words.

Figure 1 shows an example of a graph of concepts where the text segments are *tv*, *light*, *hifi*, *phone* and the service semantic descriptions are *Television*, *Light*, *DVD*, *HiFi*, *Mobile Phone* and *PDA*. Figure 1.A. represents all the distances and figure 1.B. represent only the arcs with minimum conceptual distance.

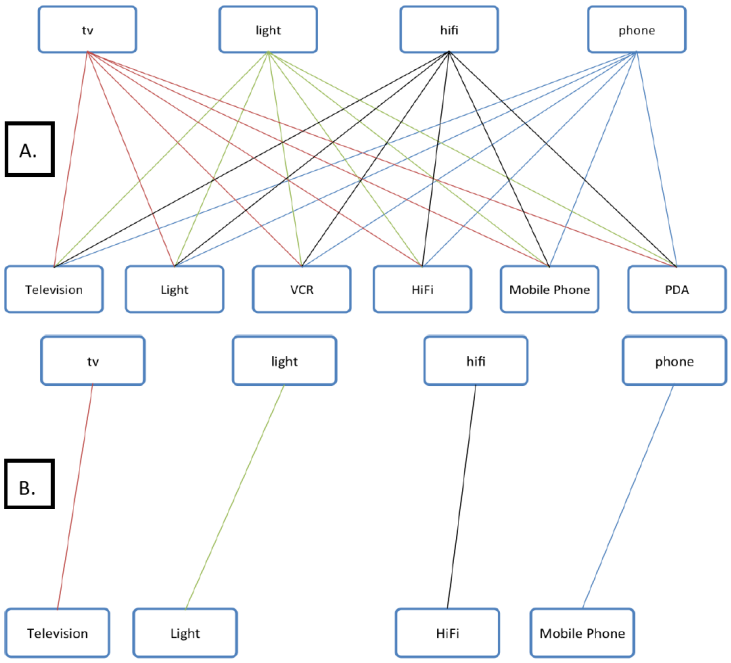


Figure 1: The graph of concepts

3.4 Knowledge structure

In order to evaluate the conceptual distance we need to find a way to classify the lexical basis of the English language. We used for this purpose a specialized dictionary called WordNet [11]. WordNet groups nouns, verbs, adjectives and adverbs in sets of synonyms, called synsets. Each synset describes a different concept. Different senses of a word are in different synsets. Most synsets are connected to other synsets via a number of semantic relations. For example, the semantic relations for nouns include:

- hypernyms: Y is a hypernym of X if every X is a (kind of) Y (mobile phone is a hypernym of phone);
- hyponyms: Y is a hyponym of X if every Y is a (kind of) X (phone is a hyponym of mobile phone);
- coordinate terms: Y is a coordinate term of X if X and Y share a hypernym (mobile phone is a coordinate term of cellular phone, and cellular phone is a coordinate term of mobile phone);
- holonym: Y is a holonym of X if X is a part of Y (mobile phone is a holonym of transmitter);
- meronym: Y is a meronym of X if Y is a part of X (transmitter is a meronym of mobile phone).

While semantic relations apply to all members of a synset because they share the same meaning, words can also be connected to other words through lexical relations, including antonyms and derivationally related, as well. Both nouns and verbs are organized into hierarchies, defined by hypernym or *IS A* relationships.

For example, the hierarchy for mobile phone is:

- cellular telephone, cellular phone, cellphone, cell, mobile phone
- radiotelephone, radiophone, wireless telephone
- telephone, phone, telephone set
- electronic equipment
- equipment

The words at the same level in hierarchy are synonyms of each other.

3.5 The concept hierarchy

The algorithm that evaluates the conceptual distance uses the WordNet lexicon to create concept hierarchies. A concept hierarchy is generated in 4 steps:

1. Find the synset that contains the concept for which the hierarchy is generated. Each word in the synset becomes a root for a tree in the concept hierarchy.
2. For each tree root, find the synsets that are in a relationship with the rootŐs synset. Each word in the related synset becomes a leaf for the tree, on the next level in hierarchy, branching from the root.
3. For each word on the current level in hierarchy, find the synsets related to the wordŐs synset and add the words in the found synsets as leafs for the tree on the next level.
4. Repeat Step 3 until the hierarchy is big enough so that the degree of generalization for the notion for which the hierarchy is built, corresponds to an accepted accuracy that produces best results. The bigger the hierarchy the longer it takes to generate it, but the smaller the hierarchy the more confusion can occur among concepts.

The hierarchy for the notion mobile phone is shown in Figure 2. The roots of each the tree are part of the same synset and each level in a tree represents the words from the synsets that are related to the word they are branching from.

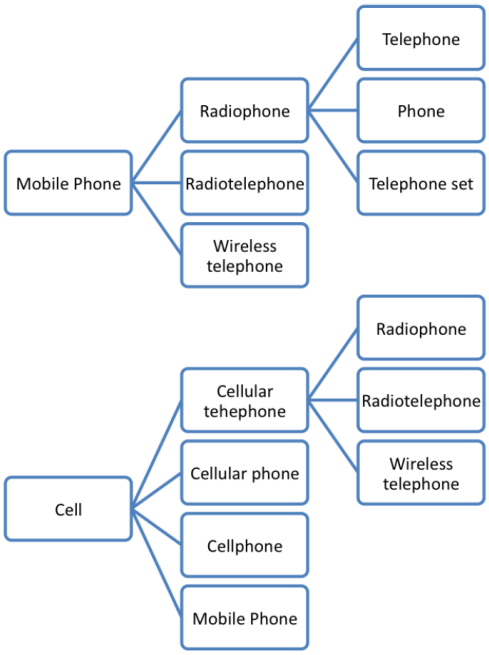


Figure 2: The concept hierarchy for the notion *mobile phone*

3.6 The conceptual distance

In order to evaluate the conceptual distance for two notions, a concept hierarchy is built for each notion. Then, the conceptual distance is calculated as follows:

- The minimum difference of levels between the common node of the 2 hierarchies and the node that represents the notion on which the hierarchy is built for, if such a common node exists.
- The maximum number of levels a hierarchy can have if thereÕs no common notion among the two.

Examples:

- D (Mobile Phone, Cell) = 0
- D (Radiotelephone, Radiophone) = 0
- D (Mobile Phone, Radiophone) = 1
- D (Mobile Phone, Telephone) = 2

3.7 Service retrieval

Finding and retrieving the closest services to the user request resumes to identifying the couples (*text segment*, *service description*) in the graph of concepts that have a minimum sum of conceptual distances. This makes it easier to take advantage of a service that only partially matches a request.

In order to find mentioned couples we need to apply tow transformations to the graph of concepts:

- Finding the sub-graph that has the minimum distance path that includes all the nodes. After this transformation each service description will be connected to 2 text segments.
- For each triplet (*service description*, *text segment 1*, *text segment 2*) remove the arc that has the maximum weight.

In order to find the minimum weight sub-graph, we use the Kruskal [15] algorithm that calculates the minimum spanning tree (MST). The nodes that contain the service descriptions resulted after the two transformations are applied, represent the services that are used to generate the composite service requested by the user. The transformed sub-graph for the graph in Figure 1 contains the nodes that are connected with the thick continuous line.

3.8 Service composition

We used a *template-based* service composition system because of its capability to handle complex interactions between components and the flexibility of choosing different sets of components. The system we used, called Aspects of Assembly (AoA) [9] is part of the WComp [8] middleware for ubiquitous computing and besides the benefits that derive from the fact that is template-based, also offers support for auto-adaptation.

These templates can be automatically selected either by the service composition system when satisfying a user request or triggered by context changes in a self-adaptive process and composed by a weaver with logical merging of high-level speciҶcations. The result of the weaver is projected in terms of pure elementary modiҶcations (PEMs) Ð add, remove components, link, unlink ports. The AoA architecture consists of an extended model of Aspect Oriented Programming (AOP) for adaptation advices and of a weaving process with logical merging.

An AoA template is structured as an aspect with a list of components involved in composition (called pointcut) and adaptation advice (a description of the architectural reconfigurations), which is specified using a domain specific language (DSL). We will examine some AoA templates and the composition process in detail in the next section.

4 Implementation and results

We used the WComp [8] platform for ubiquitous computing as the middleware of our intelligent house. WComp uses the UPnP protocol to achieve device interconnectivity and interoperability. Each UPnP device has a software proxy that acts like a software component. Using this proxy, we can treat Web services for devices similar to the UI components of a GUI designer. We added some meta-data to the UPnP service description for each device to serve as semantic description.

Using WComp, we have simulated the following devices/services: TV set, described by the *television* notion; DVD recorder, described by the *DVD* notion; Mobile phone, described by the *mobile phone* notion; PDA, described by the *PDA* notion; HiFi, described by the *HiFi* notion; Lighting system, described by the *light* notion.

The interactions between these components were specified using AoA templates. Following, is an example of such a template that is used to connect the mobile phone to the TV:

```
Pointcut
   inputDev:=/mobilePhone.*/
```

```
   outputDev:=/television.*/
Advice KeyToChannel(inputDev, outputDev):
      input.^key_Pressed ->
            ( output.set_Channel ; CALL )
```

The first 3 lines describe (defining the pointcut as in aspect programming), using filters in the AWK language, the components involved in the interaction: a mobile phone (*inputDev*) and a television (*outoutDev*). The filters of type */instanceName.*/* will find components that have their name prefixed by *instanceName*. Line 4 declares a composition schema that uses the previously described components. Lines 4-5 specify the composition mechanics: call the *tv.set_Channel* method when the *mobile phone* fires the event *key_Pressed*.

The service composition system implements a UPnP device that offers the service of designing composite services for the user in order to fit seamlessly with the WComp middleware. Requests from the user are captured by a WComp assembly of components, and then sent using the UPnP protocol to the service designer along with a description of the context where the devices are located. The service designer queries the devices for service descriptions (semantic meta-data), and then finds only those services that are relevant to the user request. Instances of the devices that provide the named services are interconnected based on the rules described in the AoA templates.

Scenario 1. "I want to use my phone to turn off the light, turn on the TV and play some music on HiFi". This phrase contains many irrelevant words to the service composition system, but the relevant words are identical (except TV) to the service semantic descriptions. Irrelevant words have an effect of increasing the time required to process the graph of concepts. All the relevant services are identified and then composed.

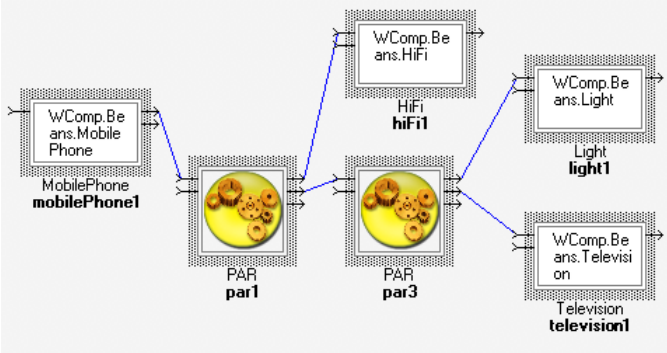


Figure 3: The dynamically composed service for Scenario 1

Scenario 2. "Use PDA for broadcasting". This user request is challenging for any composition system because it doesn't address the TV directly, but through the abstract concept of *broadcasting*. Due to the use of the specialized dictionary, the TV is found and then connected to the PDA.

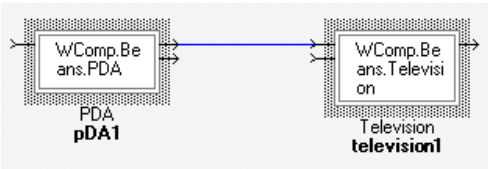


Figure 4: The dynamically composed service for Scenario 2

The WordNet (free download) dictionary is installed on a local machine thus the word search is fast. The main complexity of the algorithm is given by the conceptual distance computation. The time necessary in order to compute the distance between two concepts (similarity), based on the English dictionary, was about 6 ms (on a Dell Latitude 830 laptop, CPU Dual Core 2.2GHz, 2G RAM) with the fastest algorithm that we have found (a WordNet Similarity implementation, the java package edu.sussex.nlp.jws.JiangAndConrath). Thus, if we have M user concepts and N service concepts, we need a time equal to $\frac{M*N}{2} * 6$ ms. for instance, we can create the concept graph for 10 user concepts and 64 services in about 2 seconds. The AoA application is very fast also (less than 1 sec. for up to 350 components).

In this section we have shown that we are able to create, in practice, new services, on-demand, using real devices, by applying the patterns available for the selected set of services. The scenarios discussed above were tested in a dedicated environment for ambient computing, called *Ubiquarium* [24]. Real devices were used (a part of the services are real devices and another part are virtual ones).

5 Related work

Composing Web services on the basis of natural language requests. The solution described in [4] and [5] assumes that the user requests are expressed with a controlled subset (a narrow vocabulary) of natural language. The sentence that represents the userÕs request is transformed into a flow model using templates (e.g. *if ... then ... else*, *when ... do*). Verbs are used to identify the action and its parameters. Each available service is paired with a well-defined set of keywords. OWL-S annotations are used to provide operation semantics and an ontological classification of Web services. The operations act as nodes of a direct acyclic graph and the relations among their IOPEs (Inputs, Outputs, Preconditions and Effects) establish arcs. The graph is translated into an executable service at invocation time.

The example the authors use to sustain the proposed solution uses the phrase *"If there is any cinema showing "Big Fish" in Turin then send a SMS to Dario containing "Lets go to the movies tonight!""* An *if ... then* template is used to identify the flow model. In the next step, called *context focus*, the service types are identified: *cinema*, *SMS*. The verb (*send*) triggers the parameter retrieval stage where IOTypes recognizers based on format (Date, Time, Telephone NumbersÉ) and values (City namesÉ) are used to extract the actionÕs parameters.

This solution establishes a synergy between the semantic service descriptions and the Natural Language interpretation of user requests, through a common ontology and a consistent lexical vocabulary. Therefore it canÕt be used in active environments where new components that act as black-boxes appear and disappear from the context dynamically. Also, the use of a controlled subset of natural language makes it non intuitive for the user as he is restricted to the use of templates when expressing a request.

Semantics-based dynamic service composition. Papers [12] and [13] propose the *CoSMoS* model and the *SegSeC* platform for dynamic service composition. Their idea is to transform the semantics of the user request into a semantic graph. Nodes in the semantic graph represent operations, inputs, outputs and properties of a component, as well as their data types and concepts. Arcs (labeled links) represent the relationships among the nodes. Concepts, entities representing abstract ideas actions are used to annotate the semantics of the operations, inputs, outputs and properties of components. The user request is parsed and the components addressed by the user form a workflow.

The example in [13] uses the phrase *"Print directions from home to restaurant"*. The semantic graph contains the predicate (*print*), the target of the action (*direction*) and the parameters (*home*, *restaurant*). The workflow, containing the retrieved components, is executed as soon as

it satisfies the user request. This analysis takes place in a step called semantic matching and consists in a test that verifies that all the links that appear in the user request also appear in the graph that models the workflow.

The authors of [13] admit that their solution is not suited for environments where a large number of components are deployed. The platform lacks the feature of providing a solution in the case where the workflow doesnÕt satisfy the userÕs request. If the generated workflow doesnÕt match exactly the user request, then the dynamic service composition fails. Also, the ability of the implementation to discover certain components is to be questioned because itÕs limited to work with a narrow set of keywords and it lacks a vocabulary.

Web service with associated lexical tree. The invention claimed by Alcatel [16] relates to a method to mark-up a Web service in order to allow finding and retrieving said service via a natural language request. A lexical tree, built by deriving the service description, finding synonyms and related forms of the derived keywords, is associated to each service. Finding a service based on the user request resumes to comparing the natural language query to the lexical tree of each Web service. This method of retrieving a Web service proves to be the most appropriate when dealing with natural language requests. The invention however doesnÕt exploit the full potential of this finding, as it lacks service composition.

6 Conclusion

This paper proposes a new method for assembling services on demand, starting from the user request expressed in natural language. We use a semantic analysis of the user request, in order to identify the services described by concepts that are related to concepts from the user request. Retrieved services are then composed, based on composition patterns, called AoA (Aspects of Assembly). The uses of patterns, which assure that the new service is always valid, compensate the ambiguity of the natural language.

Another important advantage of the AoA patterns, comparing to other existent pattern-based approaches, is the fact these patterns may be superposed by composition. Thus, a large number of combinations are possible using a given set of patterns. Additionally, for a given set of services, the AoA mechanism applies only the patterns that lead to valid services.

The solution was implemented on the WComp platform and tested in a dedicated ambient computing environment, called *Ubiquarium*, using real and virtual intelligent devices/services.

The service composition is user driven, by natural language (voice) and allows the user to get the service on-demand. From this point of view, our solution is less restrictive than the other solutions described in the state of the art section.

An important advantage of our solution is the reuse of WordNet free dictionary, which is acting like ontology. Due to this, we can relax very much the limitations for the natural language, imposed by solutions where an ontology (usually restricted) must be created by the developer. Otherwise, the creation of a rich ontology is a very costly task and our solution succeeded to avoid it by reusing WordNet. This choice has another important advantage: it solves the problem of dealing with different ontologies and does not need to impose a common ontology (the only requirement is to use English).

A particular aspect of our proposal is the mixed approach: semantic and pattern-based. This approach combines the advantages of the both: thanks to composition patterns, it allows us to build complex composite services, which are always valid and functional. With other approaches (interface, logic, semantic based), that are not using patterns/templates, it is very difficult to create complex architectures that are valid and work correctly.

As future work, we intend to extend our solution for dynamic service adaptation (at runtime) and this should be feasible because WComp was designed for dynamic service reconfiguration in

pervasive environments.

Acknowledgments

This work was supported by the EcoNet project code 18826YM and the romanian national project PNCDI II code 1062, and 1083 financed by UEFISCSU.

Thanks to other members of the Rainbow team for fruitful discussions and feedback: Vincent HOURDIN, Daniel CHEUNG-FOO-WO, Eric CALLEGARI.

Bibliography

[1] A. V. Aho, B.W. Kernighan, P. J. Weinberger, *The AWK Programming Language*, Addison-Wesley, 1988.

[2] Anastasopoulos M.; Klus H.; Koch J.; Niebuhr D.; Werkman E., *DoAmI - A Middleware Platform facilitating (Re-)conŢguration in Ubiquitous Systems*, In System Support for Ubiquitous Computing Workshop. At the 8th Annual Conference on Ubiquitous Computing (Ubicomp 2006), Sep 2006.

[3] Berners-Lee, T.; Hendler, J.; Lassila, O., *The Semantic Web*, Scientific American Magazine, May 17 2001.

[4] Bosca, A.; Ferrato, A.; Corno, F.; Congiu, I.; Valetto, G., *Composing Web services on the basis of natural language requests*, IEEE International Conference on Web services (ICWS'05), pp. 817-818, 2005.

[5] Bosca, A.; Corno, F.; Valetto, G.; Maglione, R., *On-the-fly Construction of Web services Compositions from Natural Language Requests*, JOURNAL OF SOFTWARE (JSW), ISSN : 1796-217X, Vol. 1 Issue 1, pag 53-63, July 2006.

[6] E. Budanitsky, G. Hirst, *Semantic distance in wordnet: An experimental, application-oriented evaluation of five measures*, 2001.

[7] Chandrasekaran S.; Madden S.; Ionescu M., *Ninja Paths: An Architecture for Composing Services Over Wide Area Networks*, CS262 class project writeup, UC Berkeley, 2000.

[8] Cheung-Foo-Wo, D.; Tigli, J.-Y.; Lavirotte, S.; Riveill, M., *Wcomp: a multi-design approach for prototyping applications using heterogeneous resources*, In 17th IEEE Intern. Workshop on Rapid Syst. Prototyping, pag 119Ð125, Creta, 2006.

[9] Cheung-Foo-Wo, D.; Tigli, J.-Y.; Lavirotte, S.; Riveill, M., *Self-adaptation of event- driven component-oriented Middleware using Aspects of Assembly*, In 5th International Workshop on Midd leware for Pervasive and Ad-Hoc Computing (MPAC), California, USA, Nov 2007.

[10] Christensen, E.; Curbera, F.; Meredith, G.; Weerawarana, S. *Web services Description Language (WSDL) 1.1*, Website, 2001
`http://www.w3.org/TR/wsdl`

[11] Cognitive Science Laboratory, Princeton University, *WordNet Ð a lexical database for the English language*, Website, 2006
`http://wordnet.princeton.edu/`

[12] Fujii, K.; Suda, T., *Component Service Model with Semantics (CoSMoS): A New Component Model for Dynamic Service Composition*, SAINT-W '04: Proceedings of the 2004 Symposium on Applications and the Internet-Workshops (SAINT 2004 Workshops). Washington, DC, USA: IEEE Computer Society, 2004.

[13] Fujii, K.; Suda, T., *Semantics-based dynamic service composition*, IEEE Journal on Selected Areas in Communications, Vol 23(12), pag 2361- 2372, Dec 2005.

[14] Hendler, J.; McGuinness, D., *The DARPA Agent Markup Language*, IEEE Intelligent Systems, vol. 15, no. 6, Nov./Dec. 2000, pp. 72Ð73.

[15] Kruskal, J. B., *On the shortest spanning subtree of a graph and the traveling salesman problem*, Proc. Amer. Math. Soc., Vol 7, 1956.

[16] Larvet, P., *Web service with associated lexical tree*, European Patent, EP1835417.

[17] McIlraith, S. A.; Cao Son, T.; Zeng H., *Semantic Web services*, IEEE Intelligent Systems, vol. 16, no. 2, Mar./Apr. 2001, pp. 46-53.

[18] Molina A. J.; Koo H.-M.; Ko I.-Y., *A Template-Based Mechanism for Dynamic Service Composition Based on Context Prediction in Ubicomp Applications*, In Proceedings of the International Workshop on Intelligent Web Based Tools (IWBT'2007), 2007.

[19] O'Reilly, T. *What Is Web 2.0*, Website, 2005
`http://www.oreillynet.com/pub/a/oreilly/tim/news/2005/09/30/what-is-web-20.html`

[20] Rao J.; Kungas P.; Matskin M., *Logic-based Web services composition: from service description to process model*, Proceedings of the IEEE International Conference on Web services, p.446, June 06-09, 2004.

[21] Sirin, E.; Parsia, B.; Hendler J., *Template-based Composition of Semantic Web services*, In AAAI Fall Symposium on Agents and the Semantic Web, 2004.

[22] *UPnP Forum*, Website, 2008
`http://www.upnp.org/`

[23] Wu D.; Parsia B.; Sirin E.; Hendler J.; Nau D., *Automating DAML-S Web services Composition Using SHOP2*, In Proceedings of 2nd International Santic Web Conference (ISWC2003), Sanibel Island, Florida, October 2003.

[24] Hourdin, V.; Cheung-Foo-Wo D.; S.L. ; J.Y.T., *Ubiquarium informatique: Une plate-forme pour l'etude des equipements informatiques mobiles en environnement simule.*, In Proccedings of 3-eme Journees Francophones Mobilite et Ubiquite (UbiMob), Paris, September 2006.

Int. J. of Computers, Communications & Control, ISSN 1841-9836, E-ISSN 1841-9844
Vol. V (2010), No. 5, pp. 884-891

A Behavioral Perspective of Virtual Heritage Reconstruction

D.M. Popovici, R. Querrec, C.M. Bogdan, N. Popovici

Dorin-Mircea Popovici, **Crenguţa-Mădălina Bodgan**, **Norina Popovici**
OVIDIUS University of Constanta
124 B-dul Mamaia, 900527, Constanta, Romania
E-mail: {dmpopovici,cbogdan}@univ-ovidius.ro,norinapopovici@yahoo.com

Ronan Querrec
Ecole Nationale d'Ingénieurs de Brest
Laboratoire d'Informatique des Systèmes Complexes
Address: 25 rue Claude Chappe, F-29280 Plouzan, France
E-mail: querrec@enib.fr

Abstract: Our contribution focuses on the behavioral aspects that are currently used in the modeling of the virtual inhabitants of a reconstructed Greek-Roman colony in the framework of the TOMIS project. The project aims at promoting culture by the mean of the reconstruction of historical sites together with their virtual societies based on virtual and/or augmented reality technologies. Our efforts are oriented both on 3D modeling of virtual humans, animation of the virtual humans on every-day human activities and, most important, on the spicing these activities with human emotions. To this end, we iterate the most common agent-based architectures used to produce credible behavior of the virtual agents (humans or animals) in situations inspired from the real world, and emphasize their direct applicability both in humans and animal animations in order to obtain complex behavior based on atomic activities. Finally, the paper presents the technological issues related to the used motion capture technology, as source of high-definition human atomic actions, that participates in complex action plans for virtual agents activities.
Keywords: virtual reality, behavior, motion capture, fuzzy-oriented modeling, virtual agent.

1 Introduction

The 3D historical sites reconstruction is one of the most VR-supported cultural dissemination form. Besides the virtual reconstruction of the old buildings these sites are now animated by means of virtual animals and plants. More important are the virtual humans that perform some individual or collaborative activities. The navigation or exploration of the virtual environment, the triggering of the animation of some mechanisms (cranes, vehicles), and the manipulation of objects are some examples of such activities that users may share with the virtual inhabitants of an ancient site. This is the direction on which we are focused in this paper.

The paper is organised as follows: after a brief presentation of the framework of our effort, in Section 3 we bring into discussion perception, emotion and motivation as main ingredients of a behavioral architecture adopted for our virtual humans. In the sections 4 and 5 we present the application of three behavioral patterns used in the expression of virtual agent action plans. In the last section we discuss the current state of applying motion capture technology in the behavioral modeling of the agent. Finally, we mention some of our directions for the near future and conclusions.

2 TOMIS project context

The major objective of TOMIS project is the development of a multi-sensorial, interactive framework, based on VR/AR technologies that allows the recreation of historical and cultural relics that are inaccessible because of temporal constraints (they existed centuries ago) or geographic constraints (they are placed in submersed areas, great distances or off limits to the general public). To this end we are currently using and developing methodologies and techniques of digital restoration/reconstruction applied on vestiges and artifacts of historical importance, religious and cultural, on the basis of advanced VR/AR technologies. These methodologies and techniques consider also the surrounding elements of flora, fauna and geo-ecology.

The obtained multimodal pilot system will then be evaluated as experimental information technology, both in academic and public setups. To meet all these objectives, the research-development activities proposed in this project are grouped in three distinct phases: a) geometric modeling of the virtual artefacts, fauna, flora and humans; b) behavioural modeling of fauna, flora and population; and c) virtual environment setup together with interaction devices integration. In the following we focus on the second phase, more specific on population behavioral modelling.

Behavioral modelling focuses on the topology of a group, which allows the specialization of the virtual humanoids in accordance with their competences. In other words, its organization is introduced in terms of roles described as capacities (or tasks) and responsibilities of the group members. From our perspective, we consider this organization to be dynamic, explicit and sensitive to the evolution of the environment in which the virtual humans carry out their activities.

To simulate the behavior of all dinamic components which populate the environment, multi-agent systems technology is applied. Each environment element (plant, animal, member of the society or object) has associated an agent that is able to play a scenario. This allows description and compact implementation of a great variety of behaviors, from the simple animation of environmental elements (as fauna, tools or machines used by the people) to the complex interactions of daily human activities.

To this end, we will consider the lowest abstraction level of behavior and primary abilities, attached to society members and elements which populate the environment in which society evolves. On the highest levels of abstraction, we will place the actions (possible collaborative) of society members (in which the user can intervene), which are guided either by collaborative objectives or by individual ones.

3 Behavioral aspects

Due to the environment's dynamics, its own physiological and/or emotional state, and its own motivations, the agent is conditioned to evaluate in every moment of its life time, its behavioral resources, and to decide about the action it will select and express as an answer of all these factors. Consequently, the problem of action selection consists in choosing actions necessary for achieving the current goal. Therefore, frequent compromises have to be made, even independent activities have to be combined. In other words, the behavioral selection result have to permit the agent to reach its goals. To this end, the agent credibility is based on a chain of components that realises this stimulus-reaction relationship. These elements as, perception, motivation and emotion are essential to a credible action selection mechanism.

3.1 The perception

The first step in almost every agent's behavioral architecture is to obtain a sensation, which then it transforms into a perception. This perceptual image that the agent creates is dependent of its competencies, goals, knowledges and abilities. Internal or external stimuli, as active entities, produce a reaction from an excitable organism [1].

The information processing may be made in different manners. Tu and Terzopoulos use a neural architecture that maps the sensorial information in neural inputs of their fishes [2]. They implement a visual and temperature sensor and use a focalisation subsystem that elimines any non-important sensorial information. In [3], the visual sensor is based on computer vision algorithms, being inspired from primates visual aparatus. Different perceptive systems may combine by means of fusion percepts to obtains concepts of a higher level of abstraction, prooving this way a great dependence of the agent on its environment [4]. An active perceptual system can demand that some actions be realized in order to extract supplementary information from environment [5]. This way, the perception is guided by behavioral needs, so by actions that needs to be realized.

After the perception takes place in the agent's aura [6], this perceptual information is passed through the sensorial quality filter [7], and produces a separation between the environment's state and its perception by the agent.

3.2 The motivation

The motivational states are agents' emotional states that determine itself to react somehow by a specific action. Bolles' and Fanselow's [8] model explores the relation between motivational and emotional states, in particular between fear and pain, as emotions that interrupt the brain in order to impose some kind of neccessity or action. For example, if an agent is hungry, then its brains will redirect all its cognitive ressources in order to make him to search the food. This will favorise the satisfaction of its hungry.

Wright uses the motivator term, for an information subclass, such as desires, goals and intentions, which have the potential to trigger an internal or external agent's action [9]. For Aylett, motivation is a long term goal, an emotional or motor state, depending on the domain, and represents the central element of actions' planning algorithm [10]. This way, the motivational states have a major impact in the emotional and decisional processes, so in the behavioral one.

3.3 The emotion

In the theories of emotion, the individual realises a cognitive evaluation of its current state relative to a desired non-riched state in the moment of the evaluation. Reilly [11] proposes as fear model "the likelihood of failing to achieve the current goal" multiplied with "the importance of not failing", while LeDoux [12] affirms that emotion may action at a level much lower than the cognitive one, since the animals may feel the emotions without aware of the cause.

Velasquez [13] uses emotional memories in order to permit agents to chose their actions according to their emotional state. Doing so, the decisional process is directed in an emotion-dependent manner. Gratch and Marsella's [14] agents credibility is based on the obtained emotion, on the evaluation of the relations between the events that appear in a given context and the agents goals and plans. El-Nasr [15] also places the emotion in the center of its architecture. After computing the event's desirability, he uses a version of Ortony's model [16] to define, on Fuzzy logic basis, the resulting emotion against current situation and context.

These are our reasons to consider attention, emotion, and motivation as inhibited/exciting factors of the behavioral answer of the agent.

4 Agent actions

The environment that we model is an ancient Greek-Roman colony situated on the Black Sea coast. Here, the main activities of the population take place around the Tomis colony harbour site where we find different social classes of virtual humans, from sailors and commerciants to simple individuals who are looking to buy some market products. Of course, the place is also spiced by the existence of animals or tehnical devices used in market/harbour maneuvers, as ships, cranes, wheelebarrows, etc. All these elements are modeled by the means of virtual agents, as they are defined in [6]. There the agent is considered as a 5-uple:

$$\mathsf{Ag} = (\mathsf{F}, \mathsf{K}, \mathsf{Rec}, \mathsf{Eff}, \mathsf{Dec}) \tag{4.1}$$

where F is the set of the agent's attribute shapes, K represents the agent's knowledge, *Rec* the set of receptors, *Eff* the set of this effectors and *Dec* is the decisional module. While the knowledge is considered as a set of valued concepts and the decidor is a FCM-based module, the receptors and effectors are active entities responsible with the actualisation of sensorial and motors concepts in the decider (fig. 1). Moreover, the effectors, which encapsulate the agent changes as imperative methods in containers of activity, are activated/deactivated at the decider level where the agent realises the action selection.

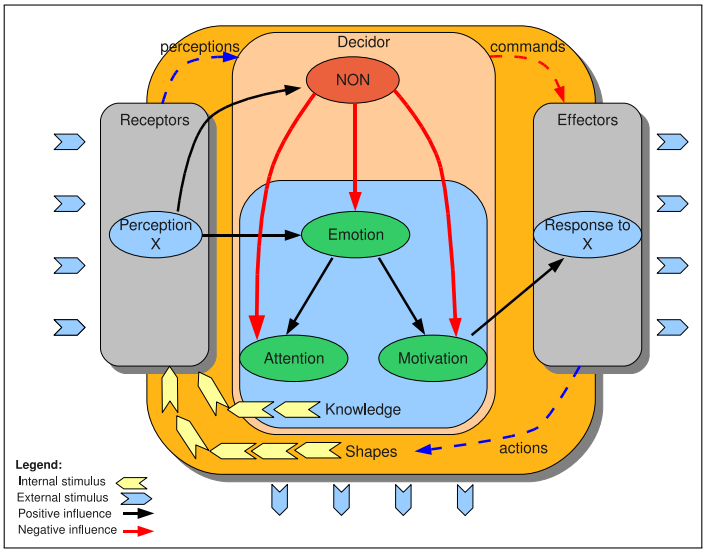


Figure 1: Affective influences in the virtual agent architecture [6]

Obtaining a perception at the level of a *Rec* produces the appearance of an emotion that triggers the increasing of the attention in what it concerns future perceptions and also the increasing of the motivation to displaying an answer at this perception, by the means of agent effectors. In the same time, a high rate of non-motivation triggered by another perception (even internal to the agent) may produce a strong inhibition of the perception and so of the emotion and answer.

The problem is that even we can determine which will be the agent's a answer to a specific perception, we cannot determine its answer in a complex situation, as in the real environment case. And this, because of the multitude of the obtained perceptions.

5 Action selection

There are two levels toward which we oriented the behavioral modeling in the Tomis colony. The simplest level does not suppose semantic abstraction and is approached by means of behavioral animation. This is the level where the simplest behaviors are implemented, and called atomical actions of the virtual agents. This is the case for seagulls, cranes, ships and other elements animation. In order to filter even this kind of animation we implemented a level of behavior filter that depends only on the distance between the agent and the viewer.

If an element may support different behavioral animations, we chose to implement a finite state machine selection mechanism in order to be able to switch from one state (that corresponds to one behavioral animation) to another. This is the case for "hitting with hammer", "draw/push a dustcart", etc. This means that it is possible that an action depends on the existence of some resources in the very near vecinity of the agent.

These lower abstracted agent capabilities are then used at the next level of abstraction, in the realisation of highly complex behaviors. The high level one is represented by actions as: Load/Unload ship, Taking goods, Guard a zone, Buy a market product, etc. Here, the action may be so complex that she needs a plan of realisation, based on simpler actions.

By using the behavioral patterns, FOF (firstOf), ALL (all) and SEQ (sequence), as defined in [6], we are able to express sequential (SEQ) actions, as well as collaborative (ALL) or even concurrential (FOF) ones. Let us give some example of such actions.

We identified two types of virtual humans: one that express individual behavior, and that plays roles as Porter, Buyer, Merchant, Publican, Teamster; and another that express group behavior, and that plays roles as GroupMember, Soldier / Guardian (despite the fact that he behaves alone, he is part of the Group), as well as Rower, Pairs, Captain. At the level of group behaviors we adopted a boid-oriented solution [17], either by introducing a leader inside the ierarchies (as for Soldier / Rower ...), either by letting the virtual agents to organise themselves (as for GroupMember) without having neccessary a leader.

No matter what is the virtual environment state, the planning of one or several virtual humans' behavior consist mainly in movement allocation by the means of their effectors.

To exemplify an action plan, we adopted a goal oriented approach and consider the high level action "Transport a thing". We chosed to decompose this complex activity into simpler actions, until we reach the atomic actions level for each task. This gives us the following actions sequences:

```
Transport<T>From<S>To<D>  |        SearchFor<S>
  - SearchFor<S>          |          - LookingFor<S>
    - Reach<S>            |          - AskingFor<S>
  - SearchFor<T>          |          - ExploreFor<S>
    - Reach<T>            |
  - Take<T>               |
  - SearchFor<D>          |        Take<T>
    - Reach<D>            |          - Tilt<T>
  - Release<T>            |          - Touch<T>
  - Explore<>             |          - Straightening<T>
```

`SearchFor<S>` means that the agent may explore the environment by looking arround for `<S>` and if he/she meets someone else it may ask for `<S>`. This is an example of using `FOF` operator: `SearchFor<T>=FOF(LookingFor<T>,AskingFor<T>,ExploreFor<T>)`.

Once the information is obtained, he/she triggers to `Reach<S>`. `Reach<S>` is considered that maybe realized just by relatively simple movements and obstacle avoidance.

Take<T> and Release<T> are two complex opposite actions because the resources they use (in order to take a big object we need both hands). In other words:

```
Take<T>=SEQ(Tilt<T>,Touch<T>,Straightening<T>)
Release<T>=SEQ(Tilt<T>,Free<T>,Straightening<T>)
```

For prooving the ALL operator we will change the action in the market place. Here we suppose to have a virtual human that have to buy several products without knowing exactly if he/she will find the products in the market and where this products are placed. So, for example, let us suppose that the human want to buy some perfume (PE), one crater[1](CR), and some wine (WI). To express this, we use the ALL operator as follows:

```
Buy<PE,CR,WI>=ALL(Buy<PE>,Buy<CR>,Buy<WI>)
```

For a "buy" activity that is supposed to give the agent the market product <T> after the payment is made, we may consider a sequence like Buy<T>=SEQ(Search<T>,Reach<T>,Take<T>). But when the agent has to buy more than one product, this sequentiality is broken by the inpredictibility of the existence and the topology of the market products. So, we chosed to express the "buy" action as follows:

```
Buy<PE,CR,WI>=SEQ( SEQ(Seach<PE>,Search<CR>,Search<WI>),
                   ALL( SEQ(Reach<PE>,Take<PE>),
                        SEQ(Reach<CR>,Take<CR>),
                        SEQ(Reach<WI>,Take<WI>) ) )
```

This means that our agent will first evaluate the environment and then he will proceed to achieve the market products according to their accesibility.

Using these behavioral patterns we are also able to decide the failure of actions. To this end, we use time restricted version of operators. So, if the agent fails to complete an action in the corresponding time interval, then the agent will drop the action and will evaluate if it is coherent to continue the current action plan or to change it.

The agent motivation and the resources accesible to the agent are essential. The agent action have to have sense, i.e. to be accorded to its internal state, its perceptions, knowledge about its environment and its capabilities; the agent have to equilibrate its actions between an opportunist behavior and the goal oriented one.

Last but not least, the agent may choose to make a compromise between multiple concurrent behaviors for satisfy a maximum number of goals in the same time.

6 Some techniques

As we have already said, the behaviors were implemented using very different techniques, starting from behavioral animation and ending with motion capture solutions.

Behavioral animations were obtained either as morphing shapes (as is the case of seagulls), either directly by procedural animation (as is the case for plants or simple artifacts, as cranes, etc). Nevertheless, a level of behavior was implemented in order to filter the displayed behaviors according to the user's field of view.

Human natural gestures and actions are obtained using motion capture technology. Here challenge was to apply the real-time captured animation to the existed models in order to enhance these 3D models with animation information. Once the animation fitted, we stored the new

[1]Crater - small container for mixing wine and water.

models in MD5 or SMD file formats for later use. In the phase of the project we used software tools as 3DS Max [18] and Blender [19] for 3D modeling and Arena software [20] for real-time motion capture. In the figure 2 are presented some examples of the human natural captured actions that are applied at the level of virtual humans in our project.

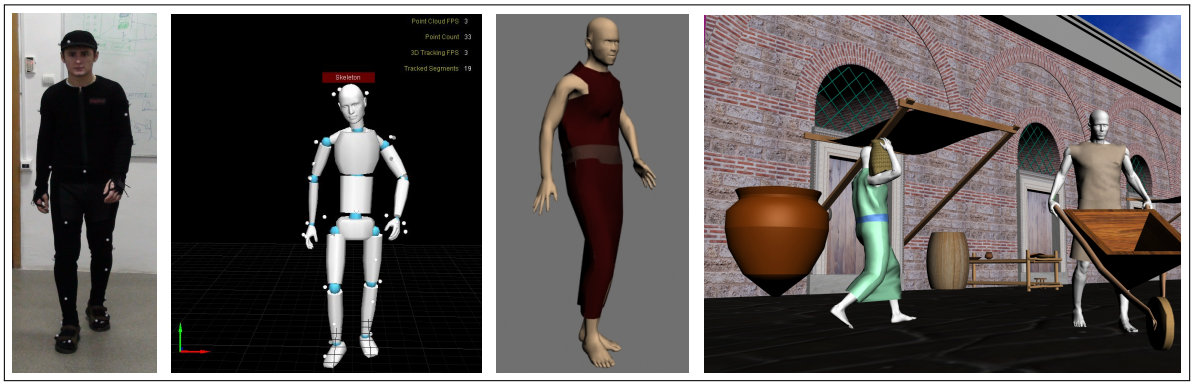


Figure 2: Simple walk action for the real actor, the skeleton, the virtual human, and some other actions applied to virtual humans in TOMIS project.

7 Conclusions and Future Works

The credibility of the user experience in the reconstructed environment is augmented by the behavior realism of the virtual humans that the user meets. To this end, we invoked in our solution both emotional aspects and technical ones, and explain how they integrates in the adopted agent architecture. Once the action selection mechanism is tested the mix of several captured motion for complex actions will provide the expected realism to the virtual humans.

8 Acknowledgements

Our work is funded by the National Centre of Programs Management through TOMIS project (PN II: 11–041/2007). We also thank to our enthousiastic volunteer members of CERVA team Alexandru Dinca, Manuel Galiu, Ciprian Ilie, Daniela Panait and Mihai Polceanu. For more information please visit our web site http:/www.cerva.ro.

Bibliography

[1] N. Richard, *Description de comportements d'agents autonomes évoluant dans des mondes virtuels*, PhD thesis, Ecole Nationale Supérieure des Télécommunications, Paris, 2001.

[2] X. Tu, D. Terzopoulos, *Artificial fishes: Physics, locomotion, perception, behavior*, Computer Graphics, 28(Annual Conference Series):43-50, 1994.

[3] J.J. Kuffner, *Autonomous Agents for Real-Time Animation*, PhD thesis, Department of Computer Science of Standford University, 1999.

[4] R.C. Arkin,*Behavior-based robotics*, MIT Press, 1998.

[5] D. Thalmann, H. Noser, Z. Huang *Autonomous virtual actors based on virtual sensors*, Lecture Notes in Artificial Intelligence (LNCS), 1997.

[6] D.M.Popovici,*Modelling the space in virtual universes*, PhD thesis, Politehnica University of Bucharest, Romania, 2004.

[7] D. Isla, B. Blumberg, *New challenges for character-based ai for games*, In AAAI Spring-Symposium on AI and Interactive Entertainment, 2002.

[8] R.C. Bolles, M.S. Fanselow, *A perceptual defensive recuperative model of fear and pain*, Behavioral and Brain Sciences, 3:291-301, 1980.

[9] I. Wright, *Emotional Agents*, PhD thesis, University of Birmingham, 1997.

[10] R. Aylett, A. Coddington, G. Petley, *Agent-based continuous planning*, In Proceedings of the 19-th Workshop of the UK Planning and Scheduling Special Interest Group (PLANSIG 2000), 2000.

[11] W.S.N. Reilly, *Believable Social and Emotional Agents*, PhD thesis, Carnegie Mellon Univ, 1996.

[12] J. LeDoux, *The Emotional Brain*, New York: Simon and Schuster, 1996.

[13] J.Velasquez, *When robots weep: Emotional memories and decision-making*, In Proceedings of the Fifteenth National Conference on Artificial Intelligence, Madison, Wisconsin, 1998. AAAI Press.

[14] J. Gratch, S. Marsella, *Tears and fears: modeling emotions and emotional behaviors in synthetic agents*, In Jorg P. Muller, Elisabeth Andre, Sandip Sen, and Claude Frasson, editors, Proceedings of the Fifth International Conference on Autonomous Agents, pages 278-285, Montreal, Canada, 2001. ACM Press.

[15] M.S. El-Nasr, J. Yen, T.R. Ioerger, *Flame-fuzzy logic adaptive model of emotions*, Autonomous Agents and Multi-Agent Systems, 3(3):219-257, 2000.

[16] A. Ortony, G.L. Clore, A. Collins, *The Cognitive Structure of Emotions*, Cambridge University Press, New York, 1988.

[17] C.W. Reynolds, *Flocks, herds and schools: a distributed behavioral model*, Computer Graphics (SIGGRAPH'87), 21(4):25-34, 1987.

[18] http://www.autodesk.com

[19] http://www.blender.org

[20] http://www.naturalpoint.com/

[21] C. Rose, B. Bodenheimer, M.F. Cohen, *Verbs and Adverbs: Multidimensional Motion Interpolation Using Radial basis Functions*, IEEE Computer Graphics and Applications, 18:32-40,1998.

Int. J. of Computers, Communications & Control, ISSN 1841-9836, E-ISSN 1841-9844
Vol. V (2010), No. 5, pp. 892-899

Information Sharing in Vehicular AdHoc Network

A. Rahim, Z.S. Khan, F.B. Muhaya, M. Sher, M.K. Khan

Aneel Rahim, Zeeshan Shafi Khan
1. Prince Muqrin Chair for IT Security,
King Saud University, Saudi Arabia
2. International Islamic University, Pakistan
E-mail: aneelrahim,zeeshanshafi@ksu.edu.sa

Fahad Bin Muhaya
Prince Muqrin Chair for IT Security,
King Saud University, Saudi Arabia
E-mail: fmuhaya@ksu.edu.sa

Muhammad Sher
International Islamic University, Pakistan
E-mail: m.sher@iiu.edu.pk

Muhammad Khurram Khan
Center of Excellence in Information Assurance,
King Saud University, Saudi Arabia
E-mail: mkhurram@ksu.edu.sa

Abstract: Relevance Technique broadcast the useful information and removes the redundant data. 802.11e protocol implementation has certain flaws and is not suitable for VANETs scenarios. Main issue in 802.11e protocol is internal sorting of packets, no priority mechanism within the queues and often lower priority traffic get more medium than high priority traffic. In this paper, the mathematical model of relevance scheme is enhanced so that it can consider the network control in real scenario by considering the impact of malicious node in network. Problems of 802.11e protocol can be resolved by making virtual queue at application level. We analyze the comparison of simple virtual queue with the over all impact of virtual queue and mathematical model. Similarly we compare the mathematical model with over all impact of virtual queue and modified mathematical model using NS-2 simulator.
Keywords: VANETs, Broadcast, 802.11e, Malicious

1 Introduction

Vehicle to Vehicle (V2V) communication enhances the safety of passenger and driver [1]. V2V communications is unreliable because of shadowing, Doppler shifts and multi-path fading. [2]

Security is an important concern in mobile adhoc network [3] [4] [5]. Attacks are easily launched on VANETs [6] because of high speed [7], no infrastructure and topology changes frequently [8] [9]. Several Security attacks are possible on safety application, which includes Denial of Service [10], Masquerade [11], fake information, false position information and ID disclosure [12]. Vehicular communication vulnerabilities are explained in [13] which include Jamming, Forgery, In-transit Traffic Tampering, Impersonation, Privacy Violation and On-board Tampering. Malicious data in VANET is because of distributed environment and unreliable components of data generation. [14] We in this paper resolve the problems of 802.11e protocol by making virtual queue at application and enhance the mathematical model of message benefit to consider

the network traffic. We measure the global benefit in real scenario by considering the impact of malicious node. We also show the comparison of simple virtual queue with the over all impact of virtual queue and mathematical model. Similarly we compare simple mathematical model with over all impact of virtual queue and mathematical model. This paper is organized as follows: In section 2, we discuss relevance based approach, its characteristics and its implementation using cross layer, 802.11e and 802.11e with virtual queue. In section 3, proposed study and results are presented using NS-2. Lastly in section 4 conclusions is given.

2 Related Work

Relevance Technique disseminates the useful information and removes the redundant data [16]. Vehicle contains huge information that can't be shared to due the high speed of Vehicles. Technique which gives High priority traffic more medium as compare to low priority traffic is a suitable approach for VANETS. So relevance technique is the only option as it forward data according to its relevance.

Relevance Techniques based upon the calculation of relevance value of message and its distribution according to its priority [15, 16]. Altruism, Application-oriented information differentiation, Controlled Unfairness is the basic characteristics of relevance based approach [15, 17, 18].

2.1 Cross Layer and 802.11 e Implementation

Relevance Technique can be implemented through cross layer design or by 802.11e protocol. In cross layer design, relevance value of every packet is measured at application layer and pass to link layer through packet header. Modified medium access control and interface queue broadcast the high priority traffic with help of application layer information. [15, 17]. 802.11e protocol implementation has certain flaws and is suitable for VANETs scenarios [18]. Main issue in 802.11e protocol is that it does not provide internal sorting of packets, no priority mechanism within the queues and performance of the network degrades as lower traffic some times get more medium than high priority traffic [15].

2.2 802.11e Implementation with Virtual Queue

802.11e protocol problems are overcome by adding four virtual queues at application level. Packets are sorted according to their priority and most importance messages are near the head of queues. Sorting is done by getting the current information from application layer and length of 802.11e is set to be one. When 802.11e is empty, one high priority packet is moved from virtual queue to 802.11e queue. Packet in 802.11e queue does not mean that it always broadcast. If get a packet in virtual queue that has higher relevance than packet in 802.11e queue than we swap both the packets in order to achieve higher global benefit [19].

2.3 Mathematically Model for Relevance Based Approach

The mathematical model for relevance based approach is given below.

$$\text{Message Benefit} = \frac{1}{\sum_{i=0}^{n} \alpha_i} * \sum_{i=0}^{n} \alpha_i * b_i(m, v, i) \text{ ---------- [18].}$$

To determine the relevance value of message, Message (m), Vehicle (v), and Information (i) context parameters are used. The N parameters are computed with the help of application dependent function bi. The N parameters are then weighted with application dependent factors ai. In the end, all parameters are sum up and divided by the sum of all a_i.

2.4 Enhanced Mathematically Model for Relevance Based Approach

As the existing model does not have the support for network control traffic. So network performance can improve by adding the network control in the mathematical model [20].

Enhanced Message Benefit $= \frac{1}{\sum_{i=0}^{n} \alpha_i} * \sum_{i=0}^{n} \alpha_i * b_i(m, v, i) + \sum_{i=0}^{n} P_i$

a) $\sum_{i=0}^{n} P_i = 0$ if it is user traffic

Where as

$\sum_{i=0}^{n} P_i = 1$ for Operational level network problem, $\sum_{i=0}^{n} P_i = 2$ for Administrative level, $\sum_{i=0}^{n} P_i = 3$ for Maintenance level

b) Message Benefit $= \sum_{i=0}^{n} P_i$ (for Network Traffic only)

If $0 \gneq \sum_{i=0}^{n} P_i \lneq 3$ Then $\frac{1}{\sum_{i=0}^{n} \alpha_i} * \sum_{i=0}^{n} \alpha_i * b_i(m, v, i) = 0$

User and Network traffic is assigned a value between zero and three, in order to handle them easily with four queues (Q0, Q1, Q2, and Q3) of 802.11e. High priority traffic is assigned Q_0 so that it can be forward before the packets in Q1, Q2 and Q3. Queues are assigned to user and network control traffic according to there relevance value. But in the existing approach there is no mechanism for priority for network control traffic. So the global benefit is enhanced by considering network traffic.

3 Proposed Study and Results

In this study we simulate the relevance based approach and calculate global benefit in ideal scenario that all nodes are doing their properly and there is no malicious node in the network. In the second scenario we consider the impact of malicious node and measure how much global benefit is decreased. The malicious nodes forward the relevant messages first but also inject some surplus information. In last scenario malicious node forward the surplus message first and ignore the relevant message. In order to validate the proposed study, we compare the performance of relevance based approach in real and ideal scenario with 802.11e protocol. NS-2, a network simulator [21], is used to simulate the behavior for relevance based approach in VANETs scenarios. We use Manhattan Mobility Model and traffic is generated by Generic Mobility Simulation Framework [22]. Vehicles are moving at a speed of 72Km/hr to 108 Km/hr within an area of 3000m x 3000m with transmission range of 300m. Performance of relevance based approach is measured by calculating the global benefit.

Table 1: Simulation Parameters

Parameters	Settings
Channel	Wireless
Vehicles	50,100,150
MAC protocol	802.11e
Time	50s
Routing Protocol	DSDV

Network Simulator is used for the simulation and different parameter used in the following study is given in Table 1.

3.1 Improvement due to Mathematical Model

In this study we simulate enhanced mathematically model of message benefit shown above with existing relevance based approach. Figure 1(a) shows that 50 vehicles are moving at high speed and share safety and comfort information with each other. Relevance based approach consider only user traffic and ignore network traffic. So its global benefit can be improved by improving the mathematical model. We now evaluate the performance of relevance based approach by adding the network control parameter in the existing formula. Figure 1(a) shows the global benefit with enhanced relevance based approach. It is clear from figure 1(a) that global benefit is low by using existing relevance based approach because network control traffic set lower priority and get less bandwidth than user traffic. So lower priority traffic can get more bandwidth than higher priority traffic. That's why the global benefit is improved by adding the network parameter in relevance based approach.

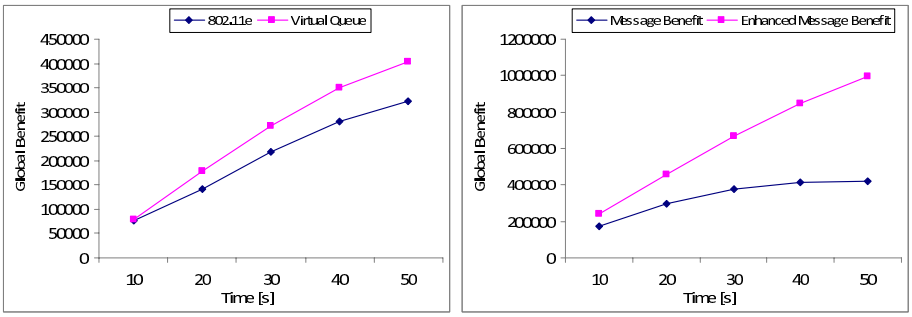


Figure 1: (a) Improvement due to mathematically (b) improvement due to virtual queue

3.2 Improvement due to Virtual Queue

Figure 1(b) shows simple 802.11e and virtual queue with 802.11e, safety messages and route messages are forwarded by vehicles. In this study 150 vehicles exchanging information with each other. In simple 802.11 e, there is no mechanism of priority assignment. This problem is resolved by virtual queue. So its global benefit is greater than simple 802.11 e because it does not allow lower priority traffic to get more medium than higher priority traffic.

3.3 Improvement due to Virtual Queue and Mathematically Model

First we check the improvement due mathematical model and virtual queue separately but now we consider the impact of both on the global benefit of network. Figure 2 shows that global benefit of existing and enhance relevance based approach due to virtual queue and mathematically

model. Enhanced relevance based approach has higher global benefit because it resolves the problem of priority mechanism and ignorance of network control traffic.

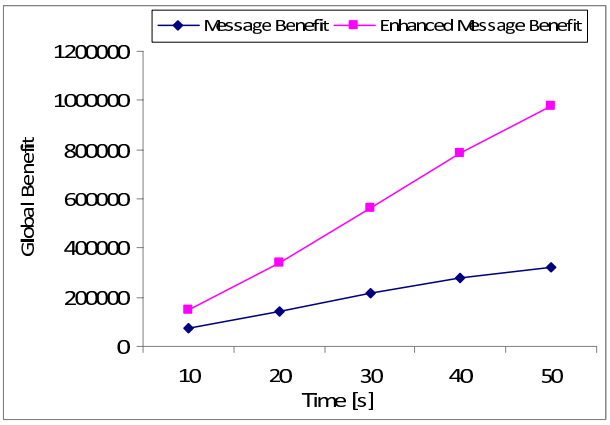


Figure 2: Improvement due to virtual queue and Mathematically model

3.4 Comparison

This study shows the comparison of simple virtual queue with the over all impact of virtual queue and mathematical model. Similarly we compare simple mathematical model with over all impact of virtual queue and mathematical model. Fig 3(a) shows the global benefit due to Message Benefit (MB), enhance message benefit (EMB) and virtual queue + EMB. It is clear from figure that global benefit by using virtual queue + EMB is greater than simple EMB because within a queue a there is no priority mechanism available.

Fig 3(b) shows the global benefit due to 802.11e, Virtual Queue and EMB + virtual queue. It is clear from figure that global benefit by using EMB + virtual queue is greater than 802.11e and simple queue because in simple queue we don't have discriminate between user traffic and network traffic.

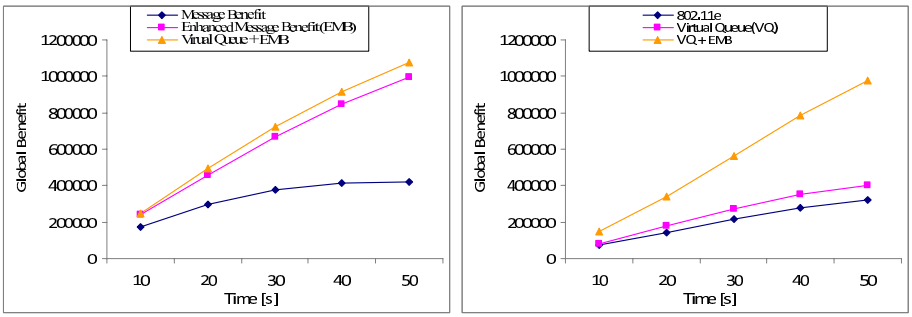


Figure 3: (a)Comparison of mathematically model with both virtual queue and mathematical model (b) Comparison of virtual queue with both mathematical models

3.5 Impact of Malicious node

In this study we consider the impact of malicious node on EMB, Virtual Queue and both (EMB + virtual queue).Figure 4(a) shows that 50 vehicles are moving at high speed and share safety and comfort information with each other. First we simulate the MB and EMB in ideal

scenario that no malicious node exists and all nodes try to improve the benefit of network rather than their own benefit. After that we simulate EMB in real scenario that malicious exist and damage the performance of the network. Figure 4(a) shows that global benefit of EMB in real scenario lies between the EMB MB in ideal scenario.

Figure 4(b) shows that 150 vehicles exchanging information with each other. First we simulate the 802.11e and virtual queue in ideal scenario that no malicious node exists and all nodes try to improve the benefit of network rather than their own benefit. After that we simulate Virtual queue in real scenario that malicious exist and damage the performance of the network. Figure 4(b) shows that global benefit of EMB in real scenario lies below than 802.11e and Virtual Queue in ideal scenario.

Figure 5 shows that 150 vehicles are moving at high speed and share safety and comfort information with each other. First we simulate the MB and VQ + EMB in ideal scenario that no malicious node exists and all nodes try to improve the benefit of network rather than their own benefit. After that we simulate VQ+ EMB in real scenario that malicious exist and damage the performance of the network. Figure 5 shows that global benefit EMB + VQ in real scenario lies between the EMB +VQ and MB in ideal scenario.

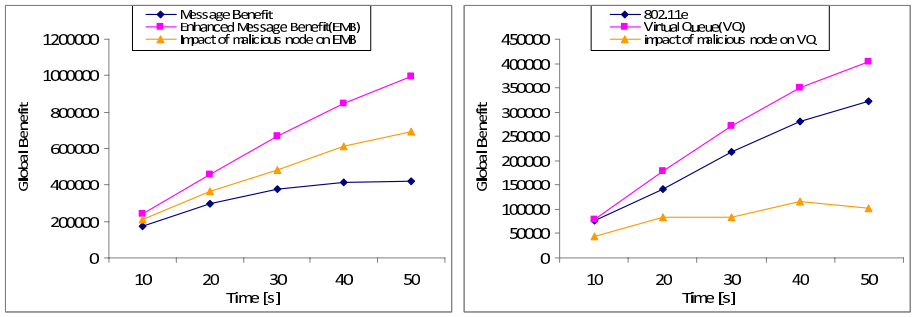


Figure 4: (a) Impact of malicious node on EMB (b)Impact of malicious node on Virtual Queue (VQ)

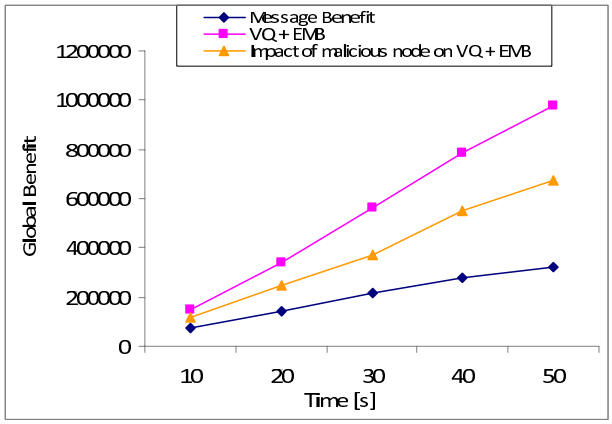


Figure 5: Impact of malicious node on Enhanced Message Benefit and Virtual Queue

4 Conclusion

Relevance scheme rely on intermediate node for communication so it consider there is no selfish node exist in network. However it is not possible in real scenario. We in this paper simulate

the relevance based approach using 802.11e, virtual queue with 802.11e and enhance message benefit in real and ideal scenario. Simulation results shows that global benefit is improved by using virtual queue with enhance mathematical model.

5 Acknowledgments

This research is supported by the Prince Muqrin Chair (PMC) for IT Security at King Saud University, Riyadh, Saudi Arabia.

Bibliography

[1] Y. Wu, L. Yang, G. Wu, J. Guo, An Improved Coded Repetition Scheme for Safety Messaging in VANETs, *IEEE*, 2009.

[2] R. K. Shrestha, S. Moh, I. Chung, D. Choi, Vertex-Based Multihop Vehicle-to-Infrastructure Routing for Vehicular Ad Hoc Networks, *IEEE, Proceedings of the 43rd Hawaii International Conference on System Sciences* 2010.

[3] H. Kumar, , R.K Singla,., S. Malhotra,, Issues and Trends in AutoConfiguration of IP Address in MANET, *International Journal of Computers Communications and Control, Volume:3, Supplement: Suppl. S, pp. 353-357*, 2008.

[4] M. A Rajan, M. G Chandra, L. C. Reddy, P. Hiremath, Concepts of Graph Theory Relevant to Ad-hoc Networks, *International Journal of Computers Communications and Control, Volume:3, Supplement: Suppl. S, pp. 465-469*, 2008.

[5] J. Sun, C. Zhang, Y. Zhang, Y. Fang, An Identity-Based Security System For User Privacy in Vehicular Ad Hoc Networks, *IEEE Transactions On Parallel And Distributed Systems*, 2010

[6] S. Dietzel, E. Schoch, B. Konings, M. Weber, Resilient Secure Aggregation for Vehicular Networks, *IEEE Network, vol 24 pp 26-31*,2010

[7] I. Jang, W. Choi, H. Lim, A Forwarding Protocol with Relay Acknowledgement for Vehicular Ad-Hoc Networks, *IEEE* 2008.

[8] H. L. Nguyen, U. T. Nguyen, Study of Different Types of Attacks on Multicast in Mobile Ad Hoc Networks, *International Conference on Mobile Communications and Learning Technologies, IEEE* , 2006.

[9] S. Mao, S. Lin, S. S. Panwar, Y. Wang, E. Celebi, Video Transport Over Ad Hoc Networks: Multistream Coding With Multipath Transport, *IEEE Journal on Selected areas in Communications, vol. 21, no. 10*, December 2003

[10] B. R. Moyers, J. P. Dunning, R. C. Marchany, J. G. Tront, Effects of Wi-Fi and Bluetooth Battery Exhaustion Attacks on Mobile Devices, *Proceedings of the 43rd Hawaii International Conference on System Sciences* , 2010.

[11] K.A. Bakar, B. S. Doherty, Evaluation of the Recorded State Mechanism for Protecting Agent Integrity Against Malicious Hosts, *International Journal of Computers Communications and Control, Vol.3, No.1, pp. 60-68*, 2008

[12] M. Raya, J-P. Hubaux. The security of vehicular ad hoc networks. *In Workshop on Security in Ad hoc and Sensor Networks (SASN)*, 2005.

[13] M. Raya, P. Papadimitratos,J.-P. Hubaux, Securing Vehicular Communications, *In IEEE Wireless Communications Magazine, Special Issue on Inter-Vehicular Communications*, October 2006.

[14] K. Sha,, S. Wang, W. Shi, RD4: Role-Differentiated Cooperative Deceptive Data Detection and Filtering in VANETs, *IEEE Transactions On Vehicular Technology, vol. 59, no. 3*, March 2010.

[15] T. Kosch, C. J. Adler, S. Eichler, C. Schroth, M. Strassberger, The scalability problem of vehicular ad hoc networks and how to solve it, *IEEE Wireless Communications* , October 2006.

[16] C. Adler, S. Eichler, T. Kosch, C. Schroth, M. Strassberger, Self-organized and Context-Adaptive Information Diffusion in Vehicular Ad Hoc Networks, *3rd International Symposium on Wireless Communication Systems*, 2006.

[17] S. Eichler, C. Schroth, T. Kosch, M. Strassberger, Strategies for context-adaptive message dissemination in vehicular ad hoc networks, *Second International Workshop on Vehicle-to-Vehicle Communications*, July 2006.

[18] C. Schroth, R. Eigner, S. Eichler, M. Strassberger, A Framework for Network Utility Maximization in VANETs, *ACM International Conference on Mobile Computing and Networking, USA* September 29, 2006.

[19] A. Rahim, M. Yasin, I. Ahmad, Z. S. Khan, M. Sher, Relevance Based Approach with Virtual Queue Using 802.11e protocol for Vehicular Adhoc Networks , *2nd International conference on Computer, Control and Communication, Karachi*, 14 Feb 2009.

[20] A. Rahim, F. B. Muhaya, Z. S. Khan, M.A. Ansari, M. Sher, Enhance Relevance based approach for Network Control Relevance, *Accepted in Infomatica Journal ISSN: 0350-5596.*

[21] Network Simulator, ns2 *http://www.isi.edu/nsnam /ns*

[22] R. Baumann, F. Legendre, P. Sommer, Generic Mobility Simulation Framework (GMSF), *ACM MobilityModels'08*, Hong Kong SAR, China, May 26, 2008

Int. J. of Computers, Communications & Control, ISSN 1841-9836, E-ISSN 1841-9844
Vol. V (2010), No. 5, pp. 900-909

E-Health System for Medical Telesurveillance of Chronic Patients

C. Rotariu, H. Costin, I. Alexa, G. Andruseac, V. Manta, B. Mustata

Cristian Rotariu, Hariton Costin
1. "Gr. T. Popa" Univ. of Medicine and Pharmacy
Kogalniceanu No. 9-13, Iasi, Romania and
2. Institute for Computer Science, Romanian Academy
Carol I No. 11, Iasi, Romania
E-mail: crotariu74@yahoo.com, hcostin74@yahoo.com

Ioana Alexa, Gladiola Andruseac
"Gr. T. Popa" Univ. of Medicine and Pharmacy,
Kogalniceanu No. 9-13, Iasi, Romania E-mail: agladi@yahoo.com

Vasile Manta
"Gh. Asachi" Technical University,
D. Mangeron No. 27, Iasi, Romania
E-mail: vmanta@cs.tuiasi.ro

Bogdan Mustata
ROMSOFT Ltd.
Sulfinei No. 18, Iasi, Romania
E-mail: mbo@rms.ro

Abstract: The current common goal in medical information technology today is the design and implementation of telemedicine solutions, which provide to patients services that enhance their quality of life. Advances in wireless sensor network technology, the overall miniaturization of their associated hardware low-power integrated circuits and wireless communications have enabled the design of low-cost, miniature, and intelligent physiological sensor modules with applications in the medical industry. These modules are capable of measuring, processing, communicating one or more physiological parameters, and can be integrated into a wireless personal area network. This paper is dedicated to the most complex Romanian telemedical pilot project, TELEMON, which has as goals design and implementation of an electronic-informatics-telecommunications system, that allows the automatic and complex telemonitoring, everywhere and every time, in (almost) real time, of the vital signs of persons with chronic illnesses, of elderly people, of those having high medical risk and of those living in isolated regions. The final objective of this pilot project is to enable personalized medical teleservices delivery, and to act as a basis for a public service for telemedical procedures in Romania and abroad.
Keywords: telemedicine, telemonitoring, biomedical devices, wireless personal area network.

1 Introduction

Telemedicine is part of the expanding use of communications technology in health care and is used in prevention, disease management, home health care, long-term care, emergency medicine, and other applications.

The proposed system, called TELEMON, enables to design a secure digital transmission (medical records, digital images, video, and text) and a secure medical records acquisition system in order to enhance the telemedical consultancy services. The main objective of this project

is to enable personalized teleservices delivery and patient safety enhancement based on an earlier diagnosis with medical telemetry using biosignals, images [1], text transmissions, and also applying the suitable treatment according to the remote medical experts' recommendations [2].

Our project allows persons with different (chronic) diseases and to elderly/lonely people to be monitored from medical and safety points of view. In this way the medical risks and accidents will be diminished. The TELEMON system will act as a pilot project destined to the implementation of a public e-health service, "everywhere and every time", in real time, for people being in different hospitals, at home, at work, during the holidays, on the street, etc.

2 Materials and Methods

The main objective of this project is the achievement of an integrated system, mainly composed by the following components in a certain area: a personal network of wireless transducers (PNWT) on the ill person (Figure 1), a data multiplexing block and a personal server (PS) in form of a Personal Digital Assistant (PDA). After local signal processing, according to the specific monitored feature, the salient data are transmitted via one of internet or GSM/GPRS to the database server of the Regional Telemonitoring Centre. The PNWT includes medical devices for vital signs (ECG, heart rate, arterial pressure, oxygen saturation, body temperature), a fall detection module, a respiration one, all these components having radio micro-transmitters, which allows an autonomic movement of the subject. The data processing will be performed by the PDA.

The results of data processing are in principal and if necessary different locally generated alarms, transmitted to the central server. Other results on server data processing will be different medical statistics, necessary for the evaluation of health status of the subject, for the therapeutic plan and for the healthcare entities.

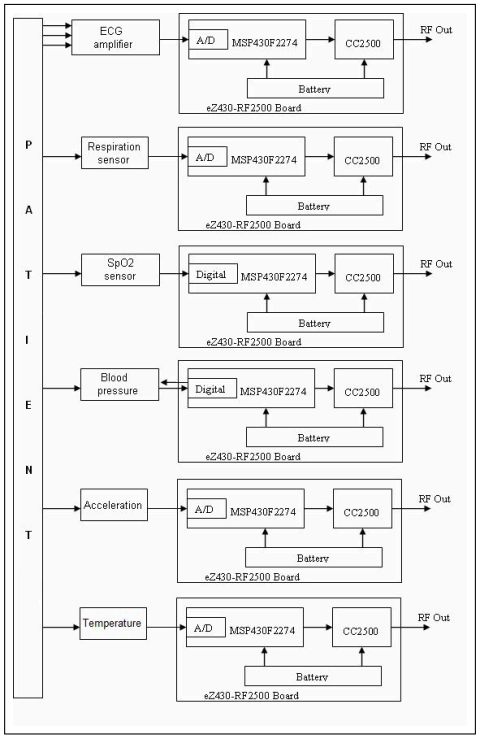


Figure 1: The local subsystem for home monitoring of the patient

a) a 3-leads ECG module records and transmits data through a radio transceiver interface;
b) the oxygen saturation module (SpO2), that also computes the cardiac rhythm;
c) the arterial pressure module, with serial interface;
d) the body temperature module;
e) the respiration module;
f) the fall detection module.

The modules (a), (d), (e) were made by our research team, while module (b), (c), and (f) were chosen from the market and were integrated in TELEMON system.

These modules transmit data to a PDA through radio transceivers, operate in the 2.4GHz band, and have 5m/10m range indoors/outdoors.

Our wireless personal area network is realized by using a custom developed sensors modules for physiologic parameters measurement and a low power microcontroller board (eZ430-RF2500 Board from Texas Instruments). The network is wirelessly connected to a personal server that receives the information from sensors.

The eZ430-RF2500 is a complete MSP430 [17] wireless development tool providing all the hardware and software for the MSP430F2274 microcontroller and CC2500 2.4GHz wireless transceiver [18]. Operating on the 2.4 GHz unlicensed industrial, scientific and medical (ISM) bands, the CC2500 provides extensive hardware support for packet handling, data buffering, burst transmissions, authentication, clear channel assessment and link quality. The radio transceiver is also interfaced to the MSP430 microcontroller using the serial peripheral interface.

The 3-lead ECG amplifier (Figure 2) is a custom-made device. It has for each channel a gain of 500, is DC coupled and has a cut-off frequency around 35 Hz. The high common mode rejection (>90 dB), high input impedance (>10 M Ω), the fully floating patient inputs are other features of the ECG amplifier.

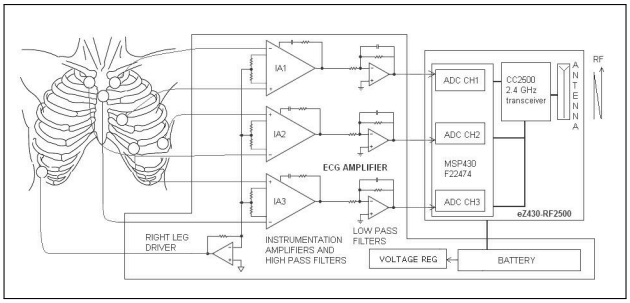


Figure 2: The ECG amplifier (block diagram)

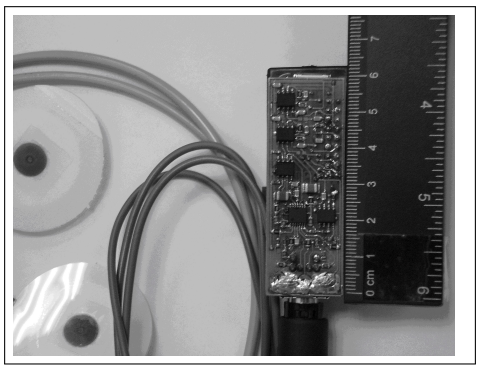


Figure 3: The 3-leads ECG module

Two AAA 1.2V rechargeable batteries power the ECG amplifier through a voltage regulator. The regulator is built around a capacitive DC/DC step-up converter.

The process of recognition of the ECG waves (Figure 4) constitutes a significant part of the most ECG analysis systems. In applications were rhythm detection is performed, only the location of the R wave is required. In other applications it is necessary to find and recognize the features of the ECG signal, such as the P and T waves, or the ST segment, for the automated classification and diagnosis. Many algorithms for the extraction of the ECG features based on the digital filters have been reported in the literature [13], [14] and [15] especially algorithms for the QRS complex recognition. The main effort in the ECG features extraction is for finding the exact location of the waves. After that, the determination of the wave's amplitudes and shapes is much simpler. The strategy for finding the exact location of the waves is to first filter the ECG signal and then recognize the QRS complexes. The baseline and the ST segment features are also computed.

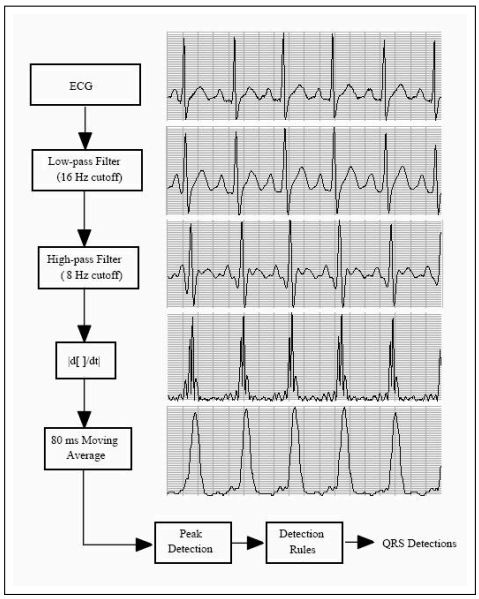


Figure 4: The ECG processing flowchart

The ECG preprocessing stage uses the raw signal to generate a windowed estimate of the energy in the QRS frequency band by using the following filters:

- Low pass filter;
- High pass filter;
- Taking the absolute value of the derivative;
- Averaging the absolute value over an 80 ms window.

The combined highpass, lowpass and derivative filters produces a bandpass filter with the bandwidth that contains most of the energy in the QRS complex. The theory and implementation of these filters are detailed in [15]. The averaging window was chosen to be the width of a typical QRS complex (80ms).

After the signal ECG filtering, the algorithm detects peaks in the signal. Each time a peak is detected it is classified as either a QRS complex or noise, or it is saved for later classification. The algorithm uses the peak height, peak location (relative to the last QRS peak), and maximum derivative to classify peaks.

The classification algorithm [16] uses the following rules:

- all peaks that precede or follow larger peaks by less than 200 ms are ignored;

• if a peak occurs, check to see whether the raw signal contained both positive and negative slopes. If not, the peak represents a baseline shift;

• if the peak occurred within 360 ms of a previous detection check to see if the maximum derivative in the raw signal was at least half the maximum derivative of the previous detection. If not, the peak is assumed to be a T-wave;

• if the peak is larger than the detection threshold call it a QRS complex, otherwise call it noise.

• if no QRS has been detected within 1.5 R-to-R intervals, there was a peak that was larger than half the detection threshold, and the peak followed the preceding detection by at least 360 ms, classify that peak as a QRS complex.

The detection threshold used in the last two rules is calculated using estimates of the QRS peak and noise peak heights. Every time a peak is classified as a QRS complex, it is added to a buffer containing the eight most recent QRS peaks. Every time a peak occurs that is not classified as a QRS complex, it is added to a buffer containing the eight most recent non-QRS peaks (noise peaks). The detection threshold is set between the average of the noise peak and QRS peak buffers according to the formula:

Det_Th = Avg_Noise_Peak + TH*(Avg_QRS_Peak - Avg_Noise_Peak),

where TH is the threshold coefficient. Similarly, the R-to-R interval estimate used in last rule is computed as the average of the last eight R-to-R intervals.

The Personal server receives the signal from the ECG module at 200Hz and computes the status of the patient for the following ECG parameters:

• Tachycardia if HR > 140bpm;
• Bradycardia if HR < 45 bpm;
• Asistola if HR = 0 bpm for at least 3 sec.;
• ST segment elevation if ST > 200l'V;
• ST segment depression if ST < - 150l'V;
• Wider QRS if QRS duration > 0,12 sec.

For the body temperature measurement we use the TMP275 temperature sensor (Texas Instruments). The TMP275 is a 0.5°C accurate, two-wire, serial output temperature sensor available in an SO8 package. The TMP275 is capable of reading temperatures with a resolution of 0.0625°C. The TMP275 is directly connected to the ez430-RF2500 using the I2C bus and requires no external components for operation except for pull-up resistors on SCL and SDA. The accuracy for the 35-45°C interval is below 0.2 °C and the conversion time for 12 data bits is 220 ms typical.

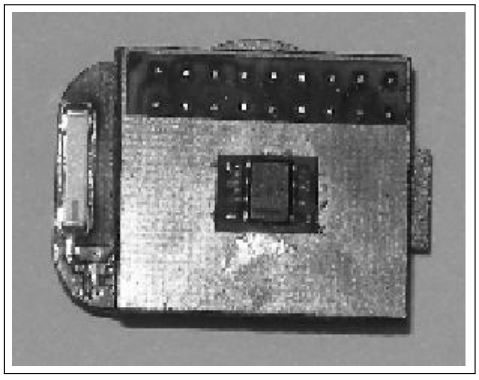


Figure 5: The thermometer module

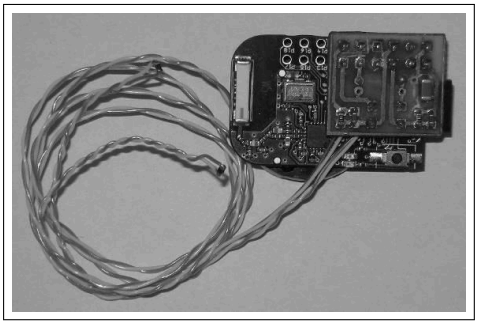


Figure 6: The respiration module

The Personal server samples the signal from the temperature sensor once per second and computes the status of the patient for the following temperature values:

- *Low temperature* - when temperature falls below 35°C;
- *High temperature* - when temperature rises above 38°C;
- *Normal temperature* - between the above values.

The *respiration module* (Figure 6) uses one of the most usual methods to sense breathing - to detect airflow using a nasal thermistor [9]. Although most applications require only breathing detection, some applications and diagnostic procedures require monitoring of the respiratory rhythm.

Our wireless respiration sensor uses a thermistor for long-time monitoring during the normal activity. The sensor is designed using MSP430F2274 microcontroller with an on-chip 10 bit A/D converter for data acquisition and CC2500 2.4GHz wireless transceiver. The thermistor detects changes of breath temperature between ambient temperature (inhalation) and lung temperature (exhalation). A thermistor placed in front of a nose detects breathing as a temperature change. The used thermistor is a 0603 SMD type and has the following characteristics: $\mathsf{R}_{\mathrm{nom}} = 10\mathrm{k}\ \Omega$ at 25°C, B = 3380, 1% tolerance.

The respiration signals are recorded using the MSP430F2274 A/D converter with 10 Hz sampling frequency.

The Personal server on patient computes the following respiration parameters:

- *Breathing amplitude* - calculated for every breathing cycle as a difference between minimum (Inhalation) and maximum thermistor voltage (Exhalation);
- *Breathing interval* - measured between two minimums representing two inhalations;
- *Breathing frequency* calculated from the breathing interval as a number of breaths per minute. Normal breathing frequency is 12-20 cycles/minute.

We consider two types of respiration:

- *Normal respiration*, when every breath lasts more than 0.5 seconds;
- *Apnoea*, when the breathing is missing for more than 10 seconds. Sleep apnoea can last more than 120 seconds.

The pulsoximeter sensor used is Micro Power Oximeter board from Smiths Medical [10] (Figure 7). The same sensor can be used for heart-rate detection and SpO_2. The probe is placed on a peripheral point of the body such as a finger tip, ear lobe or the nose. The probe includes two light emitting diodes (LEDs), one in the visible red spectrum (660 nm) and the other in the infrared spectrum (905 nm). The percentage of oxygen in the body is computed by measuring the intensity from each frequency of light after it transmits through the body and then calculating the ratio between these two intensities.

The pulsoximeter communicates with the eZ430-RF2500 through asynchronous serial channel at CMOS low level voltages. Data provided includes % SpO_2, pulse rate, signal strength, plethysmogram and status bits and is sent to the eZ430-RF2500 at a baud rate of 4800 bps, 8 bits, one stop bit and no parity.

The Micro Power Oximeter has the following measurement specifications: range 0-99% functional SpO2 (1% increments), accuracy ą2 at 70-99% SpO_2 (less than 70% is undefined), pulse range 30-254 BPM (1 BPM increments), accuracy ą2 BPM or ą2% (whichever is greater).

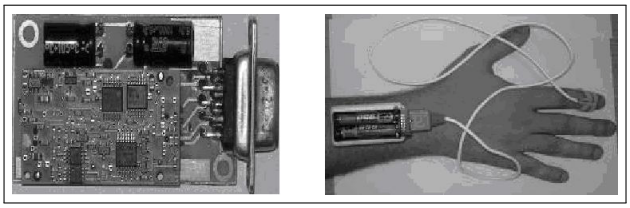


Figure 7: The pulsoximeter module

For the *blood pressure* measurement, a commercially available A&D UA-767PC BPM [11] was used. The blood pressure monitor (BPM) takes simultaneous blood pressure and pulse rate measurements. It includes a bi-directional serial port connection communication at 9600 kbps. An eZ430-RF2500 communicates with the BPM on this serial link to start the reading process and receives the patient's blood pressure and heart rate readings. Once the readings are received, the eZ430-RF2500 communicates with the network and transmits them to the Personal Server.

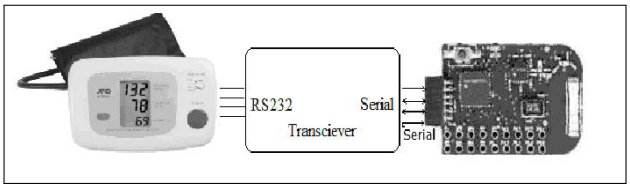


Figure 8: The blood pressure module (block diagram)

The Personal server computes blood pressure and defines the status of the patient by using the following blood pressure values:

- Hypotension: systolic $<$ 90 mmHg or diastolic $<$ 60 mmHg;
- Normal: systolic 90-119mmHg and diastolic 60-79 mmHg;
- Pre-hypertension: systolic 120-139 mmHg or 80-89 mmHg;
- Stage 1 Hypertension: systolic 140-159mmHg or diastolic 90 - 99 mmHg;
- Stage 2 Hypertension: systolic $\geqslant$ 160mmHg or diastolic $\geqslant$ 100 mmHg;

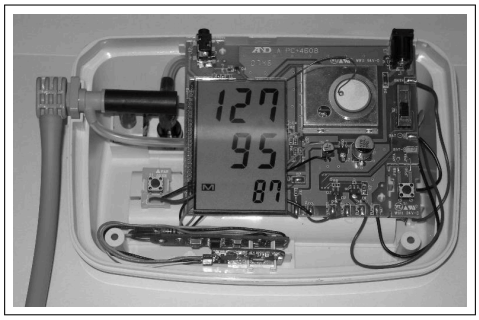


Figure 9: The blood pressure module

Our module for fall detection of humans is based on accelerometer technique. By using a tri-axial accelerometers our system can recognize patient movements. Linear acceleration are measured to determine whether motion transitions are intentional.

The algorithm for the human fall detection [3] uses the ADXL330 accelerometer and eZ430-RF2500 Wireless Module. The ADXL330 is a small, thin, low power, complete three axial accelerometer with signal conditioned voltage outputs, all on a single monolithic IC. The product measures acceleration with a minimum full-scale range of ą3g. It can measure the static acceleration of gravity in tilt-sensing applications, as well as dynamic acceleration resulting from motion, shock, or vibration.

The microcontroler calculates the aA acceleration using the formula:

$$a_A = \sqrt{a_{A_x}^2 + a_{A_y}^2 + a_{A_z}^2}$$

We determine if the subject has fallen if the condition $a_A > 0.4g$ is valid.

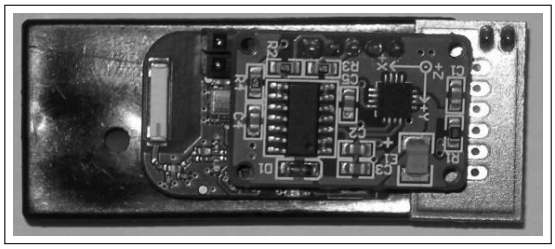


Figure 10: The fall detection module

3 Results

In the Figure 11 it is represented the personal server, that were implemented by means of a PDA (Fujitsu-Siemens Loox T830). This personal medical monitor is responsible for a number of tasks, providing a transparent interface to the wireless medical sensors, an interface to the patient, and an interface to the central server.

The USB interface (Figure 18) is realized by using a serial to USB transceiver (FT232BL) from FTDI [12] and enables eZ430-RF2500 to remotely send and receive data through USB connection using the MSP430 Application UART. All data bytes transmitted are handled by the FT232BL chip. It also contains a voltage regulator to provide 3.3 V to the eZ430-RF2500.

The software on the Personal Server [4], [5] receives real-time patient data from the sensors and processes them to detect anomalies.

The software working on the Personal Server (Figure 12) was written by using C# from Visual Studio.NET, version 8. The software displays temporal waveforms, computes and displays the vital parameters and the status of each sensor (the battery voltage and distance from the Personal Server).

The distance is represented in percent of 100 computed based on RSSI (received signal strength indication measured on the power present in a received radio signal).

If the patient has a medical record that has been previously entered, information from the medical record (limits above the alarm become active) is used in the alert detection algorithm.

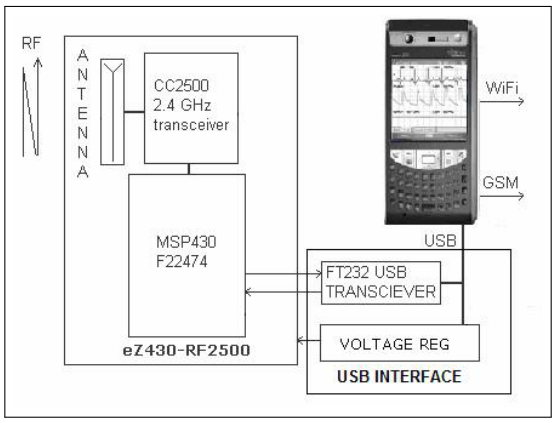


Figure 11: The Personal server (block diagram)

The following physiological conditions cause alerts:

- low SpO2, if SpO2 $< 90\%$;
- bradycardia, if HR < 40 bpm;
- tachycardia, if HR > 150bpm;
- HR change, if ?HR / 5 min $> 20\%$;

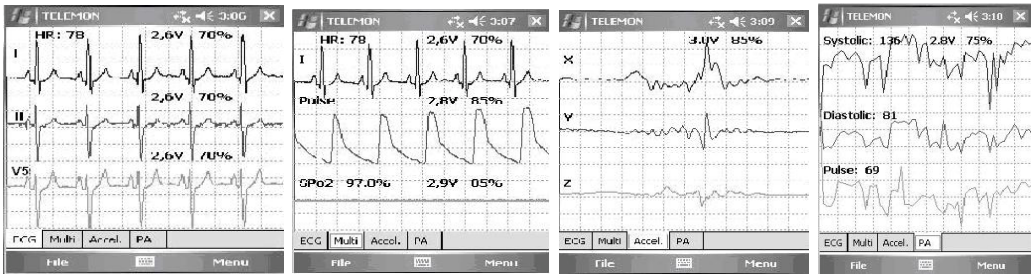


Figure 12: The Personal server interface: (a) 3 ECG traces, (b) one ECG trace, pulse waveform and SpO2, (c) 3 accelerometer traces, (d) systolic and diastolic pressure from BPM

- HR stability, if max HR variability from past 4 readings $> 10\%$;
- BP change if systolic or diastolic change is $> ą10\%$.

When an anomaly is detected in the patient vital signs, the Personal server software application generates an alert in the user interface and transmits the information to the TELEMON Server.

4 Summary and Conclusions

In this paper it is presented a project that aims to develop a secure multimedia, scalable system, designed for medical consultation and telemonitoring services. The main goal is to build a complete pilot system that will connect several local telecenters into a regional telemedicine network. This network enables the implementation of teleconsultation, telemonitoring, homecare, urgency medicine, etc. for a broader range of patients and medical professionals, mainly for family doctors and those people living in rural or isolated regions.

The Regional Telecenter in Iasi, situated in the Faculty of Medical Bioengineering, will allow local connection of hospitals, diagnostic and treatment centers, as well as a local network of family doctors, patients, paramedics and even educational entities. As communications infrastructure, we aim to develop a combined fix-mobile-internet (broadband) links.

The proposed system will also be used as a warning tool for monitoring during normal activity or physical exercise.

Such a regional telecenter will be a support for the developing of a regional medical database, that should serve for a complex range of teleservices such as teleradiology, telepathology, teleconsulting, telediagnosis, and telemonitoring. It should also be a center for continuous training and e-learning tasks, both for medical personal and for patients.

Acknowledgment
This work is supported by a grant from the Romanian Ministry of Education and Research, within PN_II programme (www.cnmp.ro/Parteneriate), contract No. 11-067/2007.

Bibliography

[1] Costin H, Rotariu C. (2004) Processing and Analysis of Digital Images. Applications in Biomedical Imagistics, Tehnica-Info Publ. House, Kishinev, Rep. Moldova, 441 pp., ISBN 9975-63-196-7

[2] European Commission, IST Directorate General (2006) Resource book of eHealth projects, Sixth Research and Development Framework Programme

[3] Giansanti D, et al, (2003) Is It Feasible to Reconstruct Body Segment 3-D Position and Orientation Using Accelerometric Data? IEEE Trans. On BME 50(4)

[4] Shorey R. (2006) Mobile, Wireless and Sensor Network: Technology, Applications and Future Directions, Ed. Wiley

[5] Wattenhofer R., (2005), Algorithms for ad hoc and sensor networks, Comp.Comm., 28, p.1498-1504

[6] Costin H, et al., (2006) A multimedia Telemonitoring Network for Healthcare. Enformatika, Transactions on Engineering, Computing and Technology, Vol. 17, Cairo, pp. 113-118

[7] Costin H, Morancea Octavia, et al. (2008) Integrated system for real time monitoring of patients and elderly people. Ukrainian Journal of Telemedicine and Medical Telematics, Vol. 6, No. 1, pp. 71-75

[8] Rotariu Cr., Costin H., Arotaritei D. and Constantinescu G. (2009) A Low Power Wireless Personal Area Network for Telemedicine, Proceedings of the 4th European Conference of the International Federation for Medical and Biological Engineering, Vol. 22, pp. 982-985

[9] Jovanov E., Raskovic D., Hormigo R., (2001) Thermistor-Based Breathing Sensor for Circadian Rhythm Evaluation, Biomedical Sciences Instrumentation, Vol. 37, pp. 493-497

[10] http://www.smiths-medical.com/Userfiles/oem/OEM.31392B1.pdf

[11] http://www.lifesourceonline.com/and_med.nsf/html/UA-767PC

[12] FT232 datasheet at http://www.ftdichip.com/FT232

[13] J. Pan and W.J. Tompkins, (1985), A real-time QRS detection algorithm, IEEE Trans. Biomed. Eng., vol. BME-32, pp. 230-236

[14] P.S. Hamilton and W.J. Tompkins, Quantitative investigation of QRS detection rules using the MIT/BIH arrhythmia database, IEEE Trans. Biomed Eng., vol. BME-33, pp. 1157-1165

[15] Tompkins WJ (ed.). (1993) Biomedical Digital Signal Processing: C-Language Examples and Laboratory Experiments for the IBM PC., Englewood Cliffs, NJ: PTR Prentice Hall

[16] P.S. Hamilton, (2002), Open Source ECG Analysis Software, E. P. Limited, Somerville, Mass, USA.

[17] MSP430 datasheet at http://www.ti.com/MSP430

[18] CC2500 datasheet at http://www.ti.com/CC2500

Int. J. of Computers, Communications & Control, ISSN 1841-9836, E-ISSN 1841-9844
Vol. V (2010), No. 5, pp. 910-918

A New Model for Cluster Communications Optimization

A. Rusan, C.-M. Amarandei

Andrei Rusan, Cristian-Mihai Amarandei
"Gheorghe Asachi" Technical University of Iaşi
Department of Computer Science and Engineering,
Adress: Bd. Dimitrie Mangeron, Nr. 53A, 700050, Iaşi, Romania
E-mail: andrei.rusan@tuiasi.ro, camarand@cs.tuiasi.ro

Abstract: Performance losses of cluster applications can arise from various sources in the communications network of computer clusters. Typically, CPU intensive applications generate a small amount of network traffic the overall influence of the network subsystem is minimal. On the other hand, a data-intensive and network aware application generates a large amount of network traffic and the influence of the network subsystem is significantly greater. This paper presents a model that aims to improve the cluster's network performance by reducing the data transfer time, this solution having the advantage that doesn't imply modifications of the original applications or of the kernel.
Keywords: cluster communications optimization, network performance

1 Introduction

The computing performance of a cluster is dependent on the performance of cluster components: computing nodes and communication infrastructure. The cluster communication infrastructure was built using network devices whose performance can be modified only by hardware changes, i.e. replacing 100Mbps Ethernet switches with gigabit switches; therefore this component of the cluster will be neglected. The performance of a computing node is defined by hardware and software performance, where the hardware performance can be considered a constant value and it can be influenced by the hardware changes only. The software can be separated in components: application and operating system. The computing performance of the cluster can be influenced by each of these components.

Cluster performance increasing by performing modifications at the applications level is a goal very hard to achieve, some applications requiring a complete rewrite for this task. Of course, there are exceptions too, like network-aware applications, which are built exactly with this goal in mind and they do not require any optimizations.

The last component that can influence the performance is the operating system through the kernel configuration and at the network layer optimizations.

Taking into account that clusters typical contains a large number of computing nodes, the solution for cluster performance improvements must be done with minimal modifications in the systems. These requirements are necessary to keep the administrative tasks at a decent level.

There is a research work involving the network tuning mechanisms or network-aware applications trying to solve these issues. The projects developed so far, like WAD (Work Around Daemon) [1] or ENABLE [2] don't meet the imposed requirements, as for WAD a modified kernel must be used, and for ENABLE the applications must be rewritten.

The WAD project provides a transparent mechanism to work around a variety of network issues, including TCP buffer size, MTU size, packet reordering, and leaky network loss [1]. The WAD goal is to eliminate the "wizard gap" representing the difference between the network performances achievable by manually handcrafting of the optimal tuning parameters, compared

to an untuned application [1]. To attain this goal, WAD requires a modified kernel from the Web100 project. This solution could not be applied in our case, because of the different Linux kernel versions the Web100 project provides a kernel patch starting from the 2.6.12. Through the use of a different kernel version that the provided one by the Linux distribution, CentOS 4.5 in our case, problems in maintaining the operating system across clusters can occur. Therefore our solution works fine no mater the kernel version used.

The ENABLE project, includes monitoring tools, visualization tools, archival tools, problem detection tools, and monitoring data summary and retrieval tools. ENABLE provides an API that makes it very easy for application or middleware developers to determine the optimal network parameters [2]. However, the solution provided by this project was not applicable in our case, because applications could not be rewritten.

This paper presents a model for self optimization of the network communications in order to improve cluster performance by shortening the data transfer time. The model implementation does not require applications and kernel structure modifications or adding new modules to the existing ones. Also, the implementation uses only the tools provided by the operating system for runtime configuration and therefore the automatic operating system and kernel updates can be applied immediately.

In the next section a brief review of the TCP transport protocol issues, Linux kernel network subsystem and the NetPIPE network performance measurement tool are presented. Section 3 describes the network self optimization model and the proposed algorithm. Section 4 presents the test environment and the experimental results. The final section summarizes author's efforts on tuning the cluster communications network and considers future extensions of the work.

2 Background

TCP protocol transmits new data into the network when old data has been received as indicated by acknowledgments from the receiver to the sender. The data rate is determined by the window size and is limited by the application, the buffer space at the sender or the receiver and by the congestion window. TCP adjust the congestion window to find an appropriate share of the network capacity of the path between source and destination. Missing or corrupted data segments are repaired by TCP by retransmitting the data from the sender's buffer. This process requires an entire window of data to fit into both sender and receiver buffers [1].

The largest TCP window can be 216=65KB because the TCP header uses 16 bits to report the receive window size to the sender. The window scale option was introduced defining an implicit scale factor used to multiply the windows size value from TCP header in order to obtain the real TCP window size, as described in [3]. These buffers have default values that may either be changed by the applications using system calls or by using tools provided by the operating system, i.e. `sysctl` tool from Linux/Unix.

The second part of this section is focused on the network subsystem of the Linux kernel. Starting from the version 2.4 of Linux kernel, an auto tuning technique is used to perform memory management. This technique simply increases and decreases buffer sizes depending on available system memory and available socket buffer space. By increasing buffer sizes when they are full of data, TCP connections can increase their window size performance improvements are an intentional side-effect [4]. On the other hand, this is done within the limitation of the available system memory and socket buffer space, and on the busy cluster that are valuable resource.

The network subsystem of the Linux operating system should be tuned in order to obtain an optimal performance of a computing system. In order to do that, changes can be operated at the following levels: network interface and kernel parameters. The kernel parameters allowing to change the network behavior can be tuned by modifying the following files located in

/proc/sys/net:

```
/proc/sys/net/core/rmem_max
/proc/sys/net/core/rmem_default
/proc/sys/net/core/wmem_max
/proc/sys/net/core/wmem_default
/proc/sys/net/ipv4/tcp_stack
/proc/sys/net/ipv4/tcp_timestamps
/proc/sys/net/ipv4/tcp_keepalive_time
/proc/sys/net/ipv4/tcp_mem
/proc/sys/net/ipv4/tcp_rmem
/proc/sys/net/ipv4/tcp_wmem
/proc/sys/net/ipv4/tcp_window_scaling
```

The network interface can also be tuned by modifying the speed and duplex settings and the MTU size. Two problems have to be addressed while setting up the cluster:

- the default kernel values don't provide the best performance for the custom environment, and
- the number of communication devices needed to be set.

Since solutions to solve these two problems are missing an optimal value for each system in cluster to get the best possible performance is proposed. By using the right tools, the network settings related changes are available immediately, the optimization algorithm presented in this paper being based on these features. The values of the send/receive buffers (tcp_wmem and tcp_rmem) can be changed by specifying minimum size, initial size, and maximum size as follows:

```
sysctl -w net.ipv4.tcp_rmem="4096 87380 8388608"
sysctl -w net.ipv4.tcp_wmem="4096 87380 8388608"
```

The third value must be the same as or less than the wmem_max and rmem_max values. The first value can be increased on high-speed, high-quality networks so that the TCP window starts out at a sufficiently high value [5]. Also, the TCP window scaling is an option to enlarge the transfer window.

Performance measurements of the cluster network can be done using a wide area of tools like Iperf [6], Netperf [7] or NetPIPE (Network Protocol Independent Performance Evaluator). Because the performance measurements must be done for both TCP and MPI layer and the NetPIPE provides a complete measurement of the communication performance on both of them, the tests were performed using this tool. The NetPIPE utility performs simple ping pong tests, bouncing messages of increasing size between two computers. To provide a complete test, NetPIPE modifies the message size, with a slight perturbation, at regular intervals and measures the point-to-point communications performance between nodes [8]. Because we want to determine of the maximum bandwidth available for different use cases, usage of different message size is a must. From all the performance measurements tools available, only NetPIPE had this feature by default, which defines it as a right tool for this kind of tests. An in depth description of the NetPIPE utility can be found in [8]- [10]. Linux kernel network subsystem information gathered by running NetPIPE on the cluster were used in order to improve the throughput.

3 Proposed model

The proposed model implies the computation of a best possible set of values for a given set of parameters. Figure 1 presents the model schematics composed from three parts: "Control logic",

"Parameter computation" and "Network test tool". The first component is responsible for sending starting set of values to "Parameter computation" and keeps the running flux under control. The second component computes the sets of values based on previously known sets of data and on the results of the current run received from the "Network test tool" and send it to the kernel in order to set the network subsystem. The "Network test tool" is responsible to run set of tests and provide the results to "Parameter computation" component. The optimization process is started by "Control logic", which sends starting values to "Parameter computation" that are set in the kernel network subsystem and starts the first set of tests through the "Network test tool". After the tests are finished, the results are sent to "Parameter computation" which computes a new set of values and the process continues until meeting the end condition.

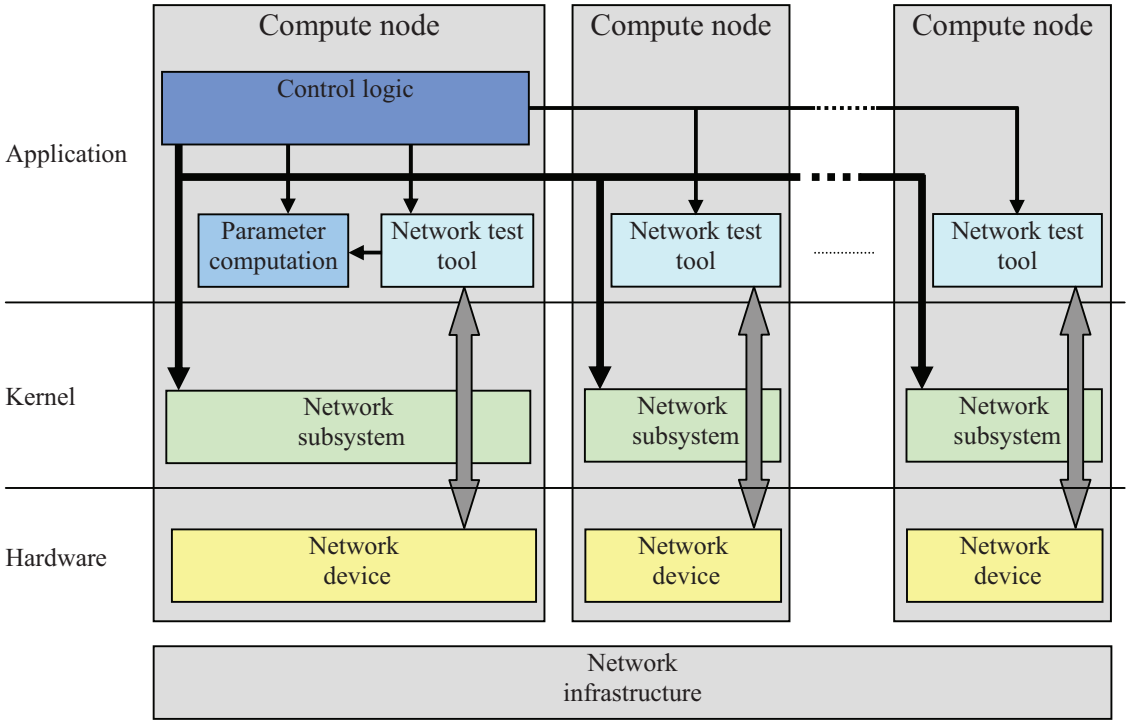


Figure 1: The cluster communication optimization model

The process provides self optimization of the kernel network subsystem, the only necessary interaction with the application being the configuration file that contains necessary values for the application startup and for the components behavior. These values can be set by administrators to meet the specific needs.

An algorithm implementing model functionality is proposed. This algorithm performs the bandwidth measurements and adjusts the sets of parameters to obtain the highest bandwidth usage for each case and running tests.

Given l, the number of tests to be performed, let it be $N = \{n_1, n_2, \ldots, n_k\}$ the set of nodes in cluster, $T = \{t_1, t_1, \ldots, t_l\}$ the set of test variables (i.e. `tcp_rmem`, `tcp_wmem`, `tcp_window_scalling`), $I = \{i_1, i_2, \ldots, i_l\}$ the kernel network subsystem parameters start values and $E = \{e_1, e_1, \ldots, e_l\}$ be the set of computed values for each test t_i. Also, given m as the number of messages, $MS = \{ms_1, ms_2, \ldots, ms_m\}$ is defined as the set of message sizes used by the testing tool, $X = \{x_1, x_2, \ldots, x_m\}$ as the best set of result value for each test t_t, the results set $R_i = \{r_1, r_2, \ldots, r_k\}_i$ one for each cluster node, $S = \{s_1, s_2, \ldots, s_m\}$, where $s_i = \sum_{j=i}^{k} r_{ij}$, and $B = \{b_1, b_2, \ldots, b_m\}$ as the set of best values for each test t_i. The algorithm computes the values for the test variables

as follows:

```
1  foreach t_i in T do
2      iter=i_i;
3      while iter<e_i do
4          param = generate_set(t_i, iter);
5          set_kernel_parameters(param);
6          foreach n_i in N do
7              start_remote_testing_program(n_i);
8              prepare_local_testing_program(n_i);
9          end
10         parallel do
11             R_i = execution_of_local_testing_programs;
12         end
13         foreach ms_i in MS do
14             foreach n_j in N do
15                 s_i[iter] += r_ij;
16             end
17             iter = get_next_iter(t_i,i_i,e_i,iter);
18         end
19     end
20     foreach ms_i in MS do
21         iter=i_i;
22         while iter<e_i do
23             if x_i<s_i[iter] then
24                 x_i = s_i[iter];
25                 b_i = iter;
26             end
27             iter = get_next_iter(t_i,i_i,e_i,iter);
28         end
29     end
30     best = get_max_count(B);
31     set_kernel_parameters(t_i,best);
32 end
```

The algorithm has two main components: the network test component corresponding to "Network test tool" in Figure 1 and implemented by lines 3-19 from the algorithm, and the computational component corresponding to "Parameter computation" in Figure 1 and implemented by lines 20-30. The methods used in the algorithm implements the following actions:

- generate_set: produce a new set of parameters used for network testing;
- start_remote_testing_program: launches the remote component of the testing application (NetPIPE in our case);
- prepare_local_testing_program: prepare the local component of the testing application, necessary to maximize the accuracy of the measured values;
- execution_of_local_testing_programs: runs the testing application;
- get_max_count: extract the parameter value corresponding to the maximal throughput.

Finally, the line 31 sequentially sets the kernel parameters to the best computed values on the entire cluster.

4 Implementation and experimental results

To improve the cluster communications, dynamic tests and adjustments for the following Linux kernel network parameters are performed: tcp_window_scalling, tcp_rmem and tcp_wmem. Bandwidth measurements and TCP parameters adjustments were carried out to obtain the highest bandwidth usage for each case and to determine the maximum bandwidth available for different use cases.

The environment used to test the proposed model consists in a grid cluster with the following configuration: one front-end computer with 4 x 3.66 GHz Intel Xeon processors, 4 x 146 GB hard drive and 8 GB of RAM and 12 computing nodes with 1 x 2.33GHz Intel Core2 Duo CPU, 1 x 160GB hard drive and 2GB of RAM with Gigabit Ethernet card connected with CAT6 cables via a Gigabit switch.

First implementation of the algorithm was made in Bash, but due the dificulties in working with data structures was switched to Perl. Also, to preserve measurement accuracy and performance the tests results were saved to files for further usage, like graphical presentation.

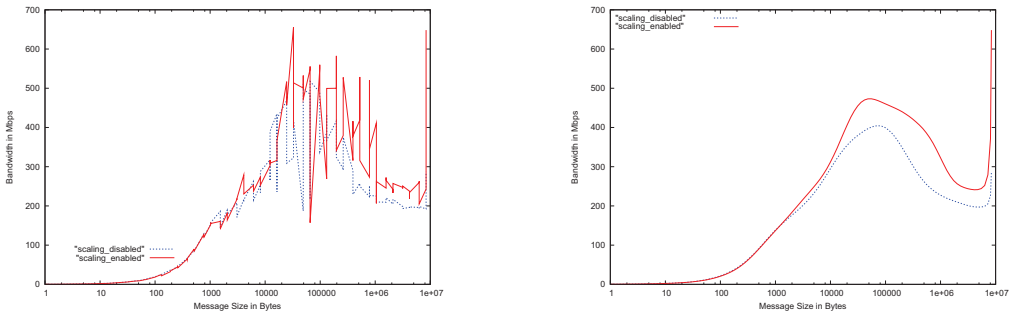
The performance results were obtained based on a test array with three elements: one for

TCP windows scaling, a second one for TCP read buffer and the third one for the TCP write buffer kernel parameters. For each of this there is a graphical presentation, where on the X and Y-axis the message size used during tests and the resulted bandwidth values are, respectively, presented.

The results for the TCP windows scaling parameter are showed in Figure 2(a), where one line is for net.ipv4.tcp_window_scaling=0, and the other one is for net.ipv4.tcp_window_scaling=1. For an easier reading of the results graph, we apply a Bezier function in order to obtain the presentation from Figure 2(b).

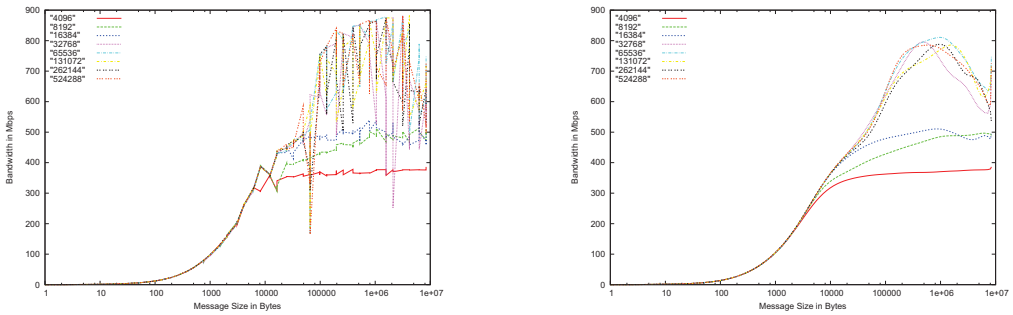
The influence of TCP read/write buffer size over the available bandwidth are presented in Figure 3(a)/Figure4(a). In this case, there are a large number of graphic representations and the observation is very difficult, so a Bezier function was applied on the results values to make the graphical presentation more readable, as shown in Figure3(b)/Figure4(b). The red line in the graphical representations corresponds to the default value for tcp_rmem (4KB) and the blue dotted line is the best value resulted using this model, with 100Mbps more than the default value.

By applying different values for TCP buffers and running tests bandwidth variation for each of them is presented (Figure 5). TCP buffers size starts from 4KB and each algorithm step doubles the previous value.



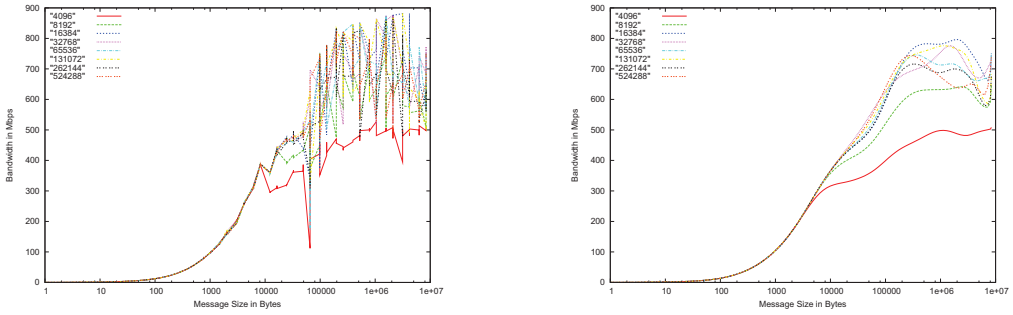
(a) (b)

Figure 2: TCP window scaling influence over bandwidth



(a) (b)

Figure 3: TCP read buffer influence over bandwidth



(a) (b)

Figure 4: TCP write buffer influence over bandwidth

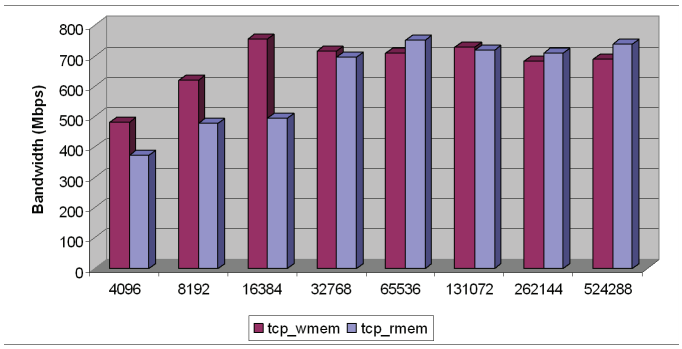
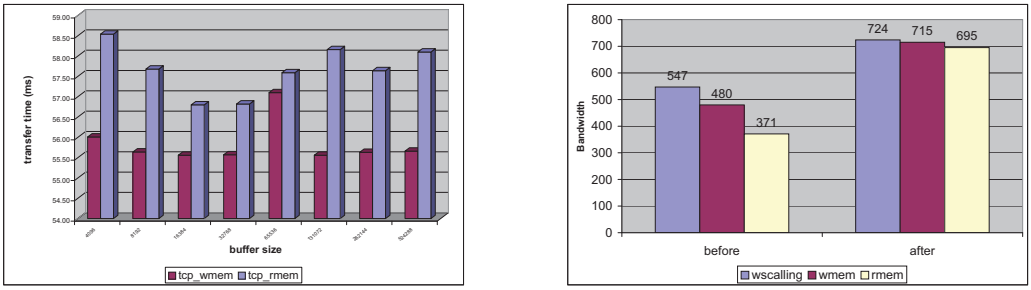


Figure 5: Bandwidth achieved for different TCP buffer size

A ram drive on all computers was built in order to test the results without any delays introduced by the hard disk drive. Because of the 2GB memory limits on the computing nodes, the file transferred and the ram drive size was 512 MB. In the Figure 6(a) transfer time for the file is shown. In this test the TCP buffers size was changed from 4KB to 512 KB and data transfer starts in both directions for each value, from the frontend to cluster nodes and back.



(a) (b)

Figure 6: (a)Transfer time for a 512MB file between cluster; (b)The benefits of parameters adjustments

The best transfer time was obtained when both tcp_rmem and tcp_wmem values were 16KB

or 32KB. During the tests, this solution provides all values for the considered TCP parameters, which can be useful for other scenarios. In Figure 6(b), the bandwidth improvement is presented.

Using only the default values, the cluster internal network available bandwidth is not optimally used with a strong impact on the overall computing performance. Using this optimization model, the bandwidth available in the cluster is efficiently used.

5 Conclusions and future work

Using the proposed model, the communication between cluster nodes has been improved. All results are considered for the specific needs of mentioned cluster, where a significant amount of data needs to be transferred between cluster nodes. For other applications, like a web server farm, the final results may be slightly different, but can be optimized by adjusting the test tool for those specific needs. The network kernel parameters computed can be used later if the use case is changed, i.e. a web server farm. The algorithm can be used for IPv6 too, however the authors doesn't implemented nor tested.

The further development of presented application will follow two directions: one is to extend its capabilities to support UDP traffic performance adjustment; and the second one is to support tuning parameters other than the ones related to the Linux kernel.

Bibliography

[1] T. Dunigan, M. Mathis, B. Tierney , A TCP tuning daemon, *Conference on High Performance Networking and Computing, Proceedings of the 2002 ACM/IEEE conference on Supercomputing*, Baltimore, Maryland, 2002

[2] B.L. Tierney, TCP tuning guide for distributed application on wide area networks, *Usenix*;login. `http://www-didc.lbl.gov/tcp-wan-perf.pdf`., 2001

[3] V. Jacobson, R. Braden, D. Borman, RFC1323 TCP Extensions for High Performance, May 1992

[4] B.L Tierney, D. Gunter, J. Lee, M. Stoufer, J.B. Evans, Enabling network-aware applications, Proceedings of the 10th IEEE International Symposium on High Performance Distributed Computing, Page: 281-288, 2001, ISBN: 0-7695-1296-8

[5] NetPIPE, webpage: `http://www.scl.ameslab.gov/netpipe/`

[6] A. Tirumala, L. Cottrell, "Iperf Quick Mode", `http://www-iepm.slac.stanford.edu/bw/iperfres.html`

[7] Netperf homepage, `http://www.netperf.org/netperf/NetperfPage.html`

[8] D. Turner, A. Oline, X. Chen, and T. Benjegerdes,Integrating New Capabilities into NetPIPE, *Lecture Notes in Computer Science*, Springer-Verlag, September 2003, pp. 37-44.

[9] D. Turner, X. Chen, Protocol-Dependent Message-Passing Performance on Linux Clusters, *Proceedings of the IEEE International Conference on Cluster Computing*, September 2002, pp. 187-194.

[10] Q.O. Snell, A. Mikler, J.L. Gustafson, NetPIPE: A Network Protocol Independent Performance Evaluator, *ASTED International Conference on Intelligent Information Management and Systems*, June 1996.

[11] H. Sivakumar, S. Bailey, R. L. Grossman,PSockets: The Case for Application-level Network Striping for Data Intensive Applications using High Speed Wide Area Networks, *Proceedings of IEEE Supercomputing 2000*, Nov., 2000, `http://www.ncdm.uic.edu/html/psockets.html`

[12] J. Postel, *RFC793 Transmission Control Protocol*, September 1981

[13] R. Braden , *RFC1122 Requirements for Internet Hosts – Communication Layers*, October 1989

[14] V. Paxson, G. Almes, J. Mahdavi, M. Mathis, *RFC 2330 Framework for IP Performance Metrics*, May 1998

[15] E. Ciliendo, T. Kunimasa, B. Braswell, *Linux Performance and Tuning Guidelines*, IBM, July 2007.

Int. J. of Computers, Communications & Control, ISSN 1841-9836, E-ISSN 1841-9844
Vol. V (2010), No. 5, pp. 919-928

A Metrics-based Diagnosis Tool for Enhancing Innovation Capabilities in SMEs

J. Sepulveda, J. Gonzalez, M. Camargo, M. Alfaro

Juan Sepulveda, Javier Gonzalez, Miguel Alfaro
Department of Industrial Engineering, University of Santiago of Chile
3769 Ecuador Ave. Santiago, Chile. PO Box 10233

Mauricio Camargo
Nancy-Universite / ERPI (Equipe de Recherche des Processus Innovatifs)
8, rue Bastien Lepage 54010 Nancy Cedex, France

Abstract: Innovation doubtless represents a main strategic lever for the development of small and medium enterprises (SMEs) in many industrial sectors and it comprises new techniques, new products and new processes, as well as new services which lead to better customer service and revenue. However, the basic question of how well the company is equipped with the necessary practices, methodologies, people, and beliefs, is far from being completely answered yet. In this paper, a metrics-based diagnosis tool for measuring and enhancing the innovation capabilities in SMEs is presented along with a set of preliminary results from case-based studies at the local industry. In this paper we propose a new method by studying the competences of SMEs in concepts tied to innovation and by using a specified framework. As a first step, all of the necessary information by using questionnaires with verbal-scales evaluations is compiled. Second, we use a non-compensatory flow-based sorting method with central profiles to identify the current level and to classify the company into predefined levels. Third, a detailed analysis of the obtained values is performed in order to make a personalized recommendation.
Keywords: technological innovation, multicriteria decision making, classification methods.

1 Introduction

1.1 Innovation process

The innovation literature is a fragmented corpus due to the contribution of many scholars with diverse disciplinary backgrounds that try to adopt different ontological and epistemological positions to investigate and analyze this complex and multimensional phenomena. Thus, a variety of approaches [1], [2] and many different measurement methods [3], [4], [5] can be found. Chiesa *et al.* in [6] describe process and performance as the two foci of innovation management measures; they overlay *core processes* with a set of *enabling processes*, the latter describing the deployment of resources, and the effective use of appropriate systems and tools governed by top management leadership and direction. A close link exists between product and process innovations: the majority of the articles address both type of innovations and only few articles considered only process innovations [7]. We can observe this relationship in the definition made by Cormican *et al.* [8] , that describe the product innovation as a continuous and cross-functional process involving and integrating a growing number of different competences inside the organization. In the area of the variables related to innovation, we can consider important works on frameworks. Adams *et al.* [3] in a survey develops a synthesized framework of the innovation management

process consisting of seven categories: *inputs management, knowledge management, innovation strategy, organizational culture and structure, portfolio management, project management and commercialization.* Boly [9] , based on the literature [7] [8], identified the most used practices by innovative enterprises and he classified them under 13 categories or groups. According to the author, these practices constitute the principal actions performed by the enterprises to define their strategy, to guide and impel the innovation processes and make evolve the organization or its methods of work; they develop these practices completely or partially and in a formal or informal way where the level of use of these practices allows to classify the enterprises according to his innovation potential.

1.2 Measuring innovation by assessment of practices

Corona in [10] defined an index of potential innovation (IIP), which is calculated by using multicriteria decision making (MCDM) tools, and uses as criteria the 13 innovation practices defined by Boly [9] . These practices are the concrete actions executed by the enterprises to define their strategy, to guide and to impel the innovation processes and to make evolve the organization or its working methods. The index will allow to obtain a classification according to the attitudes and strategies adopted by these enterprises. Based on [11] we can classify companies as: *Proactive, Preactive, Reactive and Passive.* On the other hand, Morel et al. [12] propose the use of Choquet's Integral to consider the interaction between the different innovation practices, defining an Aggregated Index of potential innovation (APII). Finally, Assielou in [13], besides of adding two new innovation practices, he makes modifications in the system of treatment and practices.

Current index-based methods are actually sorting procedures and present limitations to correctly classify enterprises in this field. In this paper we propose a new method to make the evaluation of innovation levels in the small and medium enterprises by using a specified framework. As a first step, all the necessary information by using questionnaires with verbal-scales evaluations is collected. Then, a non compensatory flow-based sorting method with limiting profiles is applied. The third step consists of a detailed analysis of the values obtained in each evaluation of an enterprise to perform a personalized recommendation for each company about areas to be improved. An example of its application is given.

2 Construction of the Gathering Tool and Reference Profiles

2.1 Gathering Tool

Based on the works by Assielou [13], Corona [10] and Camargo [14] we use categories of innovation practices c_i, each with practices belonging to a similar class, as shown below.

I. Creation / Concept Generation (c_1)
1.1 Use of tools to increase the creativity
1.2 Integration of the clients and suppliers in the conception process
1.3 Organization, compilation and management of information from the exterior
II. Conception Activities (c_2)
2.1 Use of tools of help to the conception
2.2 Existence of a methodology of help to the conception
2.3 Hardware Equipment
III. Human Resources Management (c_3)

3.1 Management of competences and the skills of the society
3.2 Innovation stimulation
IV. Strategy (c_4)
4.1 Strategy integrated to favor the innovation
4.2 Network operation
4.3 Client Importance
4.4 Financing
V. Project management (c_5)
5.1 Project administration
5.2 Management of project briefcase
5.3 Organization of tasks tied to the Innovation
VI. Capitalization of Ideas and Concepts (c_6)
6.1 Continuous Improvement of the innovation process
6.2 Politics of Management of the intellectual property
6.3 Knowledge Capitalization

In order to measure every concept on each category, verbal scales mapped onto five numerical levels are defined with values $\{0; 0.25; 0.5; 0.75; 1\}$, where 0 and 1 are the lowest and the highest level , respectively. Each level was defined in detail for each concept, in order to avoid ambiguity and make the evaluation easier to the interviewer; a survey with 18 evaluations grouped into six categories is then applied.

2.2 Reference Profiles

By observing the classification established by Godet [11] and the six categories above, it is possible to establish intervals in which we can classify the enterprises; *Passive* enterprises have the lowest values in every category whereas the *Preactive* ones obtain the highest values. The division by interval in each category allows us to establish segments of values for each innovation profile in the characteristics to be measured by having into consideration that the values included in the interval belongs to the levels expected for a company of that profile. This can give us the idea that it is possible to determine reference profiles, where for example a *Proactive* company will have all its evaluation values in the highest intervals. Thus, each of the four divisions represents to a Reference company with *Passive, Reactive, Preactive* and *Proactive* characteristics (Table 1). Is it possible to see that these Reference profiles represent companies that have homogeneous development levels in each characteristic, which is not always the case; for instance, there exist companies with high levels of development in *Innovation Stimulation*, but at the same time poor levels in *Knowledge Capitalization*, or a company obtains *Proactive* values in *Human Resources Management* and *Reactive* values in *Capitalization of Ideas and Concepts*. We cannot hope that the majority of the companies will be homogeneous, therefore it is necessary to find out a method that allows us to establish a correct classification within the four profiles using a non compensatory mathematical tool.

3 FlowSort Method

Based on the ranking methodology of PROMETHEE, a new sorting method developed by Nemery and Lamboray [15] is proposed for assigning actions to completely ordered categories;

Category	Passive	Reactive	Preactive	Proactive
I. Creation / Concept Generation	$[0 - a_2[$	$[a_2 - a_3[$	$[a_3 - a_4[$	$[a_4 - 1]$
II. Conception Activities	$[0 - b_2[$	$[b_2 - b_3[$	$[b_3 - b_4[$	$[b_4 - 1]$
III. HR Management	$[0 - c_2[$	$[c_2 - c_3[$	$[c_3 - c_4[$	$[c_4 - 1]$
IV. Strategy	$[0 - d_2[$	$[d_2 - d_3[$	$[d_3 - d_4[$	$[d_4 - 1]$
V. Project management	$[0 - e_2[$	$[e_2 - e_3[$	$[e_3 - e_4[$	$[e_4 - 1]$
VI. Capitalization of Ideas	$[0 - f_2[$	$[f_2 - f_3[$	$[f_3 - f_4[$	$[f_4 - 1]$

Table 1: Interval Distribution of Category values

these categories are defined either by limiting profiles or by central profiles (also named centroids). The assignment of an action into a Category is based on the relative position of this action with respect to the defined reference profiles in terms or incoming or outgoing net flows. We denote by $A : (a_1, ..., a_n)$ the set of n actions to be sorted. These actions are evaluated on q criteria $g_j (j = 1, ..., q)$ that have to be maximized. We denote the categories to which the actions must be assigned by $C_1, C_2, ..., C_k$. These categories are either delimited by two boundaries, in the case of limiting profiles, or by centroids in the case of central profiles. This Categories are ordered as $C_1 > ...C_l > C_k$, where $C_h > C_k$, with $h < l$, which denotes that C_h is preferred to category C_l. We denote $R = (r_1..., r_{k+1})$ as the set of limiting profiles in the case when a category is defined by an upper and lower profile, represented as r_{h+1} . On the other hand, when we define a category by one central profile, the centroid is denoted by $\tilde{R} = (\tilde{r}_1, ..., \tilde{r}_k)$, where $\tilde{r}_j$ is the centroid of category C_j. We also define $\pi(x, y)$ as the preference of action x over an action y, which is used in the same way as in PROMETHEE. Thus, on the basis of these preference degree, positive, negative and net flows of each action x of R_i, are computed by equations (3.1),(3.2),(3.3), where $R_i = R \bigcup \{a_i\}$.

$$\phi^+_{\dot{R}_i} = \frac{1}{|\dot{R}_i| - 1} \sum_{y \in \dot{R}_i} \pi(x, y) \tag{3.1}$$

$$\phi^-_{\dot{R}_i} = \frac{1}{|\dot{R}_i| - 1} \sum_{y \in \dot{R}_i} \pi(y, x) \tag{3.2}$$

$$\phi_{\dot{R}_i} = \phi^+_{\dot{R}_i} - \phi^-_{\dot{R}_i} \tag{3.3}$$

In this case we use $\dot{R}_i$ when no difference can be made between a set of limiting profiles and a set of centroids.

The Flow-Based Assignment rules differ in the use of limiting profiles and central profiles. In the case of limiting profiles, the rules of the positive and negative flow assignment are defined as follows:

$$C_{\phi^+}(a_i) = C_h, \text{ if } \phi^+_{R_i}(r_h) \geq \phi^+_{R_i}(a_1) > \phi^+_{R_i}(r_{h+1}) \tag{3.4}$$

$$C_{\phi^-}(a_i) = C_h, \text{ if } \phi^-_{R_i}(r_h) < \phi^-_{R_i}(a_1) \leq \phi^-_{R_i}(r_{h+1}) \tag{3.5}$$

If we want to strictly impose the assignment to one category, using the net flow we can define the assignment rule by (3.6).

$$C_{\phi}(a_i) = C_h, \text{ if } \phi_{R_i}(r_h) \geq \phi_{R_i}(a_1) > \phi_{R_i}(r_{h+1}) \tag{3.6}$$

In the case of central profiles, the Flow-Based Assignment rules of positive and negative flows are defined by (3.7) and (3.8).

$$\tilde{C}_{\phi^+}(a_i) = C_h, \text{ if } \frac{\phi^+_{\tilde{R}_i}(\tilde{r_h}) + \phi^+_{\tilde{R}_i}(\tilde{r}_{h+1})}{2} < \phi^+_{\tilde{R}_i}(a_1) \leq \frac{\phi^+_{\tilde{R}_i}(\tilde{r_h}) + \phi^+_{\tilde{R}_i}(\tilde{r}_{h-1})}{2} \tag{3.7}$$

$$\tilde{C}_{\phi^-}(a_i) = C_h, \text{ if } \frac{\phi^-_{\tilde{R}_i}(\tilde{r_h}) + \phi^-_{\tilde{R}_i}(\tilde{r}_{h+1})}{2} \geq \phi^-_{\tilde{R}_i}(a_1) > \frac{\phi^-_{\tilde{R}_i}(\tilde{r_h}) + \phi^-_{\tilde{R}_i}(\tilde{r}_{h-1})}{2} \tag{3.8}$$

Here, we can also strictly impose the assignment to one category using the net flows with the assignment rule (3.9).

$$\tilde{C}_{\phi}(a_i) = C_h, \text{ if } \frac{\phi_{\tilde{R}_i}(\tilde{r_h}) + \phi_{\tilde{R}_i}(\tilde{r}_{h+1})}{2} < \phi_{\tilde{R}_i}(a_1) \leq \frac{\phi_{\tilde{R}_i}(\tilde{r_h}) + \phi_{\tilde{R}_i}(\tilde{r}_{h-1})}{2} \tag{3.9}$$

4 Application

In this section the method is explained in four main steps, by using information obtained from seven SMEs, E_j, with $(j = 1, ..., 7)$ taken from the metalworking industry located at Santiago of Chile, as follows.

E_1: appliances manufacturing such as refrigerators, gas and kerosene heaters.
E_2: appliances manufacturing such as gas and electric stoves and heaters.
E_3: appliances manufacturing such as home boilers and sinks.
E_4: faucets and gas valves manufacturing.
E_5: vehicle transforming maker such as for ambulances and safety vehicles.
E_6: safety deposit box manufacturing with electronic controls.
E_7: vending machine refurbishment and adapting for industrial utilization.

4.1 First Step: Determination of the weights and references profiles.

By observing the weight used in each one of the practices in [13], we construct our own weight distribution, considering that our definition of categories and concepts is a grouping and in some cases the division of the innovation practices. This weight distribution can be observed in Table 2 , where also for every category the local weight of each concept of a category is indicated.
To establish the reference profiles, we will define four areas for each of the six categories using the construction of limiting profiles; this will allow us to establish min and max values in every category c_i . In this case we constructed a symmetrical division in each of the categories because the necessary information to establish central profiles was not available (Table 3).

4.2 Second step: Survey application and data processing.

In this step we work with the information collected from the surveys made to each enterprise. According to the evaluation of each of the concepts, an evaluation for each category c_i is made by the weighted sum $\sum e_{ij} w_{ij}$, where e_{ij} is the evaluation between $[0, 1]$ of the j-th concept in the i-th category , and w_{ij} is the local weight (Table 4).

Category (c_i)	Local Concept Weights (w_{ij})	Global Category Weight (w_i)
I. Creation / Concept Generation	{0.26; 0.33; 0.41}	0.175
II. Conception Activities	{0.43; 0.19; 0.38}	0.107
III. Human Resources Management	{0.47; 0.53}	0.068
IV. Strategy	{0.05; 0.51; 0.27; 0.17}	0.232
V. Project management	{0.01; 0.47; 0.52}	0.194
VI. Capitalization of Ideas and Concepts	{0.43; 0.29; 0.37}	0.224

Table 2: Local and Global weights concepts

Limiting Profile	c_1	c_2	c_3	c_4	c_5	c_6
r_1	1	1	1	1	1	1
r_2	0.75	0.75	0.75	0.75	0.75	0.75
r_3	0.5	0.5	0.5	0.5	0.5	0.5
r_4	0.25	0.25	0.25	0.25	0.25	0.25
r_5	0	0	0	0	0	0

Table 3: Limiting Profiles defined in Flow-Sort Method

A	c_1	c_2	c_3	c_4	c_5	c_6
E_1	0.31	0.32	0.12	0.13	0	0
E_2	0.36	0.59	0.40	0.86	0.15	0.19
E_3	0.28	0.16	0.40	0.58	0	0.10
E_4	0.17	0.22	0.12	0.43	0.11	0.19
E_5	0.26	0.42	0.24	0.34	0.15	0
E_6	0.52	0.78	0.52	0.98	0.64	0.42
E_7	0.17	0.2	0.18	0.29	0.34	0.11

Table 4: Evaluation results

4.3 Third Step: Flow-Sort Aplication

By defining the set of actions for the seven enterprises as $A = \{E_1, E_2, E_3, E_4, E_5, E_6, E_7\}$, which have been evaluated in the six criteria already defined, and the four classification categories {*Passive, Reactive, Preactive, Proactive*} defined by the five limiting profiles of Table 3, we start calculating the preference degrees, showed in Table 5 between the reference profile and the seven enterprises in order to obtain positive and negative flows. With these calculations we can measure positive, negative, and net flows for each enterprise by using equations (3.1), (3.2), and (3.3).The calculations of all of the flows for the enterprises is shown in Table 6, where for example the positive flow of enterprise 1 with respect to limiting profile r_4 is calculated as

$$\phi^+_{R_1}(r_4) = \frac{\sum \pi(r_4, r_j) + \pi(r_4, E_1)}{|R_4| - 1} = \frac{1 + 0.67}{6 - 1} = 0.334$$

The assignment to each category in Table 6 were obtained by equations (3.4) and (3.5); for example for assigning Enterprise 1 the flow is between profile limits r_4 and r_5 as indicated by $\phi^+_{R_1}(r_4) \geq \phi^+_{R_1}(E_1) \geq \phi^+_{R_1}(r_5)$ and $\phi^-_{R_1}(r_4) < \phi^-_{R_1}(E_1) \leq \phi^-_{R_1}(r_5)$. Thus Enterprise 1 is classified as *Passive*. In the cases when positive and negative flows difers, for example Enterprise 3 and 5, we must apply the equation (3.6) to obtain an unique classification.

	r_1	r_2	r_3	r_4	r_5
$\pi(E_1, r_j)$	0	0	0	0.33	0.67
$\pi(r_j, E_1)$	1	1	1	0.67	0
$\pi(E_2, r_j)$	0	0.17	0.33	0.67	1
$\pi(r_j, E_2)$	1	0.83	0.67	0.33	0
$\pi(E_3, r_j)$	0	0	0.17	0.50	0.83
$\pi(r_j, E_3)$	1	1	0.83	0.50	0
$\pi(E_4, r_j)$	0	0	0	0	1
$\pi(r_j, E_4)$	1	1	1	1	0
$\pi(E_5, r_j)$	0	0	0	0.50	0.83
$\pi(r_j, E_5)$	1	1	0	0.50	0
$\pi(E_6, r_j)$	0	0	0	0.50	0.83
$\pi(r_j, E_6)$	1	1	1	0.50	0
$\pi(E_7, r_j)$	0	0	0	0.33	1
$\pi(r_j, E_7)$	0	0	0	0.67	0

Table 5: Preference degrees between the reference profile and the actions

		r_1	r_2	r_3	r_4	r_5	E_i	Class
R_1	ϕ_+	1	0.8	0.6	0.344	0	0.173	Passive
	ϕ_-	0	0.2	0.4	0.656	0.916	0.744	Passive
	ϕ_{net}	1	0.6	0.2	-0.313	-0.916	-0.571	Passive
R_2	ϕ_+	1	0.754	0.532	0.284	0.0	0.431	Reactive
	ϕ_-	0	0.246	0.468	0.716	1	0.569	Reactive
	ϕ_{net}	1	0.507	0.064	-0.433	-1	-0.139	Reactive
R_3	ϕ_+	1	0.8	0.554	0.305	0	0.303	Passive
	ϕ_-	0	0.2	0.446	0.695	0.961	0.659	Reactive
	ϕ_{net}	1	0.6	0.107	-0.39	-0.961	-0.356	Reactive
R_4	ϕ_+	1	0.8	0.6	0.354	0	0.246	Passive
	ϕ_-	0	0.2	0.4	0.646	1	0.754	Passive
	ϕ_{net}	1	0.6	0.2	-0.293	-1	-0.507	Passive
R_5	ϕ_+	1	0.8	0.6	0.297	0	0.258	Passive
	ϕ_-	0	0.2	0.4	0.703	0.955	0.697	Reactive
	ϕ_{net}	1.0	0.6	0.2	-0.406	-0.955	-0.439	Passive
R_6	ϕ_+	1.0	0.732	0.445	0.2	0	0.623	Preactive
	ϕ_-	0.0	0.268	0.555	0.8	1	0.377	Preactive
	ϕ_{net}	1.0	0.464	-0.11	-0.6	-1	0.246	Preactive
R_7	ϕ_+	1.0	0.8	0.6	0.315	0	0.285	Passive
	ϕ_-	0.0	0.2	0.4	0.685	1	0.715	Passive
	ϕ_{net}	1.0	0.6	0.2	-0.37	-1	-0.43	Passive

Table 6: Flow-Sort Results

The assignment of an enterprise into one of the four enterprises profiles can be easily displayed using the positive and negative flows diagram. In Figure 1, which shows the flows of the Enterprise 1, it is possible to view that there is not ambiguity in the classification of this enterprise into the *Passive* profile , since the positive and negative flows allocate this enterprise into the same profile. In other case , in the Figure 2 which shows the flows of the Enterprise 3, we can observe that the positive flow classifies the enterprise into the *Passive* profile and the negative flow classifies the enterprise into the *Reactive* profile, but the final net flow classify this enterprise into the *Reactive* profile. This final classification can be explained taking into consideration the

proximity of the positive and negative action score to the *Reactive* area classification over the Passive area classification.

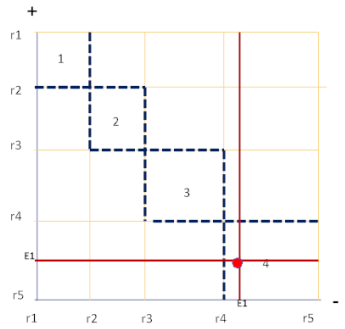


Figure 1: Flow Diagram Enterprise 1

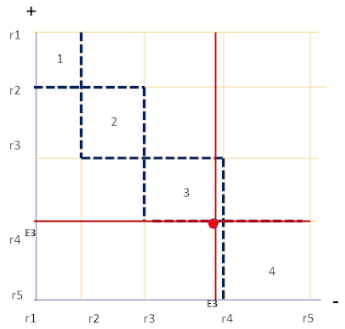


Figure 2: Flow Diagram Enterprise 3

For the Enterprise 6 identified as *Preactive* in this method, the net flow action obtains a positive value, which is opposed to the net values obtained by the enterprises classified as *Passive* which have the most negative value (Figure 2). It is important to note that the profile areas where there is not exist any ambiguity depends only on the singular comparison between an enterprise and all of the reference profiles. The former explains the variety of these areas that we see in all these flows diagrams.

4.4 Fourth Step:Analysis of the results.

The analysis of the results and the search of possible alternatives of innovation progress in these enterprises can be analyzed by observing the detailed net score flows for each category, which show the net significance of all six Categories in obtaining the final net flow value for every enterprise. We can observe that in most cases the category that contributes with the highest negative flows is the *Category Capitalization of Ideas*, which is responsible for the third part of the total net flow. The second category with the most negative flows is the category *Project Management*. In most enterprises, these two categories grouped together represent between the 40% or 50% of the final significance in the net flow action (Table 7). This can give us an idea

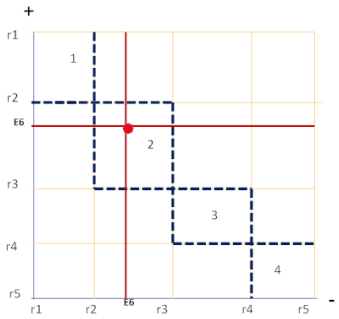


Figure 3: Flow Diagram Enterprise 6

that any improvement in the values obtained in the values of these categories can be critical to obtain a better evaluation. In other words, any improvement in any of the concepts that belongs to the categories above mentioned may produce an advance of the innovation process. Thus, these enterprises would be more qualified to move to a higher profile. On the other hand, in the enterprises that obtained a classification as *Reactive* and *Preactive*, the values associated to the Category Strategy have positive net flows values having a positive significance into the total net flow action; thus we can say that enterprises with good evaluations in the concepts concerning to the *Strategy* can present features of a *Reactive* or a *Preactive* enterprise profile.

	c_1	c_2	c_3	c_4	c_5	c_6
E_1	6.13%	3.75%	7.15%	24.39%	27.19%	31.39%
E_2	7.61%	4.65%	2.96%	30.26%	25.30%	29.22%
E_3	7.80%	14.30%	3.03%	10.34%	34.58%	29.95%
E_4	20.70%	12.66%	8.04%	9.15%	22.95%	26.50%
E_5	7.97%	4.87%	9.29%	10.56%	26.50%	40.80%
E_6	10.43%	19.13%	4.05%	41.48%	11.56%	13.35%
E_7	24.44%	14.94%	9.50%	10.80%	9.03%	31.28%

Table 7: Detailed Net Flow Significance

5 Conclusion

In this paper a new enterprise classification method by using a non compensatory flow-based sorting method with limiting profiles has been proposed. The method called Flow-Sort was used to identify the current level of the enterprise and to classify it into four predefined levels of innovation: *Passive, Reactive, Preactive*, and *Proactive*. The aim of the method, besides classifying the enterprise into a profile, is to give new ideas to formulate an improvement strategy to allow the company to increase its innovation performance. Along this work, we established many observations, as (a) the importance of defining a precise framework that allow us to evaluate all the characteristics and concepts in a precise and structured form, (b) the correct construction of the tool for gathering data, since this is a key element for obtaining the necessary information input, (c) the use of a mathematical tool that allows a comparison against an established profile; the tool is independent of the universe of enterprises to be measured so that it can be applied on a reduced or a large number of enterprises without changing its effectiveness, (d) the possibility of establishing the parameters for an innovation improvement strategy for each of the enterprises individually according to their obtained values and the analysis of significance of each of the categories and concepts.

As a limitation of the application used in the example above, we have the highly arbitrary definition of the four reference profiles since we used four homogenous zones for the six categories when using limiting profiles, as shown in Table 3. More accurate knowledge on a company may allow different values for the profiles, or better the use of central profiles. The latter may be continuously refined as the analysis is repeated in the mid and long term. A classification into new innovation levels can be made by using the proposed method. In such a case, the use of either central or limiting profiles will exclusively depend on the certainty about the characteristics of each new level.

Another comment is related to the interpretation of results and the creation of incentive policies towards better innovation performance levels; these aspects are left to the policy makers at each company since they cannot be predefined in a standardized way. However, an empirical study

with a broader universe of companies by using the method here proposed could give a better answer on the suitability of best practices of this field.

Bibliography

[1] Kerssens-van Drongelen, I.C. and Bilderbeek, J., R&D performance measurement: more than choosing a set of metrics, *R& D Management* vol 29, nş1, p. 35-46, 1999

[2] Koberg C., Detienne D., Heppard k.,An empirical test of environmental, organizational, and process factors affecting incremental and radical innovation, *Journal of High Technology Management Research* 14 p 21-45, 2002

[3] Adams, R., Bessant, J., Phelps, R., Innovation management measurement: A review, *International Journal of Management Reviews*, vol 8 , nş1, p. 21-47. 2006

[4] Bremser W., Barsky N., Utilizing the balanced scorecard for R&D 5 performance measurement, *Management* vol 34 nş3, p. 229-238, 2004

[5] Wang C.,Lu I., Chen I, Evaluating firm technological innovation 15 capability under uncertainty, *Technovation* vol 28, p. 349-, 2008

[6] Chiesa, V., Coughlan, P., Voss, C.A., Development of a technical innovation audit, *Journal of Product Innovation Management* 13 (2), p 105-136, 1996

[7] Becheikh N., Landry, R., Amara, N., Lessons from innovation empirical studies in the manufacturing sector: A systematic review of the literature from 1993-2003, *Technovation*, vol 26, nş5-6, p. 644-664,2005

[8] Cormican k.,Sullivan D.,Auditing best practice for effective product innovation management,*Technovation* vol 24, p. 819-829, 2004

[9] Boly, V., *Ingénierie de l'innovation : organisation et methodologies des entreprises innovantes*,Lavoisier, Paris, France, 2004

[10] Corona A. José Ramón, *Innovation et metrologie : une approche en terme d'Indice d'Innovation Potentielle*, Thčse de Doctorat, Institut Nationale Polytechnique de Lorraine, février, 2005

[11] Godet, M., *Manuel de prospective stratégique. Tome 2. L'art et la methode*, Ed. Dunod, Paris, France, 1997

[12] Morel L., Camargo M.,Comparison of multicriteria analysis techniques to improve the innovation process measurement, IAMOT 2006, Beijing, China, 8 pages, may 22-26

[13] Assielou G, *Metrologie des processus d'innovation*, thčse de Doctorat, Institut Nationale Polytechnique de Lorraine, 2008

[14] Camargo M., Morel L., Fonteix C., Evolutionary based methodology to integrate product innovation degree on a firm technological strategy, *IAMOT* 2007 Proceedings

[15] Nemery P, Lamboray C. *FlowSort: a flow-based sorting method with limiting or central profiles*, TOP 16:90-113, Springer-Verlag, 2008

Int. J. of Computers, Communications & Control, ISSN 1841-9836, E-ISSN 1841-9844
Vol. V (2010), No. 5, pp. 929-938

Network Coded Transmission in a Wireless Grid Network with an Energy Constraint

R. Stoian, A.V. Raileanu, L.A. Perisoara

Rodica Stoian, Adrian Victor Raileanu, Lucian Andrei Perisoara
Politehnica University of Bucharest
Romania, 061071 Bucharest, 1-3 Iuliu Maniu
E-mail: {rodicastoian2004, adrianrai, lperisoara}@yahoo.com

Abstract: In wireless networks, routing based on packet forwarding does hardly yield optimum transmission performance in terms of network utilization and throughput. As an alternative to routing, network coding has been introduced in the recent years, where nodes are mixing the data instead of forwarding. In applications, random linear network coding is the most used method, due to its decentralized mode, and due to preserving the achievability of multicast capacity bounds. In this paper, we study the performance of network coding used for multicast transmission of messages in a wireless grid network with an energy constraint. Several energy saving schemes have been proposed in the literature, but in this study we will focus on duty cycling scheme, in which nodes are not always in *on* state. The performance is measured as the end-to-end delay, i.e. the duration until each node can decode the message sent by the source, and the CDF of observations is used to make analysis.
Keywords: network coding, energy efficient, end-to-end delay, duty cycling.

1 Introduction

Energy saving is an important factor in wireless transmissions, especially in autonomous devices, i.e. battery operated nodes. In applications like battlefield surveillance, environment and habitat monitoring, sometimes it is hostile, hazardous or impractical to replace or recharge the batteries. The performance of wireless network applications highly depends on the lifetime of the network. For practical applications we expect the lifetime to be from several months to several years, so energy saving is crucial in designing the network.

Energy consumption in a network node can be due to useful sources (transmitting, receiving or processing data) or wasteful sources (channel idle listening, retransmissions due to packet collisions, overhearing, control packets used for errors control). The critical issue is to minimize the energy consumption of network nodes while meeting the application requirements.

This paper is organized as follows. In Section II we explain how network coding is applied for wireless networks, marking some advantages of using it. In Section III we describe the scenario of a general multicast transmission in a wireless network. In Section IV we present the problem of energy consumption optimization using duty cycling. Finally, in Section V we present simulation results.

2 Network Coding and Wireless Networks

Network coding is a recent field of Information Theory that breaks the classical assumption about the routing in the networks. Instead of simply forwarding the packets, the intermediate nodes recombine several input packets into one or several output packets. In [1], Ahlswede et

al. showed that network coding achieves the multicast capacity, which is defined as a maximum data rate which is achieved for a multicast transmission. In [2], it is shown that the maximum multicast capacity can be achieved by using linear encoding functions at each node, which implies to solve linear equations at the receiver.

In Figure 1 we show a simple example of using network coding to reduce the number of transmissions used to exchange two bits b_1 and b_2, the operation applied being XOR. With network coding, the first node can recover the bit b_2 from the received bit b_1+b_2 and the known bit b_1. Similarly, b_1 can be recovered at the second node. Network coding can reduces the traffic without increasing delay and so it can save energy by reducing the amount of transmitted data.

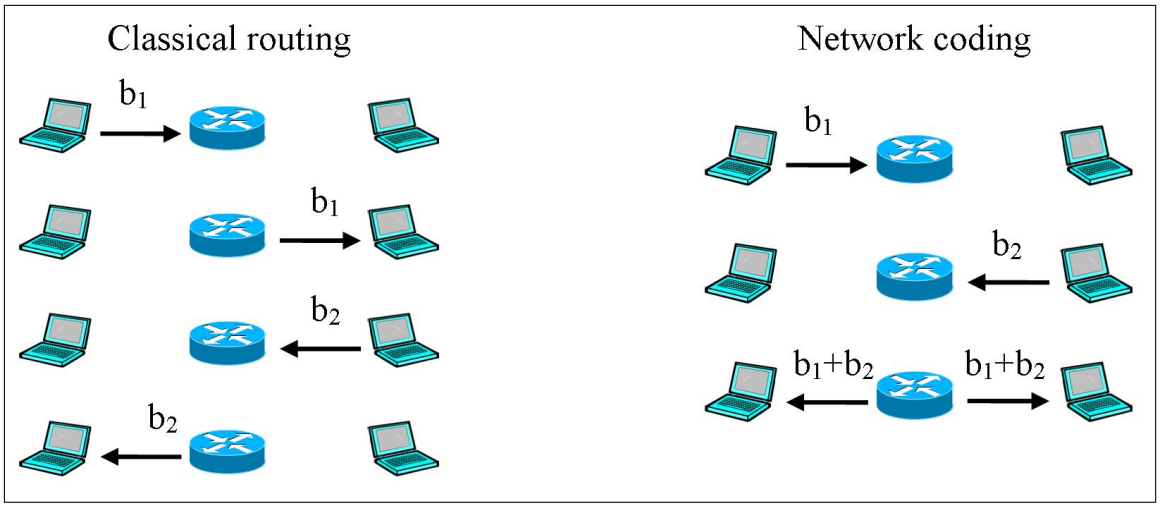


Figure 1: An example of decreasing the transmission time using network coding

Let suppose for a network that source node s emits K information packets $\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_K$, each of length L symbols from a finite field $GF(q)$ to N receivers $t_1, t_2, ..., t_N$. For linear network coding, each node combines a number of received packets into one or several output packets:

$$\mathbf{y} = \sum_{i=1}^{K} \alpha_i \mathbf{x}_i \tag{2.1}$$

where the summation is applied for every symbol position. For random linear network coding, the coefficients α_i of the linear combination are generates in a random manner, which assures with high probability a linear independence of the output packets from a node for a sufficiently large size $q = 2^m$ of the finite field $GF(q)$, as it was proved in [3]. The encoding coefficients forms the encoding vector $\alpha = (\alpha_1, \alpha_2, ..., \alpha_K)$, which belongs to a K-dimensional vector space over $GF(q)$. All encoding vectors associated with the output edges of all intermediate nodes from s to a specific node t forms the encoding matrix.

When we refer to a network code we must specify the all encoding vectors which should be used for the encoding process, for all edges of the network. The encoding coefficients are send to the destination in the packet header, so the destination nodes can decode the packet without knowing the network topology or the encoding rules, even if the nodes are added or removed in an ad-hoc manner.

Assume a node t has received M coded packets $\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_M$ and the encoding vectors $\alpha_1, \alpha_2, ..., \alpha_M$. To decode the packets, it need to solve a linear system with M equations and K unknowns $\mathbf{x}_i$, derived from (2.1). To solve this system, the node wait until he receives at least $M \geq K$ linearly independent packets, equivalently with M linearly independent encoding vectors.

This condition is assured using buffers for each input of the node. The buffer stores only the innovative packets (packets which are not a linear combination of the already stored packets in the buffer). Non-innovative packets are discarded by Gaussian elimination because they do not provide any new information to the node.

Some advantages of using network coding in wireless networks are throughput and capacity improvements [1], bandwidth and energy savings [4], robustness to noise [5], reduced traffic.

3 A Scenario of General Multicast

We introduce a grid wireless mesh network, which is a generalization of a local network (e.g. office), or a special purpose network deployed over a rectangular area for monitoring or as point of presence (e.g. information panels). The type of communication is one to many, as we consider one original source (e.g. a gateway to another network, a controller, etc.), that sends messages to all the other points, called nodes. This is called general multicast, as one message is sent to a number of participants (e.g. all participants), but because not all nodes are in the radio range of the source, the message is relayed node by node.

Transmission over radio is simplified, considering only one channel, and without collisions. The radio signal is affected by distance attenuation and a simple exponentially distributed noise floor. The network stack is reduced to simple MAC/IP layers, with the purpose of taking into consideration only MAC latency and for identifying nodes by addresses.

The position of the source is at one corner of the rectangle (see Figure 2), as it simulates a gateway or a controller. At the same time, the position was chosen as it provides the worst case scenario, where the source has the lowest number of neighbors possible in the given situation.

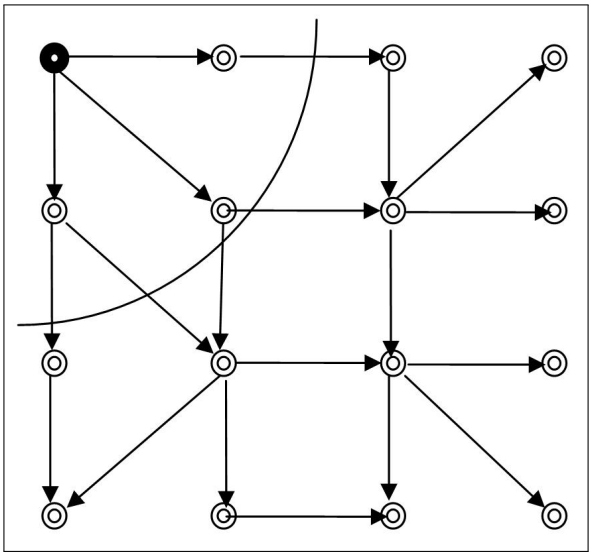


Figure 2: General multicast relay in a wireless grid network, with source in corner position. The radio range of the source is not large enough for broadcast

The message unit is considered 1 byte. The entire transmission is an M byte message generated by the source. In normal networks, this would take M consecutive transmissions from the source, and some additional ones when there is no acknowledge received. Our model uses random network coding to disseminate information, so that the source emits linear combination of all the message bytes at each transmission. A number of $K = 32$ random coefficients from GF(256) is used for messages with $M < K$. The original message is padded with zeros and each

byte is multiplied the corresponding coefficient k_i and summed (i.e. XOR-ed). The result is one mixed byte and a list of coefficients that are used for decoding at the receiving nodes. For bandwidth saving, coefficients can be chosen from lower Galois fields, e.g. GF(2), reducing the size of the coefficient list. However, using lower Galois fields increases the probability that two packets will be linear dependent, and will not contribute to the decoding process.

Each receiving node accumulates linear independent packets (novel packets), based on analysis of the coefficient list included in each packet. When it has received at least K such packets, it decodes the message using matrix inversion. A node that receives a novel packet will re-encode and send it using a linear combination of all its received packets. If a node receives a packet that is linear dependent to its received packets, it will discard it, without forwarding it.

When a node decodes the entire message, it becomes a source, generating packets with its own generated random coefficients. Generating acknowledge messages is out the scope of this paper.

We evaluate transmission performance using the end-to-end delay (ETED) metric, calculated as the time since the source emitted the first packet until a specific node received and decoded the entire message. The overall end-to-end delay (OETED) is the maximum end-to-end delay, i.e. calculated at the last node that received and decoded the message. The minimum end-to-end delay is the time until the first successful transmission (i.e. first node that decodes the full message). The average end-to-end delay is the arithmetic average of all the successful transmission times.

4 Energy Consumption Optimization

The two main operations that require optimizations of energy are transmission and reception. Transmission energy is optimized through radio power adjustment, so that a tradeoff is done between the range and battery life. In our scenario, all the nodes use the same transmission power. In non-centralized networks, however, reduction of transmission energy can be achieved also by reducing the number of redundant sending operations in nodes. In flooding based routing, nodes tend to forward each packet received, creating a lot of duplicated information inside the network. Widmer et. al [12] proposed network coding algorithms that reduce the number of transmissions per each node compared to flooding. In [13], Fragouli et al. studied distributed algorithms to achieve optimum number of transmissions in grid and random topologies. The results show that network coding can achieve up to 30% energy reduction compared to probabilistic routing.

The current study focuses on the optimization of reception and decoding, analyzing how classical energy saving schemes affect performance of network coded transmissions. Idle listening and overhearing are major sources of energy consumption in a wireless network. While the first one can be easily optimized, the second one is of great importance to the network coding overall architecture. Reducing overhearing in a network coded system may reduce its performance. Our article studies the degradation of performance when implementing different levels of energy saving. The method is a simplified version of the S-MAC protocol, which use a listen/sleep cycle. The energy consumption is reduced by switching on and off each node independently, so that reception is performed only in some discreet time windows. During *power on* states, a node is able to receive, decode and retransmit data. During *power off* states, a node is not able to receive any packet, so it will be lost. This should not be an issue with our current wireless network coding scenario, as neighbor nodes can be in different power states, and take advantage of all the packets.

Denote T, a full *power on - power off* time interval. We consider an energy duty cycle, the percentage of power on states. For example, a duty cycle of 25% means that the node is on for

time T/4, then off for time 3/4T, then again on for T/4, and so on. The power states have random time offsets, so that to avoid situations when all the nodes are in power off state, see Figure 3.

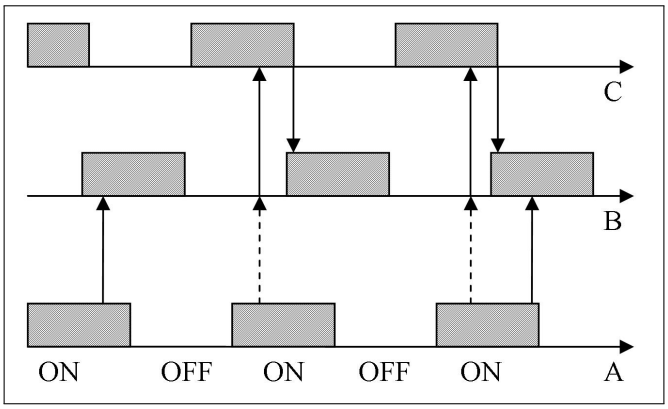


Figure 3: Transmission from node A to neighbors B and C. Packets that were missed by B in off state (dotted arrow), are recovered later from neighbor C

5 Simulations and Results

This article studies the effects of energy optimizations on the end-to-end delay in the wireless grid network, using random network coding.

Based on [11], we developed a wireless network simulator, which uses network coding for transmission. We chose a fixed number of nodes $N = 100$, that are arranged on a 10×10 square area. The source message is made of packets of length $M = 32$ bytes, equal to the number of coefficients, $K = 32$. The main simulator functional modules are: source, node and scheduler. The source module is responsible to encode the original message and continuously send differently encoded packets to the other nodes. The other nodes contain logic for packet decoding and rank calculation, but also include a packet re-encoder and sender, for relay. The scheduler is responsible for sequencing the packet transfer in the network. Transmission hops are determined based on a physical model (as defined in [14], [15]), where instant noise level and attenuation decide whether a node is in the transmission range or not. For studying energy efficiency algorithms, the scheduler has been enhanced to support on-off node states. The simulator engine is event based, so that each transmission and reception has a different timestamp. The link rate is constant for all transmissions, and is simulated as an inter-sending time interval. The time difference is made more realistic by introducing MAC latency, propagation time and a random jitter for all other factors that are not explicitly simulated.

The first series of simulations were used to analyze the impact of the distance between nodes on the end-to-end delay (ETED). We run simulations for distance $d = 40$ m to $d = 179$ m. For $d < 40$ m the networks is close to a broadcast network, (i.e many more than 50% of the nodes are in the source radio range), so it is out of the scope of the current paper. At $d = 180$ m all the nodes lost connectivity. Each simulation was repeated three times.

Figure 4 shows that for large and medium range coverage, given by $d = 80$ m and $d = 120$ m, the nodes completion time increases almost linearly. At very short range, where connectivity is available only with closest neighbors, the last set of nodes is completed in almost exponential time.

Due to the geometry of the network grid, Figure 5 shows two flat regions, for both average and overall ETED.

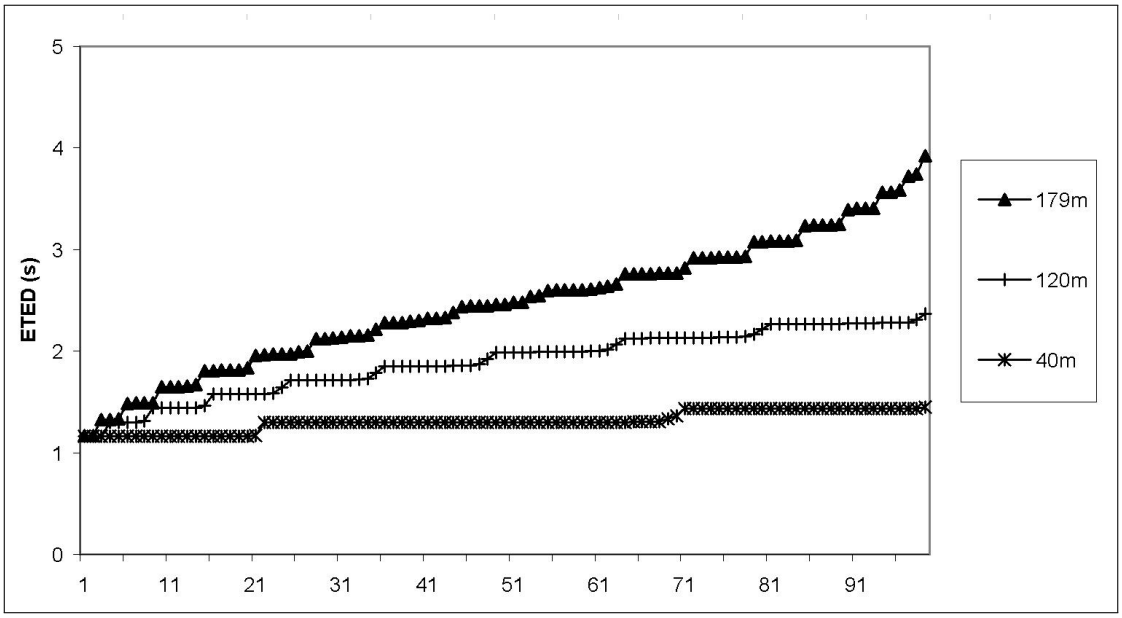


Figure 4: End-to-end delay measured for each node in the network, for different values of d

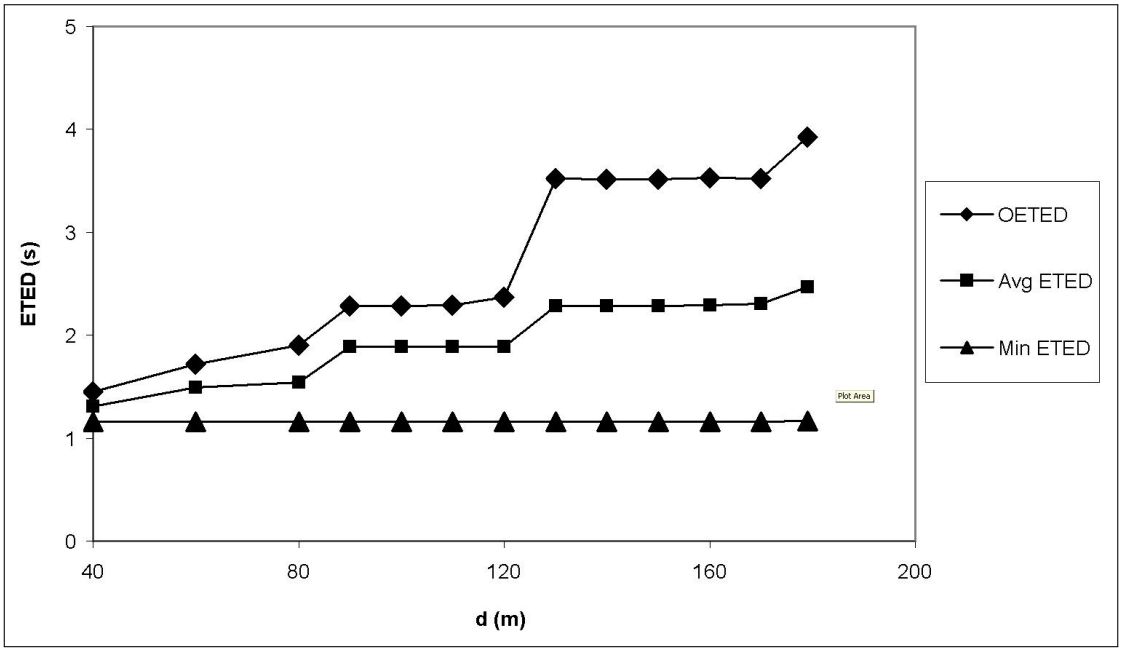


Figure 5: End-to-end delay dependency on different values of d

These are explained by the number of nodes that are available in one range, for a given d. Choosing an optimum d that would maximize the area covered by the network, would be the largest value in any of the flat regions, e.g. 80 m, 120 m, 170 m. Now, observing the ratio between average and overall ETED between different flat regions, we can note that only the third region (120 - 170m) has a linear node completeness behavior, while the other ones tend to be logarithmic. In a network where there should not be too many nodes late than the average, choosing $d < 130$ m would be the right choice. The ratio between the areas covered by the network for the three values of d (80, 120, 170), is 0.22:0.5:1. The ratio observed on average ETED is 0.67:0.82:1 and for overall ETED is 0.54:0.67:1. If average ETED is of interest,

maximizing the area leads to a value of $d = 170$ m. If the overall ETED is important, than the value of $d = 120$ m is offering half of the maximum area but with a 30% reduction of delivery time.

In the next stage, for a fixed $d = 120$ m, we have added to the simulation scenario a power cycle with a duty of 50%. We performed simulations for different values of the power on-off time interval, with T in a range from 10 ms to 500 ms. Values lower than 10 ms may not be efficient due to electrical / logical reasons in the device, i.e. network processing chip switching on-off too fast. Values above 500 ms are of no interest to the current study, as it reduces the number of switches per experiment to less than 10.

In Figure 6, at $T = 0$ are drawn the results when there is no power cycle, for reference. As, expected results show better behavior at lower T values, as the nodes have a quicker opportunity to recover lost packets from their neighbors.

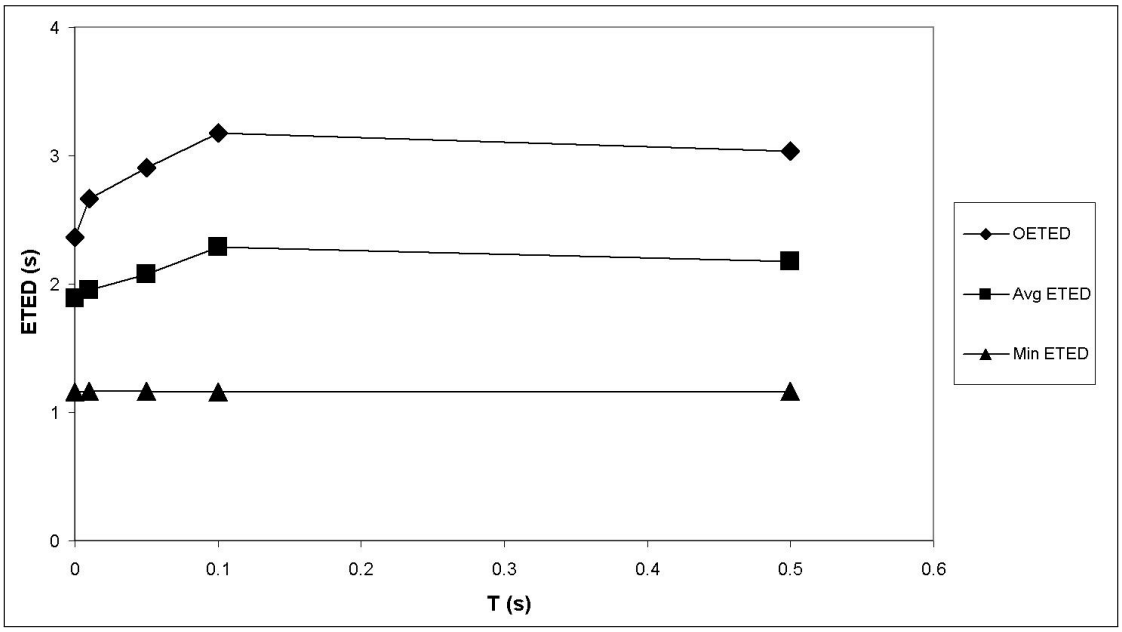


Figure 6: End-to-end delay dependency on different values of T

In the next experiment, we studied the effect of different duty cycles, ranging from 10% to 90%. We observed the behavior of the network for three values of $T = [0.010, 0.100, 0.500]$ s. Sometimes, the energy advantage comes from using a duty cycle lower than 50% (i.e. less than half of the energy used).

Figures 7 and 8 show that it is possible to preserve low values of ETED even at a duty cycle of 30%. However, for $T > 0.100$ s, the optimal duty cycle is close to 60%.

We performed another set of measurements with fixed $T = 0.100$ s, for different duty cycles, and for different values of distance d = [80, 120, 170] m.

Figures 9 and 10 show that the 80 m and 120 m networks are more sensitive to values lower than 30%, while the 170m network has a higher threshold of 70%.

With network coding, the fact that there are only a few nodes connected to each other, does not allow improvement of energy savings. As soon as the number of connections per node doubles, the energy consumption can be optimized up to 30%.

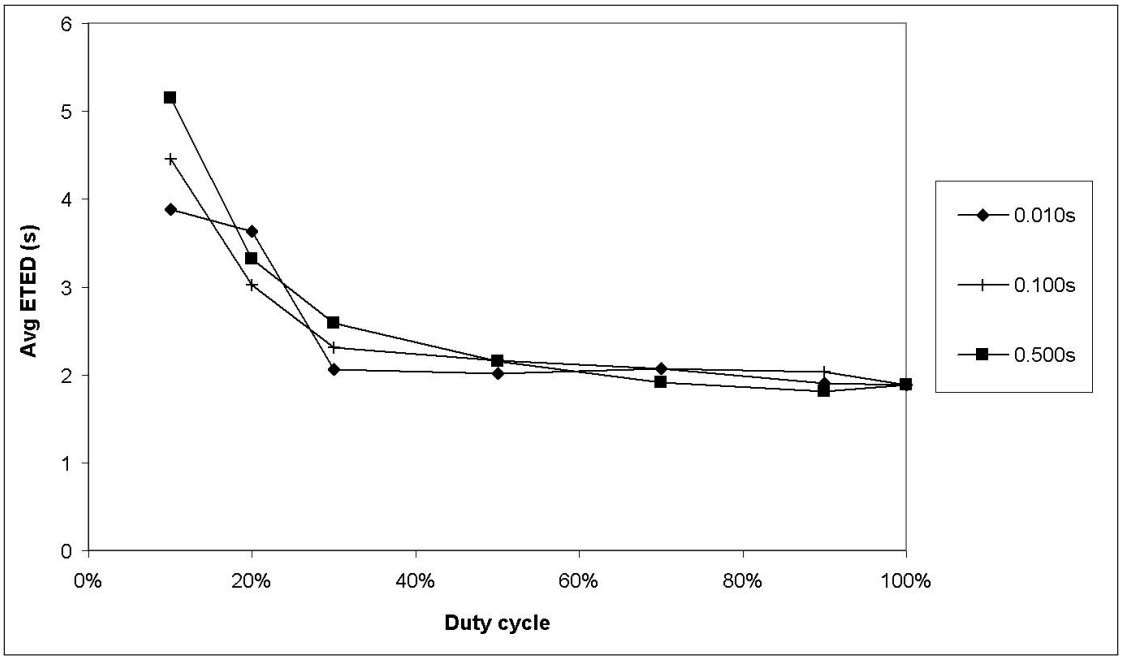


Figure 7: Average end-to-end delay dependency on different values of the duty cycle, for $T =$ [0.010, 0.100, 0.500] s

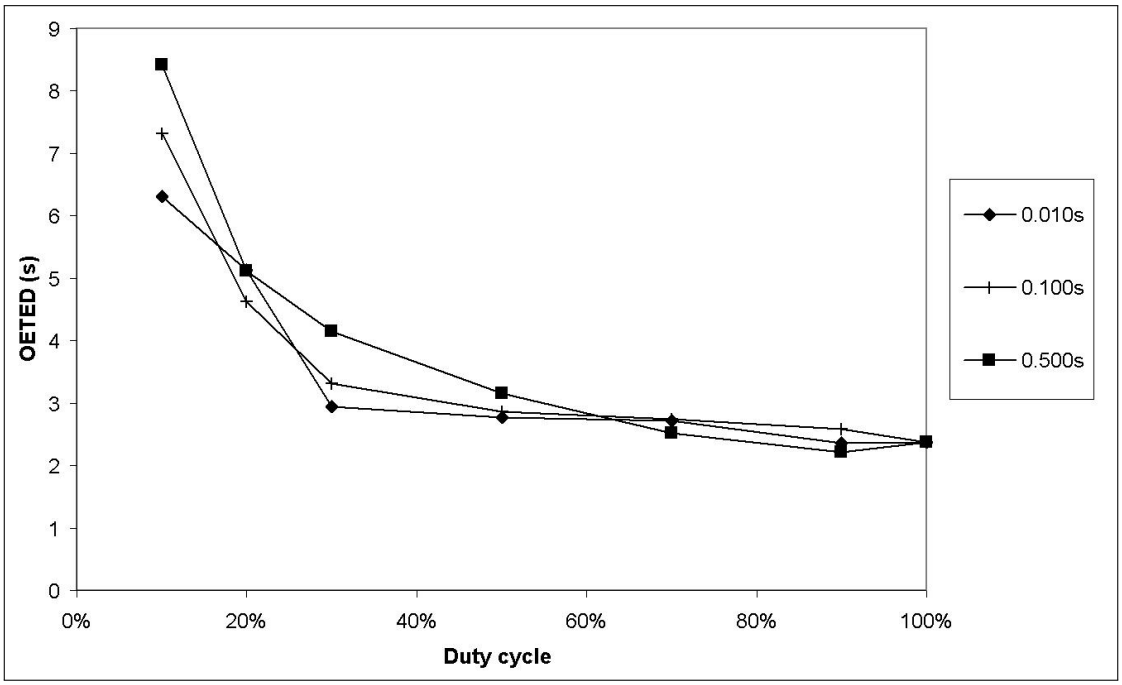


Figure 8: Overall end-to-end delay dependency on different values of the duty cycle, for $T =$ [0.010, 0.100, 0.500] s

6 Conclusions

This article presents methods for saving energy in a wireless network, in the case of general multicast transmission. The routing of packets inside the network is not store and forward, but mix and forward, namely network coding. We observed the end-to-end delay inside a wireless

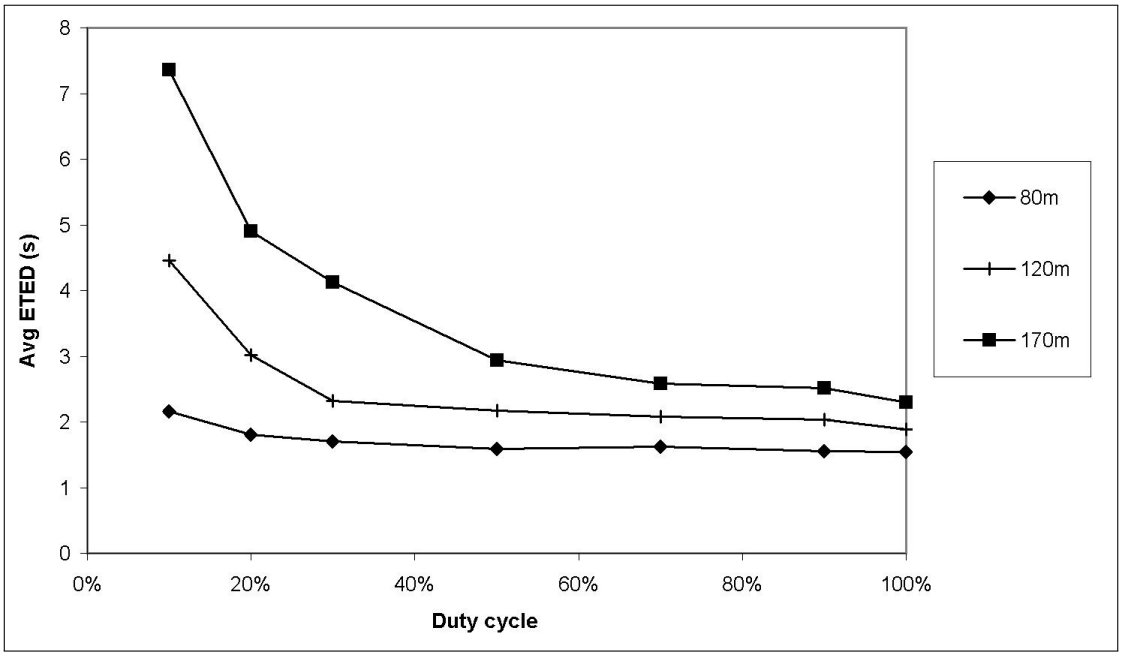


Figure 9: Average end-to-end delay dependency on different values of the duty cycle, for $d = [80, 120, 170]$ m

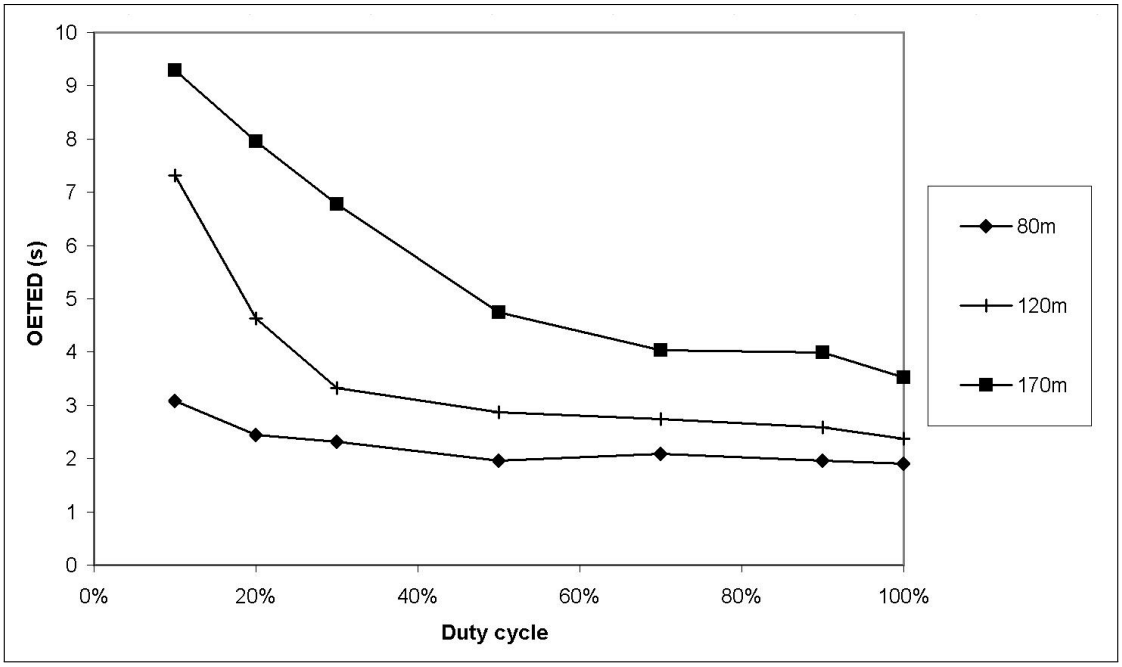


Figure 10: Overall end-to-end delay dependency on different values of the duty cycle, for $d = [80, 120, 170]$ m

grid network that uses network coding, for different area sizes (i.e. different node density) and for various reception windows for energy saving. A good tradeoff between area and end-to-end delay is to choose half of the maximal area, while obtaining at least 30% improvement in speed. Simulations show that energy consumption can be reduced to between 30% and 50%, depending on the duration of the full power on - power off cycle. For sparse networks, the energy

consumption can be reduced up to 70%.

Further study may reveal important results for other network topologies, other type of transmissions (e.g. unicast), other energy saving schemes or other realistic features (variable range, transmission power, interference, multiple radio channels, etc).

Bibliography

[1] R. Ahlswede, N. Cai, S. Y. R. Li, and R. W. Yeung, *Network information flow,* IEEE Trans. Inform. Theory, vol. 46, no. 4, pp. 1204-1216, July 2000.

[2] S. Y. R. Li, R. W. Yeung, and N. Cai, *Linear network coding,* IEEE Trans. Inform. Theory, vol. IT-49, no. 2, pp. 371-381, Feb. 2003.

[3] Y. Wu, P. A. Chou, and K. Jain, *A comparison of network coding and tree packing,* in Proc. ISIT 2004, Chicago, June 2004.

[4] Y. Wu, P. A. Chou, and S. Y. Kung, *Minimum-energy multicast in mobile ad-hoc networks using network coding,* IEEE Trans. Communications, vol. 53, no. 11, pp. 1906-1918, Nov. 2005.

[5] Rodica Stoian, L. A. Perisoara, and Radu Stoica, *Random Network Coding for Wireless Ad-Hoc Networks*, in Proc. on Int. Symp. on Signals, Circuits and Systems (ISSCS 2009), Iasi, Romania, vol. 2, pp. 469-472, July 9-10, 2009.

[6] G. Lu, N. Sadagopan, B. Krishnamachari, and A. Goel, *Delay efficient sleep scheduling in wireless sensor networks*, in Proc. of IEEE INFOCOM 2005, March 2005.

[7] Ming Xiao, Tor M. Aulin, *Energy-Efficient Network Coding for the Noisy Channel Network*, ISIT 2006, Seattle, USA,pp. 778-782, July 9-14, 2006.

[8] W. Ye, J. Heidemann, and D. Estrin, *An Energy-Efficient MAC Protocol for Wireless Sensor Networks,* in Proc. of IEEE INFOCOM, June 2002.

[9] Y. Yang, B. Krishnamachari, and V.K. Prasanna, *Energy-latency trade offs for data gathering in wireless sensor networks,* in Proc. of IEEE INFOCOM, March 2004.

[10] J. H. Chang, and L. Tassiulas, *Energy Conserving Routing in Wireless Ad-Hoc Networks*, in Proc. of IEEE INFOCOM, March 2000.

[11] SlimSim simulator, http://cs.anu.edu.au/ aaron/sim.php

[12] J. Widmer, C. Fragouli, and J.Y. Le Boudec, *Low-complexity energy-efficient broadcasting in wireless ad-hoc networks using network coding*, First Network Coding Workshop (NETCOD), Riva del Garda, Italy, 2005.

[13] C. Fragouli, J. Widmer, J.Y. Le Boudec, *A network coding approach to energy efficient broadcasting: From theory to practice*, in IEEE INFOCOM, Barcelona, Spain, Apr. 2006

[14] P. Gupta, and P. R. Kumar., *The Capacity of Wireless Networks,* IEEE Transactions on Information Theory, 46(2):388ď˙ž˝404, March 2000.

[15] Rodica Stoian, and A. Raileanu, *How to Choose a Model for Wireless Networks,* Scientific Bulletin of the "Politehnica" University of Timişoara, Romania, Tom 51(65), Fasc. 2, pp 109-112, Sept. 2006.

Int. J. of Computers, Communications & Control, ISSN 1841-9836, E-ISSN 1841-9844
Vol. V (2010), No. 5, pp. 939-945

On Polar, Trivially Perfect Graphs

M. Talmaciu, E. Nechita

Mihai Talmaciu, Elena Nechita
University of Bacău, Romania
E-mail: mtalmaciu@ub.ro, enechita@ub.ro

Abstract: During the last decades, different types of decompositions have been processed in the field of graph theory. In various problems, for example in the construction of recognition algorithms, frequently appears the so-called weakly decomposition of graphs.
Polar graphs are a natural extension of some classes of graphs like bipartite graphs, split graphs and complements of bipartite graphs. Recognizing a polar graph is known to be NP-complete. For this class of graphs, polynomial algorithms for the maximum stable set problem are unknown and algorithms for the dominating set problem are also NP-complete.
In this paper we characterize the polar graphs using the weakly decomposition, give a polynomial time algorithm for recognizing graphs that are both trivially perfect and polar, and directly calculate the domination number. For the stability number and clique number, we give polynomial time algorithms.
Keywords: Polar graphs, trivial perfect graphs, weakly decomposition, recognition algorithms, optimization algorithms.

1 Introduction

Polar graphs are a natural extension of some classes of graphs like bipartite graphs, split graphs and complements of bipartite graphs.

According to ([3]), a graph $G = (V, E)$ is called polar if the set V of its vertices can be partitioned into (S,Q) (S or Q possibly empty) such that S induces a complete multipartite graph (that is a join of stable sets) and Q is a disjoint union of cliques. In ([3]) has been proved that the problem of recognizing an arbitrary graph to be polar is NP-complete.

Recently some important result concerning polar graphs have been proven. Hereby, in ([9]) the authors give a characterization through forbidden subgraphs of polar cographs and a polynomial algorithm that finds the largest induced subgraph in a cograph. In ([8]) is presented a polynomial algorithm to recognize the polar property for triangulated graphs. In ([7]), a polynomial algorithm for the recognition of graphs that are both polar and permutation is given. In ([16]) they assert that polynomial algorithms for independent set are unknown and algorithms for domination number are NP-complete for split graphs (see ([2] and [4]).

Both problems to find independent maximal set of maximum and minimum weight are NP-hard, in general. In ([12]) are given polynomial time algorithms that solve the problems formulated above for classes of polar graphs.

2 Definition and notation

Throughout this paper, $G = (V, E)$ is a connected, finite and undirected graph, without loops and multiple edges ([1]), having $V = V(G)$ as the vertex set and $E = E(G)$ as the set of edges. $\overline{G}$ is the complement of G. If $U \subseteq V$, by $G(U)$ or $[U]_G$ we denote the subgraph of G induced by U. By $G - X$ we mean the subgraph $G(V - X)$, whenever $X \subseteq V$, but we simply write $G - v$, when

$X = \{v\}$. If $e = xy$ is an edge of a graph G, then x and y are adjacent, while x and e are incident, as are y and e. If $xy \in E$, we also use $x \sim y$, and $x \not\sim y$ whenever x, y are not adjacent in G. A vertex $z \in V$ distinguishes the non-adjacent vertices $x, y \in V$ if $zx \in E$ and $zy \notin E$. If $A, B \subset V$ are disjoint and $ab \in E$ for every $a \in A$ and $b \in B$, we say that A, B are *totally adjacent* and we denote by $A \sim B$, while by $A \not\sim B$ we mean that no edge of G joins some vertex of A to a vertex from B and, in this case, we say that A and B are *non-adjacent*.

The *neighbourhood* of the vertex $v \in V$ is the set $N_G(v) = \{u \in V : uv \in E\}$, while $N_G[v] = N_G(v) \cup \{v\}$; we simply write $N(v)$ and $N[v]$, when G appears clearly from the context. The neighbourhood of the vertex v in the complement of G will be denoted by $\overline{N}(v)$.

If $D \subset V$ and every vertex from $V - D$ has at least one neighbour in D, then D is called a *dominating set* of G. The minimum size of a dominating set is the **domination number** $\nu(G)$.

A **complete** graph is a graph in which every vertex is adjacent to every other.

The neighbourhood of $S \subset V$ is the set $N(S) = \cup_{v \in S} N(v) - S$ and $N[S] = S \cup N(S)$. A *clique* is a subset Q of V with the property that $G(Q)$ is complete. The *clique number* of G, denoted by $\omega(G)$, is the size of the maximum clique.

An **independent** set or **stable** set is a set of vertices of which no pair is adjacent. The **independence number** $\alpha(G)$ of a graph G is the size of a largest independent set of G.

By P_n, C_n, K_n we mean a chordless path on $n \geq 3$ vertices, a chordless cycle on $n \geq 3$ vertices, and a complete graph on $n \geq 1$ vertices, respectively.

The **distance** $d_G(u, v)$ between two (not necessary distinct) vertices u and v in a graph G is the length of a shortest path between them.

A graph is called **triangulated** if it does not contain chordless cycles having the length greater or equal to four.

A graph is called **cograph** if it does not contain P_4.

A graph is a **split** graph if the vertex set can be partitioned into a clique and a stable set.

A graph G is **trivially perfect** ([10]) if for each induced subgraph H of G, the number of maximal cliques of H is equal to the maximum size of an independent set of H.

Let $n \geq 1$ and π be a permutation over $\{1, ..., n\}$. We will denote π equivalently as a **permutation** sequence $(\pi(1), ..., \pi(n))$. The **inversion** graph of π has vertex set $\{1, ..., n\}$ and two vertices u, v are adjacent if $(u - v)(\pi^{-1}(u) - \pi^{-1}(v)) < 0$. A graph is a **permutation** graph if it is isomorphic to the inversion graph of a permutation sequence.

Let F denote a family of graphs. A graph G is called F-*free* if none of its subgraphs is in F. The *Zykov sum* of the graphs G_1, G_2 is the graph $G = G_1 + G_2$ having:

$$V(G) = V(G_1) \cup V(G_2),$$
$$E(G) = E(G_1) \cup E(G_2) \cup \{uv : u \in V(G_1), v \in V(G_2)\}.$$

When searching for recognition algorithms, frequently appears a type of partition for the set of vertices in three classes A, B, C, which we call a *weakly decomposition*, such that: A induces a connected subgraph, C is totally adjacent to B, while C and A are totally nonadjacent.

The structure of the paper is the following. In Section 3 we recall the notion of weakly decomposition. In Section 4 we present a new characterization of polar, trivially perfect graphs. In Section 5 we give a recognition algorithm for this class of graphs. In Section 6 we give combinatorial optimization algorithms for polar, trivially perfect graphs. In the last section we have our concluding remarks.

3 Preliminary results

At first, we recall the notions of weakly component and weakly decomposition.

Definition 1. ([6], [13], [14]) *A set* $A \subset V(G)$ *is called a weakly set of the graph* G *if* $N_G(A) \neq V(G) - A$ *and* $G(A)$ *is connected. If* A *is a weakly set, maximal with respect to set inclusion, then* $G(A)$ *is called a weakly component. For simplicity, the weakly component* $G(A)$ *will be denoted with* A.

Definition 2. ([6], [13], [14]) *Let* $G = (V, E)$ *be a connected and non-complete graph. If* A *is a weakly set, then the partition* $\{A, N(A), V - A \cup N(A)\}$ *is called a weakly decomposition of* G *with respect to* A.

Below we remind a characterization of the weakly decomposition of a graph.

The name of "*weakly component*" is justified by the following result.

Theorem 1. ([5], [13], [14]) *Every connected and non-complete graph* $G = (V, E)$ *admits a weakly component* A *such that* $G(V - A) = G(N(A)) + G(\overline{N}(A))$.

Theorem 2. ([13], [14]) *Let* $G = (V, E)$ *be a connected and non-complete graph and* $A \subset V$. *Then* A *is a weakly component of* G *if and only if* $G(A)$ *is connected and* $N(A) \sim \overline{N}(A)$.

The next result, that follows from Theorem 1, ensures the existence of a weakly decomposition in a connected and non-complete graph.

Corollary 1. *If* $G = (V, E)$ *is a connected and non-complete graph, then* V *admits a weakly decomposition* (A, B, C), *such that* $G(A)$ *is a weakly component and* $G(V - A) = G(B) + G(C)$.

Theorem 2 provides an $O(n + m)$ algorithm for building a weakly decomposition for a non-complete and connected graph.

Algorithm for the weakly decomposition of a graph ([13])
Input: A connected graph with at least two nonadjacent vertices, $G = (V, E)$.
Output: A partition $V = (A, N, R)$ such that $G(A)$ is connected, $N = N(A)$, $A \not\sim R = \overline{N}(A)$.

```
begin
   A := any set of vertices such that
   A ∪ N(A) ≠ V
   N := N(A)
   R := V − A ∪ N(A)
   while (∃n ∈ N, ∃r ∈ R such that nr ∉ E ) do
      begin
         A := A ∪ {n}
         N := (N − {n}) ∪ (N(n) ∩ R)
         R := R − (N(n) ∩ R)
      end
end
```

Corollary 2. *For* $G = (V, E)$ *a connected non-complete graph, and* (A, N, R) *a weakly decomposition with* $G(A)$ *the weakly component the following relation holds:*

$\alpha(G) = \max\{\alpha(G(A)) + \alpha(G(R)), \alpha(G(A \cup N))\}$.

In ([13]) some applications of weakly decomposition have been depicted. Let $G = (V, E)$ be connected, non-complete graph and (A, N, R) a weakly decomposition, with A the weakly component.The following hold:
a) G is $P_4 - free$ if and only if $A \sim N \sim R$ and $G(A)$, $G(N)$ and $G(R)$ are $P_4 - free$;
b) G is triangulated if and only if N is a clique and R and $G - R$ are triangulated.
Each of the results above lead to recognition algorithms for the specified graphs.
c) If G is triangulated then $\alpha(G) = \alpha(G(A)) + \alpha(G(R))$ and this leads to the algorithm that determines $\alpha(G)$.

4 Characterization of polar, trivially perfect graphs

In this section, using the weakly decomposition, we present a recognition algorithm for the polar trivially perfect graphs. At first we remind a characterization in terms of forbidden subgraphs of polar cographs and two characterization of trivially perfect graphs.

Theorem 3. ([9]) *For a cograph* G*, the following statements are equivalent:*
a) G *is polar;*
b) Neither G *nor* $\overline{G}$ *contains any one of the graphs* H_1*,* H_2*,* H_3*,* H_4 *as induced subgraphs, where* $H_i = G_i \cup F_i$ *(*$1 \leq i \leq 4$*), and every* G_i *is a* P_3 *and* F_i*, described as sequences of degrees, are:* $F_1 : (4,3,3,3,3)$*;* $F_2 : (5,3,2,2,2,2)$*;* $F_3 : (4,4,3,3,3,3)$*;* $F_4 : (5,5,3,3,3,3)$.

Theorem 4. ([15]) *If* G *is a connected, non-complete graph and* (A, N, R) *is a weakly decomposition with* $G(A)$ *the weakly component, then* G *is trivially perfect if and only if:*
i) $A \sim N \sim R$*;*
ii) N *is clique;*
iii) $G(A)$*,* $G(R)$ *are trivially perfect if and only if it contains no vertex subset that induces* P_4 *or* C_4.

Theorem 5. ([12]) *A graph is trivially perfect if and only if it contains no vertex subset that induces* P_4 *or* C_4.

Theorem 6. *Let* $G = (V, E)$ *be a connected, non-complete graph and* (A, N, R) *a weakly decomposition with* $G(A)$ *the weakly component. Let also* G *and* $\overline{G}$ *be trivially perfect graphs.* G *is polar if and only if* $G(A)$ *and* $G(R)$ *are polar graphs.*

Proof. If G is a polar graph then $G(A)$ and $G(R)$ are polar graphs, as every induced subgraphs of a polar graph is also polar. Conversely, suppose that $G(A)$ and $G(R)$ are polar graphs. We show that G is a polar graph. Because G is trivially perfect it follows that $A \sim N \sim R$ and N is a clique. Suppose that $X \subset V$ still exists such that $G(X)$ is isomorphic to one of the following four graphs: H_1, H_2, H_3, H_4, where $H_i = G_i \cup F_i$ $(1 \leq i \leq 4)$, and every G_i is a P_3 and every F_i, described as a sequence of degrees, is: $F_1 : (4,3,3,3,3)$; $F_2 : (5,3,2,2,2,2)$; $F_3 : (4,4,3,3,3,3)$; $F_4 : (5,5,3,3,3,3)$. Because G is trivially perfect it follows that G is $\{P_4, C_4\}$-free. If x is the vertex of degree 4 in F_1, z and t are the vertices of degree 4 in F_3, u and v are the vertices of degree 5 in F_4 then $F_1 - \{x\}$ is isomorphic to C_4, $F_3 - \{z, t\}$ is isomorphic to C_4, $F_4 - \{u, v\}$ is isomorphic to C_4, which is a contradiction. We know that the complement of a polar graph is a polar graph and that $\overline{G}$ is C_4-free, because it is trivially perfect. If y is the vertex of degree 5 in F_2 and a is one of the vertices of degree 2 adjacent to the vertex of degree 3 in F_2 then $F_2 - \{a, y\}$ is isomorphic to C_4, which is a contradiction.

5 Recognition of polar trivially perfect graphs

Theorem 6 leads to the following recognition algorithm.

Input: $G = (V, E)$ a connected graph satisfying the conditions in Theorem 6
Output: An answer to the question: Is G a `Polar` graph ?

```
begin
    L = {G}  // L is a list of graphs
    while (L ≠ ∅)
        begin
          extract an element H from L
          find a weakly decomposition (A, N, R) for H
          if (A ⊀ N ⊀ R) then G is not trivially perfect
                         else  introduce in L the connected, non-complete components of
                               G(A), G(R)
```

end
Return: G is Polar
end

In what follows, we give some remarks on the algorithm.
Because the operation inside the body of while loop that takes the longest execution time is the weakly decomposition (namely $O(n+m)$) it follows that the total execution time of the algorithm is $O(n(n+m))$.

6 Combinatorial optimization algorithms for polar, trivially perfect graphs

In this section we calculate the domination number, give $O(n(n+m))$ algorithms to calculate the stability number and clique number.

Theorem 6 leads to the following result.

Corollary 3. *Let* $G = (V, E)$ *be a connected, non-complete graph and* (A, N, R) *a weakly decomposition with* $G(A)$ *the weakly component. If* G *is trivially perfect and also polar then the following hold:*
i) $\alpha(G) = \alpha(G(A)) + \alpha(G(R))$*;*
ii) $\omega(G) = |N| + \max\{\omega(G(A)), \omega(G(R))\}$*;*
iii) $\nu(G) = 1$.

Proof. Let $T \subset A \cup N$ such that T is stable and $|T| = \alpha(G(A \cup N))$. Because N is a clique it follows that $|T \cap N| \leq 1$. If $T \cap N = \emptyset$ then $T \cup \{r\}$ is a stable set in $G(A \cup R)$, and if $T \cap N = \{n_0\}$ then $(T - \{n_0\}) \cup \{r\}$ is a stable set in $A \cup R$, for every $r \in R$. It follows that in the relation in Corollary 2, the maximum is obtained only for the first component. So i) holds.
Because $A \sim N \sim R$ and N is a clique it follows that $\omega(G) = \omega(G(N)) + \max\{\omega(G(A)), \omega(G(R))\}$, but $\omega(G(N)) = |N|$. So ii) holds.
Because $A \sim N \sim R$ and N is a clique it follows that a domination set of minimum cardinal is a set determined by any vertex in N. So iii) holds.

Corollary 3 implies an algorithm for the construction of a stable set of maximum cardinal and of a clique of maximum cardinal in a trivially perfect, polar graph.

Input: $G = (V, E)$ a connected graph satisfying conditions in Corollary 3
Output: A stable set S with $|S| = \alpha(G)$ and a clique Q with $|Q| = \omega(G)$
begin
$S = \emptyset,\ Q = N$
$L = \{G\}$ // L is a list of graphs
while $(L \neq \emptyset)$
begin
extract an element H from L
if (H is complete) then
Return:
$S = S \cup \{v\},\ \forall v \in V(H)$
$Q = Q \cup N$
else
Determine a weakly decomposition (A, N, R) for H
Put $[A]_H$ and the connected component of $[R]_H$ in L
end
end

Facility location analysis deals with the problem of finding optimal locations for one or more facilities in a given environment ([11]). A type of problems in facility location analysis concerns the determination of a location that minimizes the maximum distance to any other location in the network.

The following centrality indices are defined in ([11]):
the eccentricity of a vertex u is $e_G(u) = max\{d(u,v)|v \in V\}$;
the radius is $r(G) = min\{e_G(u)|u \in V\}$;
the center of a graph G is $C(G) = \{u \in V|e_G(u) = r(G)\}$.

Using Theorem 6 we obtain the following result.

Corollary 4. *Let* $G = (V, E)$ *be a connected, non-complete graph and* (A, N, R) *a weakly decomposition with* $G(A)$ *the weakly component. If* G *is both trivially perfect and polar the following hold:*
(i) $e_G(u) = 2$ *for* $u \in A \cup R$*; (ii)* $e_G(u) = 1$ *for* $u \in N$*; (iii)* $r(G) = 1$*; (iv)* $C(G) = N$.

7 Conclusions and future work

Using the weakly decomposition we have obtained polynomial time recognition algorithms for polar graphs and directly calculated the domination number, while for the stability number and for density we give polynomial algorithms.

Future work concerns on other classes of graphs, characterized by forbidden subgraphs.

Bibliography

[1] C. Berge, *Graphs*, North-Holland, Amsterdam, 1985.

[2] A. A. Bertoss, *Dominating sets for split and bipartite graphs*, Inform. Process. Lett., 19:37-40, 1984.

[3] Z. A. Chernyak, A. A. Chernyak, *About recognizing* (α, ω)*-classes of polar graphs*, Discrete Mathematics, 62:133-138, 1986.

[4] D. G. Corneil, Y. Perl, *Clustering and domination in perfect graphs*, Discrete Appl. Math., 9:27-39, 1984.

[5] C. Croitoru, E. Olaru, M. Talmaciu, *Confidentially connected graphs*, Proceedings of the international conference "The risk in contemporany economy", in Annals of the University "Dunarea de Jos" of Galati, Suppliment to Tome XVIII (XXII), 2000.

[6] C. Croitoru, M. Talmaciu, *A new graph search algorithm and some applications*, presented at ROSYCS 2000, Univ. "Al.I.Cuza" Iaşi, 2000.

[7] T. Ekim, P. Heggernes, D. Meister, *Polar permutation graphs*, Report in Informatics, Report no 385, March 2009.

[8] T. Ekim, P. Hell, J. Stacho, D. de Werra, *Polarity of Chordal Graphs*, Discrete Applied Mathematics, Volume 156, 2469-2479, 2008.

[9] T. Ekim, N. V. R. Mahadev, D. de Werra, *Polar cographs*, Discrete Applied Mathematics, Volume 156, 1652-1660, 2008.

[10] M. C. Golumbic, *"Trivially perfect graphs"*, Discrete Math. 24:105-107, 1978.

[11] D. Koschutzki, K. Lehman, L. Peaters, S. Richter, Tenfelde-Padehl and O. Zlotowski, *"Centrality Indices"*, Lecture Notes in Computer Sciences, Springer Berlin / Heidelberg, Network Analysis, 16-61, 2005.

[12] V. V. Lozin, R. Mosca, *Polar Graphs and Maximal Independent Sets*, Rutcor Research Report 4-2004, February, 2004.

[13] M. Talmaciu, *Decomposition Problems in the Graph Theory with Applications in Combinatorial Optimization - Ph. D. Thesis*, University "Al. I. Cuza" Iasi, Romania, 2002.

[14] M. Talmaciu, E. Nechita, *Recognition Algorithm for diamond-free graphs*, Informatica, Vol. 18, No. 3, 457-462, 2007.

[15] M. Talmaciu, E. Nechita, *An algorithm for the bisection problem on trivially perfect graphs*, BICS 2008, Amer. Inst. Physics, Volume 1117:60-66, 2008.

[16] http://wwwteo.informatik.uni-rostock.de/isgci/classes/gc_314.html.

Int. J. of Computers, Communications & Control, ISSN 1841-9836, E-ISSN 1841-9844
Vol. V (2010), No. 5, pp. 946-952

Contributions to the Study of Semantic Interoperability in Multi-Agent Environments - An Ontology Based Approach

I.F. Toma

Iulian-Florin Toma
University of Pitesti
Romania, 110040 Pitesti, 1 Targu din Vale
E-mail: tif@tif.ro

Abstract: This paper details the results of our work in the field of multi-agent ontology-based environment simulation. We analyze the impact of introducing some techniques for the alignment / translation / mapping of agent ontologies, which allows for collaborative understanding of distributed ontologies. In the end, we analyze the difficulties / gaps that are to be filled for an actual deployment of the technology / concept in a real life environment.
Keywords: ontology, semantic interoperability, agents.

1 Introduction

The ability to communicate depends on understanding the syntax and the semantics of a language. We used an ontology model to facilitate semantic interoperability in a simulated multi-agent environment. Current studies relieve the fact that using ontologies associated with the agents allows for building semantic-aware distributed multi-agent systems. An ontology is a formal specification for shared understanding of a domain of interest and offers the possibility of building a formal and machine manipulable model of a domain of interest. Ontologies describe entities and relations between them, classes of objects and their attributes, and are complemented by logical rules that constrain the meaning assigned to the terms. These constraints are represented by inference rules that can be used by agents to perform the reasoning on which the autonomy and proactiveness of the agents are based. Thus semantic interoperability facilitates the increase of the autonomy of agents. The reasoning process of agents in multi-agent ontology-based environments is mainly focused on the alignment of their own ontologies with ontologies exposed by other agents. Having considered the multi-agent environment as heterogeneous, we analyse the introduction of guidelines for ontology development and evolution which should facilitate ontology reuse that may underpin a usage model for ontologies.

2 Multi-agent environment

The multi-agent environment we propose is part of an e-commerce scenario where a user's agent AGU is activated to find, request and gather offers for a specific product from a series of suppliers' agents AGS1, ..., AGSn. In our implementation, the suppliers' agents (AGS agents) are published in the FIPA compliant Directory Facilitator (DF) of the JADE platform. FIPA is an IEEE Computer Society standards organization that promotes agent-based technology and the interoperability of its standards with other technologies. JADE is a software framework to develop agent applications in compliance with the FIPA specifications for interoperable intelligent multi-agent systems. We propose the implementation of a Mediator Agent (AGM) to make the interactions with the DF and to handle the negotiation process between agents.

The AGS agents (suppliers' agents) use their own ontologies. Suppliers can adhere to centralized/standardized/publicly available ontologies, can extend such ontologies based on their needs,

but can also build their own ontologies from scratch. Facing this heterogeneity of ontologies used by the AGS agents, the AGU agent has to achieve semantic interoperability to negotiate with each AGS agent. FIPA recommends using an external Ontology Agent (AGO) for handling the interoperability problems in heterogeneous multi-agent environments [9]. This agent handles the tasks of matching ontologies and translating the values of some attributes (translate notions based on their specified language, convert currencies, convert units of measure).

Figure 1 illustrates the architecture of the proposed system:

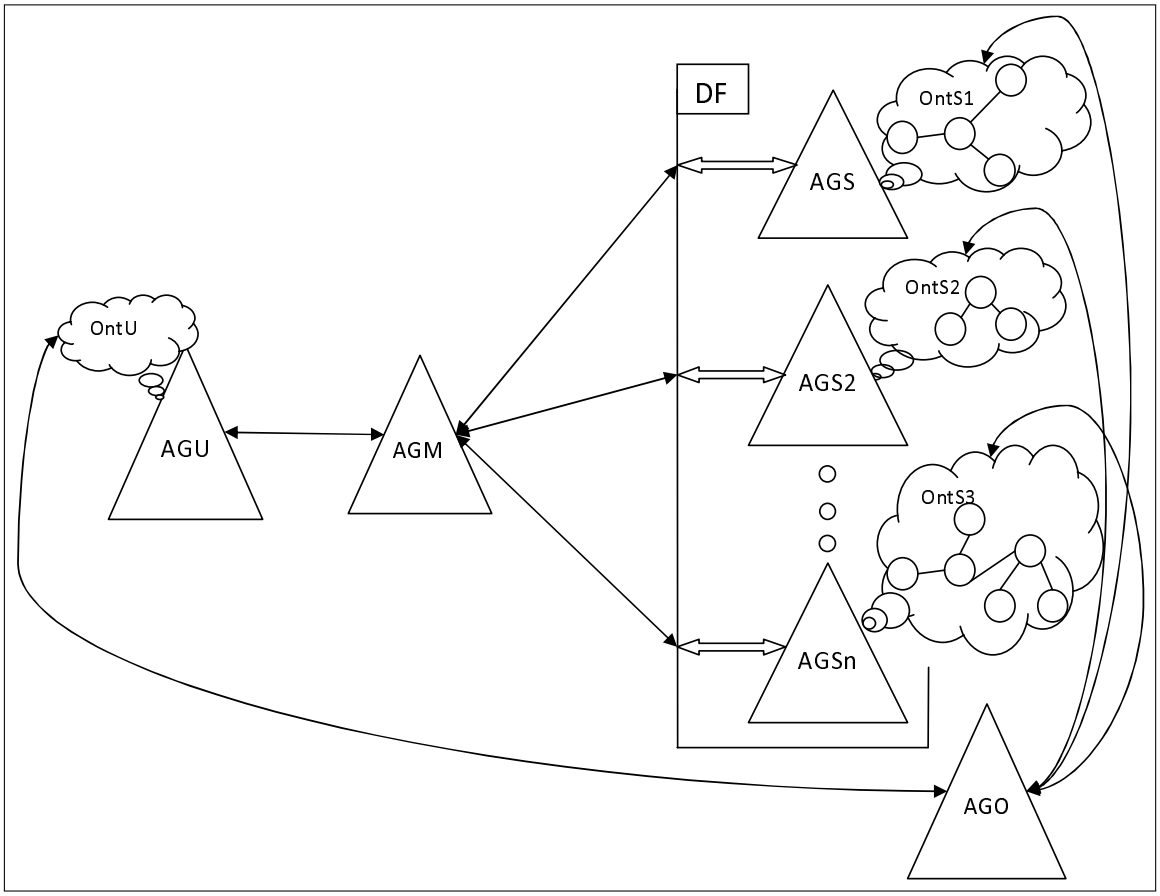


Figure 1: Ontology-based multi-agent environment

3 Ontologies

The semantical description of an entity offers a better understanding of the entity and puts the entity in a context. The idea is not necessarily to achieve the level of understanding and association to other entities that a human subject would have for that entity, but to improve the understanding of a software regarding that entity only to satisfy specific goals. The main heterogeneity problem in our simulated environment is that the agents will typically use different ontologies. Ontology-based interoperation provides a solution in environments with heterogeneous semantics. Ontologies can capture both the structure and semantics of information environments [7]. An ontology-based search agent like AGU can handle both simple keyword-based queries as well as complex queries on structured data. Figure 2 contains some sample ontologies exposed by 2 book suppliers' agents [1].

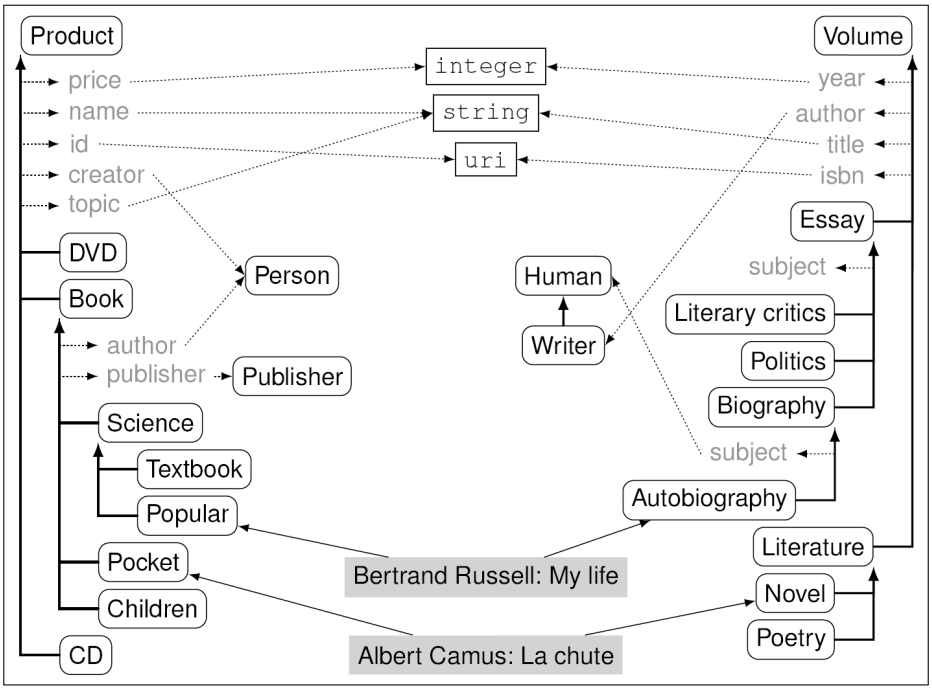


Figure 2: Sample ontologies [1]

The ontologies can be designed in any ontology editor, like the Protégé software [6] (Figure 3). A large variety of formal languages exists for describing the ontologies (e.g. OWL [2] [8]).

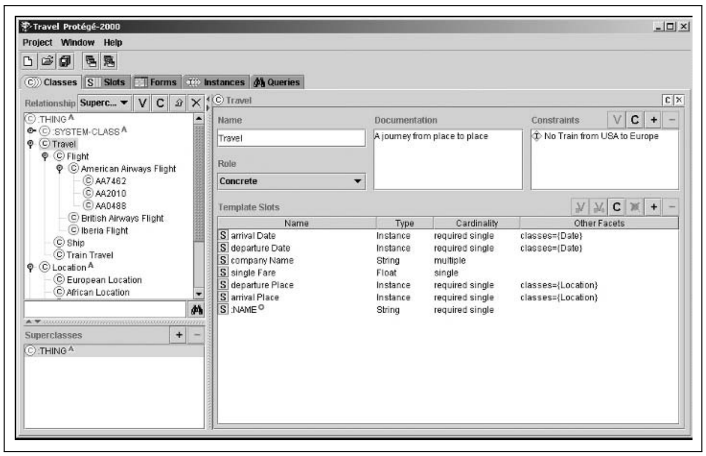


Figure 3: Ontology editing in Protégé 2000

4 Implementation

The test simulation can be implemented using the JADE environment. The JADE environment has direct support for ontologies only if they are represented as Java classes. These classes can be written by hand of generated using the BeanGenerator plug-in of the Protégé software. Although the Java classes can be generated automatically, this approach has the disadvantage that it is very rigid when conceptual changes are necessary. Under this approach, modifications of the ontology in Protégé have to be followed by the regeneration of the java classes using the

BeanGenerator tool, the re-integration in the main project and the recompiling. Any modification in an ontology represented using Java classes may also require modifications in the Java code of the agent. A solution to simplify this process is to access the Jena framework for writing ontology-based applications. This framework allows keeping the ontologies in the OWL format generated by Protégé, eliminating any further need to transform the ontology in another format. The parser of this framework has direct access to the OWL files and provides access to the classes of the ontology using Java language. Using this framework, the agent can access the instances of the classes, retrieve the value of attributes, determine subclasses etc.

5 Negotiation

The negotiation is implemented using the FIPA Contract Net Interaction Protocol [10]. The workflow for this protocol is represented in Figure 4.

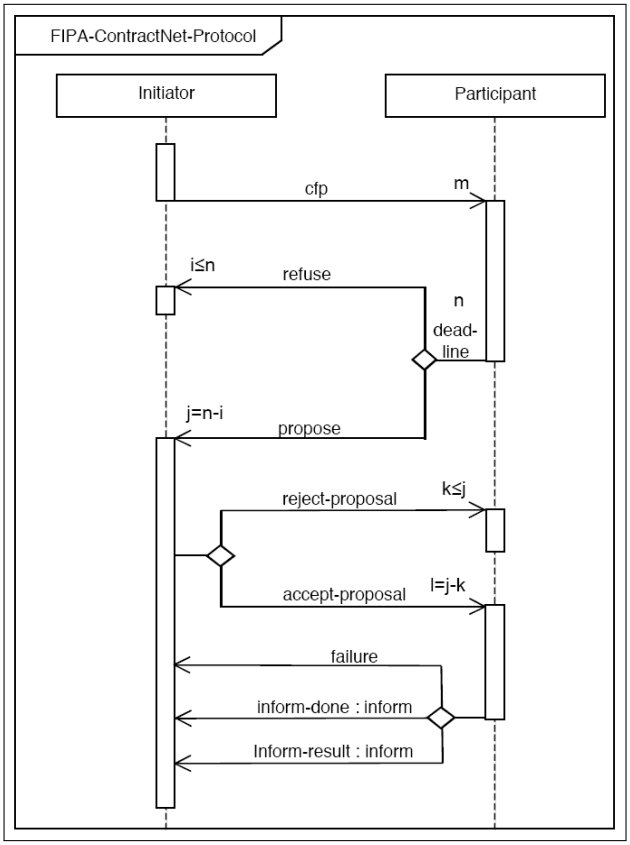


Figure 4: The workflow of thhe FIPA Contract Net Interaction Protocol

The semantic analysis of the ontologies takes place between the call for proposals (represented as "cfp") and the receival of the refusals/proposals from the suppliers' agents. The semantic analysis is performed by the dedicated agent AGO . The suppliers' agents AGS1...AGSn (noted: AGS) receive the CFP. If the concept/item is understood by AGS, it starts to evaluate de conditions/constraints contained in the CFP. If the conditions match, AGS returns a proposal to the AGU agent, otherwise it returns a REJECT-PROPOSAL message. The messages may be sent using predicates defined in a separate common / generalized / standard centralized ontology (OntReqSTD) specific to the process of requesting offers of products . Agents AGU and AGS could adhere to this ontology and AGS could then respond with a more descriptive message

like "AvailableOnlyWithPreOrder", "StockEmpty" rather than responding with a simple yes/no message. The types of ontologies proposed for our model are represented in Figure 5.

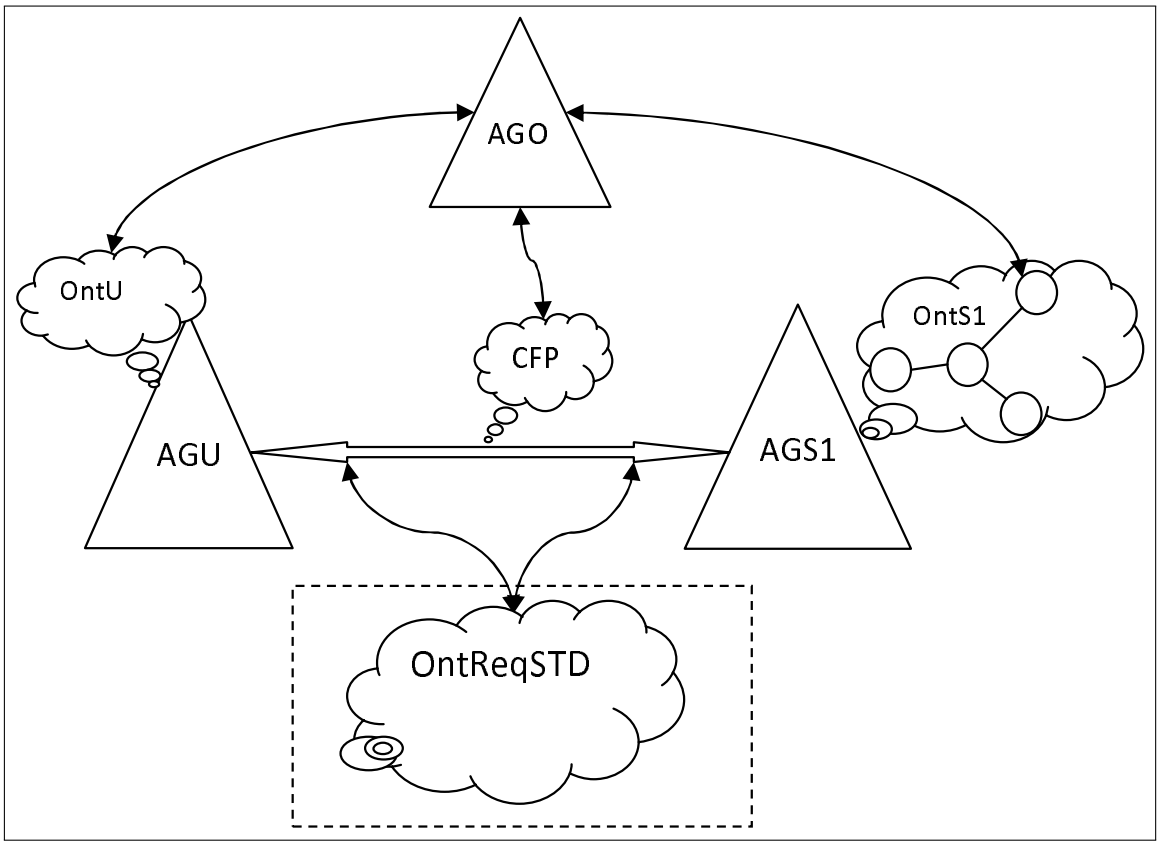


Figure 5: The role of the AGO (Ontology Agent)

Ontology matching analysis starts if the AGS does not "understand" the concepts/attributes inside the CFP. In this case, the AGS agents call the AGO agent. The AGO agent requests contact details to AGS to contact the AGU agent. Then AGO queries the AGU agent about its exposed ontology. AGO starts a matching process to obtain an alignment between the AGU request (cfp), AGU ontology (OntU) and AGS ontology. If the matching succeeds (with a high confidence mark), AGO reformulates the CFP to the AGS and the AGS may start the simple quantitative evaluation and send an answer to AGU accordingly to the result. AGO may also play the role of converting units of measure, currencies, translating notions between languages etc.) and use the values to reformulate the CFP.

6 Ontology mapping

The same concept/product may be described using different attributes, structures and relations, conducting to distinct ontologies. Even so, taking into consideration that the concept is part of a specific domain, there exists a set of common characteristics that should be mandatory when representing a specific product, like "width", "height", "material", etc. Thus lexical measures can be used to compare attributes and relations between concepts. Some of the lexical measures that can be applied to achieve ontology-matching are n-gram similarity, Hamming distance, Levenshtein distance, Jaro measure, Jaro-Winkler measure and the token-based distances like Cosine similarity, Term frequency-Inverse document frequency (TFIDF) described in [1].

The AGO agent can also make use of the language-based methods by calling external resources like Lexicons and Thesauri. WordNet [3] is such an electronic lexical database for English, based on the notion of synsets or sets of synonyms. A synset denotes a concept or a sense of a group of terms. WordNet also provides an hypernym (superconcept/subconcept) structure as well as other relations such as meronym (part of relations). It also provides textual descriptions of the concepts (gloss) containing definitions and examples. There are lots of techniques available for analyzing similarities using WordNet [4] [5].

In addition to comparing their names or identifiers, the structure of entities that can be found in ontologies can be compared. This comparison can be subdivided into a comparison of the internal structure of an entity, i.e., besides its name and annotations, its properties or, in the case of OWL ontologies, the properties which take their values in a datatype, or the comparison of the entity with other entities to which it is related.

Using these well-known techniques, the AGO agent can run a detailed analysis on the matching of the OntU and OntS ontologies. If an alignment is achieved, the AGO agent can reformulate the CFP (call for proposals) to match the OntS ontologies of the AGS agents.

7 Conclusions and the future of semantic interoperability

In this paper, we proposed a feasible implementation of a multi-agent environment which makes use of ontologies and ontology mapping to achieve semantic interoperability. The ontology-based multi-agent environments and semantic-enabled communications are, in the present (2009-2010), under theoretical study only. There are numerous attempts to bring this technology to public use e.g. building public libraries of ontologies, but this field still resides only in the boundaries of the scientifical community. To bring it to public use, tools should emerge, that would facilitate the semantic annotation and semantic description of entities, by any internet user, in various environments like web services, web pages, distributed software, reusable components etc. These tools should be easy to use by any internet user. They should incorporate automatic validation based on the recommendations and restrictions of the ontology engineering discipline. Only then, when the advantages of these technologies will overcome the difficulties of defining and maintaining the ontologies, will semantic-enabled communications be a part of common software and human activities.

Bibliography

[1] J. Euzenat, P. Shvaiko, *Ontology Matching*, Springer, 2007.

[2] *OWL 2 Web Ontology Language*, W3C Recommendation, http://www.w3.org/TR/owl2-overview/, October 27, 2009.

[3] G.A. Miller, WordNet: A lexical database for English, *Communication of ACM*, 38(11):39-41, 1995.

[4] Pedersen, Patwardhan, and Michelizzi, WordNet::Similarity - Measuring the Relatedness of Concepts, *Proceedings of the Nineteenth National Conference on Artificial Intelligence (AAAI-04)*, pp. 1024-1025, July 25-29, 2004, San Jose, CA (Intelligent Systems Demonstration)

[5] A. Budanitsky, G. Hirst, Semantic distance in WordNet: An experimental, application-oriented evaluation of five measures, *Proceedings of the Workshop on WordNet and Other*

Lexical Resources, Second Meeting of the North American Chapter of the Association for Computational Linguistics, Pittsburgh, USA, 2001.

[6] J. Gennari, M.A. Musen, R.W. Fergerson, W.E. Grosso, M. Crubezy, H. Eriksson, N.F. Noy, S.W. Tu, *The Evolution of Protege: an Environment for Knowledge-Based Systems Development*, Technical Report, SMI Report Number: SMI-2002-0943, 2002.

[7] M.P. Singh, M.N. Huhns, *Service-Oriented Computing: Semantics, Processes, Agents*, WILEY, 2005

[8] A. Gómez-Pérez, M. Fernández-López, O. Corcho , *Ontological engineering : with examples from the areas of knowledge management, e-commerce and the semantic web* , Springer, 2004

[9] *Ontology Service Specification*, FIPA-OSS, http://www.fipa.org/specs/fipa00086

[10] *FIPA Contract Net Interaction Protocol Specification*, FIPA TC Communication, SC00029H, http://www.fipa.org/specs/fipa00029/SC00029H.html.

Int. J. of Computers, Communications & Control, ISSN 1841-9836, E-ISSN 1841-9844
Vol. V (2010), No. 5, pp. 953-960

Using QSPS in Developing and Realization of a Production Line in Automotive Industry

N. Tudor, V.C. Kifor, C. Oprean

Nicolae Tudor
Continental Automotive Systems
Test Department
E-mail: nicolae.tudor@continental-corporation.com

Vasile Claudiu Kifor
"Lucian Blaga" University of Sibiu,
Director of the research department

Constantin Oprean
"Lucian Blaga" University of Sibiu
Rector

Abstract: Using the QSPS (Quality System for the production software) for industrial projects and not only therefore, has led to accurate running of the production line from beginning of the SOP (Start of Production). This paper presents the application way of the QSPS at one of the strongest European automotive company. By using of this system several significant costs savings and quality improvement can be observed.
The content of this paper will show step by step how to use QSPS for the integration of a production line in the traceability system from a big company in automotive industry.
The production line involved contains 56 production equipments, which have to be passed trough by the product before being packet and deliver to the customer.
The control of the line is done by this traceability system, so the impact of this system with the quality of the product is very high.
The structure of this system contains 7 steps. All of these steps are followed and executed in each System (test, pilot and production environment).
Keywords: quality improvement, control, savings, efficiency, capability.

1 Introduction

The traceability software for the production lines becomes more and more important and is a very significant process while running complex production lines with a high difficulty degree. The entire customer claims issues, statistics, the control of the process, CPK (process capability index) studies, FPY (First Pass Yield) and PPM (defective Parts Per Million) reports are done very simple, based on a strong traceability system.

Therefore a lot of techniques for implementing the traceability software were tested and checked, some of them successful and some less successful depending on the kind of the project. The paper below describes the practical usage of the QSPS system in real situation for the implementation of the traceability software of a complex line which contains 56 stations. The QSPS system, should improve the quality of work, cost and time saving at one side, but on the other side the system should be very simple and easy to be used in order to simplify the work activities and not to make it more complicated.

Also the system should cover and prevent all the risk which can appear in all 3 QSPS Phases (test, pilot and production) because of non conformity of improper software.

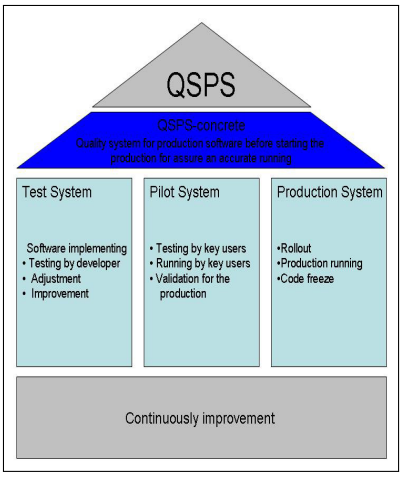


Figure 1: QSPS Framework

The next chapters describe step by step the implementation strategy of the traceability software, based on QSPS, of the already specified production line. All of the steps are described in the test (model) system, which have to be implemented and build up much closed (or even identical) to the characteristics of the production system. Those steps should be followed in all of the 3 environments systems of the QSPS (test, pilot and production).

2 Step I: Time and cost evaluation

This phase of the project should help to understand the deepness of the project, the difficulty level, and cost of the project, to make very clear and transparent the resources needed (time resources and headcounts) and to define a due date very closed to the reality.

Therefore the project should be split out in work packages, which can be better followed up, controlled and evaluated. The **Table 1** shows how the project was divided in work packages and subprojects.

Subproject	**Work packages**	**Difficulty level**
Process definition		2
Process description		2
Process review		2
Technical requirements specification		3
Target specification		3
Software implementation concept		3
Programming		3
Documentation of the software		2
Installation and configuration of the 56 modules		2
Testing the 56 Modules		2
Advising and user support		2
Project evaluation(the time for evaluation itself)		2
Internal software release	compliance to technical requirements	3
	compliance to the software ISO norms	2
	calibration and adjustments(bug solving)	3
	capability measurements	2
	usage work instruction	2
	user manual	2
	ergonomics and design	2
Customer software release	efficiency analyze	3
	defects analyze	3
	corrective actions	2
	analyze after corrective actions	2
	customer release reports	2
Validation and verification of the implemented software package	lifetime test	3
	test procedure for modification issues	2

Continued on next page

Table 1 – **continued from previous page**

Subproject	Work packages	Difficulty level
	approval procedure	2
	final releases	2
Functional follow up of the software during production	software audit	2
	statistical rate of the occurred errors	2
Customer claims	Problem recording	1
	Analyze and solving procedure	2
	Analyze report	1

Table 1 - Project segmentation

With this information the time evaluation can be done more accurate using the PERT algorithm. The results of the time evaluation can be observed in **Figure 2**.

Resource, time and costs evaluation

Project Name Traceability project of production line

Version Date 24.11.2009

Work Package (WP)	Number Employees	Rep.	Time[h] VO	VN	VP	VA	Total	Deviation	Dificulty Level	Costs (Euro)
Process definition	1,0		30	30	32	30,3	30,3	0,1	2	2821
Process description	1,0		16	20	20	19,3	19,3	0,4	2	1798
Process review	1,0		8	10	10	9,7	9,7	0,1	2	899
Technical requirements specification	1,0		80	90	90	88,3	88,3	2,8	3	14486,66667
Target specification	1,0		40	50	50	48,3	48,3	2,8	3	7926,666667
Software implementation concept	1,0		80	85	85	84,2	84,2	0,7	3	13803,33333
Programming	1,0		160	160	170	161,7	161,7	2,8	3	26513,33333
Documentation of the software	1,0		40	50	50	48,3	48,3	2,8	2	4495
Installation and configuration of the 56 modules	1,0		40	50	50	48,3	48,3	2,8	2	4495
Testing the 56 Modules	1,0		40	50	50	48,3	48,3	2,8	2	4495
Advising and user support	1,0		8	10	10	9,7	9,7	0,1	2	899
Project evaluation(the time for evaluation itself)	1,0		4	5	5	4,8	4,8	0,0	2	449,5
Internal software release	1,0					0,0	0,0	0,0		0
compliance to technical requirements	1,0		40	50	50	48,3	48,3	2,8	3	7926,666667
compliance to the software ISO norms	1,0		8	9	9	8,8	8,8	0,0	2	821,5
calibration and adjustments(bug solving)	1,0		16	18	18	17,7	17,7	0,1	3	2897,333333
capability measurements	1,0		8	10	10	9,7	9,7	0,1	2	899
usage work instruction	1,0		8	8	8	8,0	8,0	0,0	2	744
user manual	1,0		8	8	8	8,0	8,0	0,0	2	744
ergonomics and design	1,0		8	8	8	8,0	8,0	0,0	2	744
Customer software release	1,0					0,0	0,0	0,0		0
efficiency analyze	1,0		4	4	4	4,0	4,0	0,0	3	656
defects analyze	1,0		4	4	4	4,0	4,0	0,0	3	656
corrective actions	1,0		4	4	4	4,0	4,0	0,0	2	372
analyze after corrective actions	1,0		4	4	4	4,0	4,0	0,0	2	372
customer release reports	1,0		2	3	3	2,8	2,8	0,0	2	263,5
Validation and verification of the implemented software package	1,0					0,0	0,0	0,0		0
lifetime test	1,0		8	9	9	8,8	8,8	0,0	3	1448,666667
test procedure for modification issues	1,0		4	5	6	5,0	5,0	0,1	2	465
approval procedure	1,0		8	8	8	8,0	8,0	0,0	2	744
final releases	1,0		8	8	8	8,0	8,0	0,0	2	744
Functional follow up of the software during production	1,0					0,0	0,0	0,0		0
software audit	1,0		4	5	6	5,0	5,0	0,1	2	465
statistical rate of the occurred errors	1,0		4	6	6	5,7	5,7	0,1	2	527
Customer claims	1,0					0,0	0,0	0,0		0
Problem recording	1,0		0,5	0,5	0,5	0,5	0,5	0,0	1	30
Analyze and solving procedure	1,0		3	3	3	3,0	3,0	0,0	2	279
Analyze report	1,0		0,5	0,5	0,5	0,5	0,5	0,0	1	30
TOTAL			626	705	717	693,8	693,8	21,2		93549

Figure 2: Project time and cost evaluation

3 Step II: Internal software release

This step is the first evaluation of the software package from quality point of view. At the same time it is the first control tool used to signalize if the software was done in the right way, accordingly to the requirements, to the quality norms and not at least to assure the production of qualitative products. Therefore each of the steps below has to be verified.

3.1 Compliance to customer specification

This check has to be done based on the target specification. Each step from this chapter must be checked. For help a check table can be used and the characteristics from the target specification should be proved. For the line in our example following characteristics were verified:

- Program flow;
- Special cases in real life (like emergency stop, current interruption);
- Fail /pass /scrap situations;
- Handshake protocol between the industrial equipment and traceability software;
- Communication syntax;
- Repair scenarios;
- Limitation of the nr. of repair process;
- Check in/Check out process;
- Process parameter;
- Process results, CPK Measurements;
- Statistics of the process;
- Statistic of the failure;
- Product logistic on the line (process flow);
- Line flow control;
- Back flushing to the main database, in order to administrate material stocks;
- Label scanning;
- Label syntax;
- Packing. Check if the product pass all the processes in the line before being packed;
- Quality alert;
- Alert notification.

3.2 Check the implemented software package for software techniques and standards

For this point it was used the ISO/IEC 9899, for programming language C. Based on technical corrigenda 1:2001 the source code was reviewed. Also the next criteria were inspected at the very early beginning:

- Is the right software?
- The target is reached?
- Interface to the users is clear and unique while using it?
- At least one diagnostic message?
- The programming editor was used correctly? (Tabulator, empty spaces etc.)
- Requirements resulted from coded character set.
- Requirements resulted from binding techniques.
- Comments in the source code.
- Documentation.
- Usage documentation.

3.3 Software calibration and adjustments (bugs solving)

After the two chapters above were effectuated and all of the situations where verified, so all the bugs could be resolved, bugs which can occurred if the programmer doesn‘t consider a state from the program logic.

3.4 Capability measurements (using Pareto distribution, CPK, ISO 15504)

This measurement was done even in the test system. Production units can be created virtual for a good simulation. Knowing the handshake protocol between the traceability system and the

industrial equipment, a simulation can be done very closed to the reality situation. With these results the real running of the production can be compared. The upper, respectively the lower limits are in this case the cycle time of the process admitted by the customer while processing a production unit. CPK was done as described in **Table 2**, respectively the Pareto distribution for the analyses of the software package in case of rejected software, **Table 3** respectively **Figure 3**.

Unit Id	Measured time	Lower Limit 31	Upper Limit 36
1	33,5	30	36
2	33,2	30	36
3	34,1	30	36
4	33,9	30	36
5	33	30	36
6	34,3	30	36
7	33,2	30	36
8	32,9	30	36
9	33,8	30	36
10	33,6	30	36
11	33,7	30	36
12	33,6	30	36
13	32,9	30	36
14	33,1	30	36
15	33,8	30	36
16	34,3	30	36
17	33,9	30	36
18	33,6	30	36
19	33,2	30	36
20	33,3	30	36
21	32,7	30	36
22	33,5	30	36
23	33,8	30	36
24	34	30	36
25	33,6	30	36
26	33	30	36
27	33,8	30	36
28	32,9	30	36
29	33,4	30	36
30	33,9	30	36
Average	33,51666667		
Deviation	0,435560149		
CP	1,913245127		
CPK	1,90049016		

Table 2: CPK Calculation

3.5 Software ergonomic and designs

Therefore we based on the ISO 9241. This ISO Norm is a standard of Ergonomics of Human System Interaction. Following rules were proved:

- suitability for the task (the software interface should be suitable for the user‘s task and skill level);
- self-descriptiveness (the interface should make it clear what the user should do next);
- controllability (the user should be able to control the pace and sequence of the interaction);
- conformity with user expectations (it should be consistent);
- error tolerance (the software should be forgiving);
- suitability for individualization (the software should be able to be customized to suit the user) and
- suitability for learning (the software should support learning).

4 Step III: Customer software release (Run@Rate)

For this step we run the line with 100 production units (in the test system we simulate the running of the line with virtual units) and observe the behavior of the software, of the users

Software return reasons			
Problem	Count	Percent of total	Cumulative Percent
Not compatible	21	30,67	30,67
Does not perform as expected	18	18	50,33
Missing hardware resources	15	16	54
Not suitable for self learning	11	14	65,4
Missing user manual	10	7	71
Bad user interface	8	5,33	78
Too complicated	7	4	88,9
Bad cycle time	4	3	98,3
Bad Backup of recorded data	3	2	100
TOTAL	**97**	**100**	**100**

Table 3: Pareto analyze

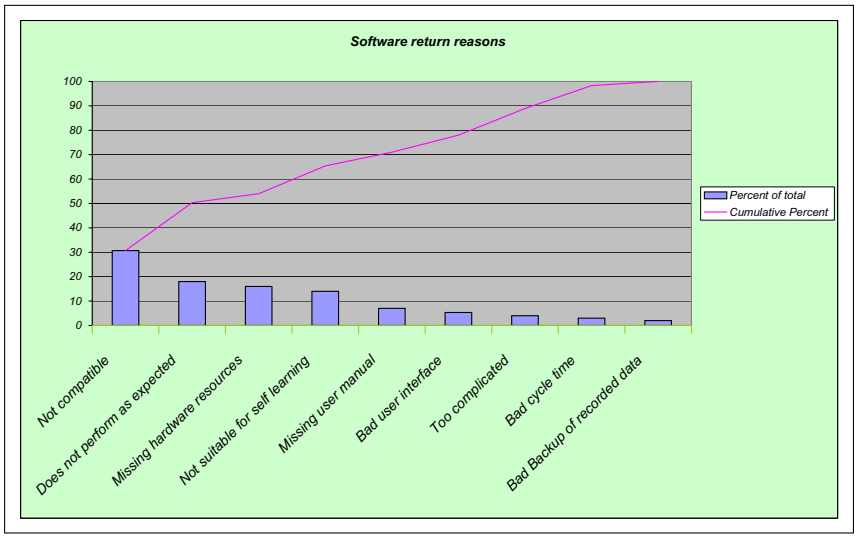


Figure 3: Pareto distribution

(operators) while using it in different situation and intentionally we cause extreme situation (like emergency stop of the equipment, current interruption... etc.) for a better software rating. A statistic is made for each equipment, inclusively the packing station. This statistic contains the number of the parts being processed on each station during one hour.

Also a pareto was realized similar to the **Figure 3**. The following issues were also effectuated:

- Efficiency analyze (how many parts per hour per equipment);
- Defects (bugs) analyze;
- Causes and corrective actions;
- Reevaluation procedure after corrective actions;
- Creating reports (like 8D report).

5 Step IV: Validation and verification of the implemented software package

After release and after the customer agreed the implemented software, the final release is made by the project manager. That means that the software is officially registered as a production software. At this level the project manager analyze the robustness of the software package by making some research about the following properties:

- Lifetime (how long the software will run without any modifications or maintenance), ISO 12207

  - Acquisition analyze (how was the software required);
  - Supply analyze (how was the software provided);
  - How it was develop;
  - How is the operating mode;
  - How much maintenance is needed.
- Modifications handling (how will be the modification make and tested);
- Documents like PPAP (product part approval process) and PSW (part submission warranty) are made and signed from the project manager.

6 Step V: Functional follow up of the software during the production

After the software was accepted and installed in the production line, periodical (every 2 weeks) software audits were made to assure that the software is running in conformity with the requirements and the production criteria are fulfilled.

In this software audit the main criteria (see chap. 3.1) are checked again.

Statistics are made for the cycle time, and errors occurred, in order to make the difference between software errors and process error.

The main aim of this step is to find improvements for the software in order to make it more robust and efficient.

7 Step VI: Customer claims

Customer claim escalation plan is also created like in the Figure 4 in order to obtain the best reaction time in case of software error which could appear, and disturb the normal running of the production. This will drive to very small solving and reaction times for eventually downtimes on the line because of software problems.

8 Conclusions

Using the QSPS system, the downtime of the line (containing the 56 production equipments) because of production software is decreased to almost 0%. The target of 0% is only possible when the technical requirements from the customer are done accurate and to 100% complete. Even with incomplete requirements, using the QSPS system the possibility to reach the target of 0% is possible. Therefore the implementation and running of the software should have a much longer period of time running in the pilot environment, to assure a safe lunch in the production. In most all of the cases this is not possible because the production should always start even if some problems are unsolved, because of time pressure. QSPS should eliminate that inconveniences, provoked by time pressure or other factors, due to a much easier structure, better and faster to be used, compared to other systems met till know in other studies.

The advantage of using the QSPS is that during the 3 environments (test, pilot and production), using the rules and ISO norms for each of them, the almost or all troubles which occurred are filtered before running the real production. That means that after the SOP (Start of Production) the interruption of the production line because of the software is almost inexistent.

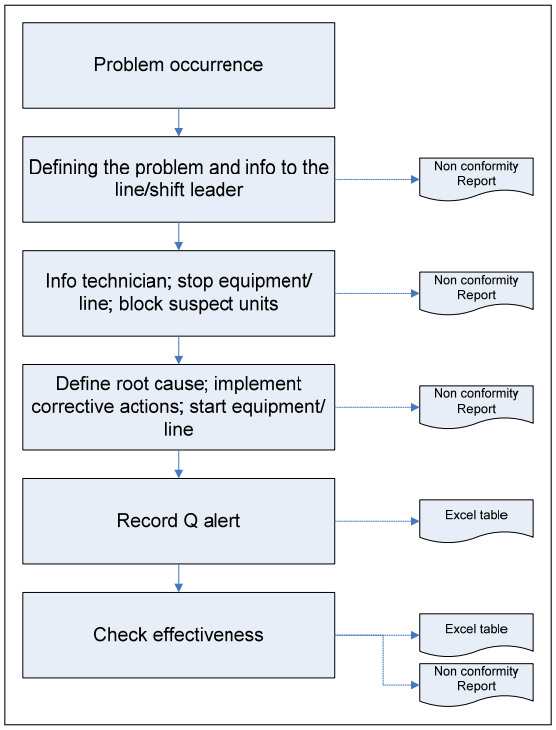


Figure 4: Escalation Plan

Bibliography

[1] Kai-Yuan Cai, Bey-Bey Yin, *Software execution processes as an evolving complex network*, Information Science, April 2008.

[2] Georg Kühner, Torsten Bluhm, *Employing industrial standards in software engineering for W7X*, Fusion Engineering and Design, 2009.

[3] J.-W. Li, *Modeling a quality assurance information system for product development projects with the UML approach*, Vol. 20, Nr.4 of International Journal of Computer Integrated Manufacturing, June 2007.

[4] N. Tudor, C. Kifor and C. Oprean, *Quality System for Production Software - QSPS*.

[5] Oprean, C., Kifor C. V., Suciu, O., *Managementul integrat al calităţii*, Sibiu, Editura Universităţii Lucian Blaga din Sibiu, 2005, ISBN 973-739-034-2.

[6] N. Tudor, D. Dumitrascu, *The Benefits of Project Structuring in Sub-projects and Work Packages*, http://imtuoradea.ro/auo.fmte/files-2008/MIE_files/TUDOR%20NICOLAE%202.pdf.

[7] N. Tudor, D. Dumitrascu, *Advance Estimate Expenses for Project Execution Time*, http://imtuoradea.ro/auo.fmte/files-2008/MIE_files/TUDOR%20NICOLAE%201.pdf.

Int. J. of Computers, Communications & Control, ISSN 1841-9836, E-ISSN 1841-9844
Vol. V (2010), No. 5, pp. 961-967

Secure Data Retention of Call Detail Records

F. Vancea, C. Vancea, D. Popescu, D. Zmaranda, G. Gabor

Florin Vancea, Codruţa Vancea, Daniela Elena Popescu
Doina Zmaranda, Gianina Gabor
University of Oradea, Romania
410087 Oradea, St. Universităţii 1
E-mail: {fvancea,cvancea,depopescu,zdoina,gianina}@uoradea.ro

Abstract: In today's world communication is relying heavily on electronic means, both for voice and other native data. All these communication sessions leave behind journaling information by the very nature of the underlying services. This information is both sensitive with respect to user's rights and important for law enforcement purposes, so proper storage and retrieval techniques have to be taken into consideration. The paper discusses such techniques in relation with recent EU recommendations and suggests some methods for achieving good performance while preserving the required security levels.
Keywords: call detail records, secure retention, row chaining.

1 Introduction

During the last years we have seen an increasing interest in secure storage and retrieval of journaling archives of various transactions. The possibility of fraud or inappropriate usage of resources have led to laws and regulations regarding storage of transaction history for phone calls, SMS, e-mail traffic and other forms of communication. Specifically, the European Parliament and the Council of the European Union have adopted a directive (2006/24/EC) [1] which gives guidelines for retention of data collected from publicly available communication services or public communication networks, in order to ensure that the data is available for investigation, detection and prosecution of serious crime.

Usually, the devices that perform the actual communication function are generating logging data that can identify the peers, time attributes and other relevant details of the transaction. These logs have been traditionally used for billing and eventually for statistical purposes by the network operators themselves. Since this information was stored anyway by the operators, the law enforcement agencies could make use of them for specific purposes within a frame more or less regulated by each country laws. Currently, many states are requiring the voice and data operators to store these logs for a well-defined period of time and to present them timely when legally requested by various law-enforcement entities. In order for this information to be useful for investigation and furthermore for prosecution, there is an implicit requirement that the information contained by the logs cannot be tampered with.

Even if the operators should not (or actually may not) capture and store the actual content of the communication, there are also increasing concerns about the privacy of the service users related to their communication partners. Wholesale collection of user communication actions and habits can be viewed as a breach of privacy, so strong protection against casual browsing or unauthorized intentional usage of this information should be provided.

When combining the long-term storage and efficient retrieval requirements with privacy, non-repudiation, integrity, performance and cost requirements, the task of data retention becomes non-trivial.

2 System structure, main requirements and challenges

The communication network operator has to manage during normal operation several pieces of information about each communication session. At the bare minimum this information is used to establish communication channels and to ensure correct traffic flow but it may also be stored and used for billing purposes. We will call such information about one user session Call Detail Record (CDR) as this is the name commonly used by voice service providers. For simplicity we will also name the communication session call but all reasoning below applies also for data sessions (e-mail, Internet sessions).

The records originate in the network equipment (central switch, access infrastructure), which have limited ability to store data, so periodically the records are transferred into database systems where they are processed (e.g. for billing) and finally discarded. If CDR retention is desired, the records are also stored for a rather extended period of time in some sort of semi-persistent storage until lawfully required through a query or until the retention period expires.

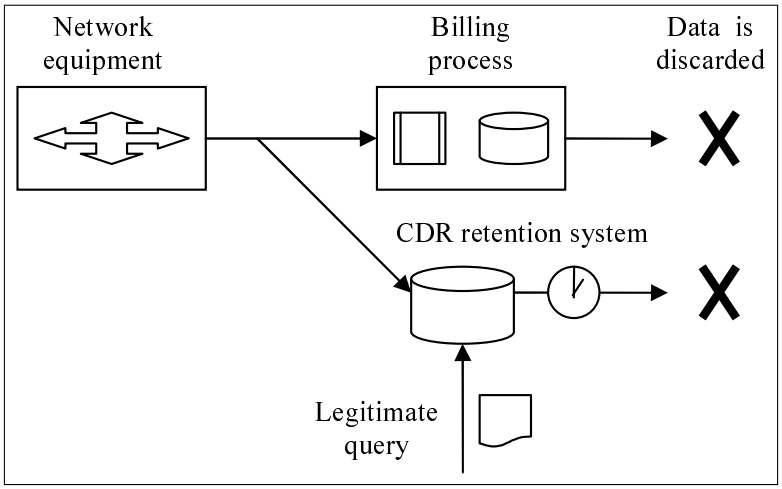


Figure 1: CDR data flow and data retention

For the purpose of this discussion we will not consider the billing path because this is part of the existing business model of the operator and is already well-established and less likely to accept any change. User data may be exposed there but this is an old threat, which is supposedly already handled by operator's procedures. Furthermore, the lifecycle of user sensitive data there is relatively short (one month or at most a couple of months) and the impact of integrity loss is only eventually financial to the operator but not deeply legal to the user.

The CDR data can be attacked:

- Before entering the retention system;
- While stored in the retention system;
- During query;
- After supposed deletion.

We will consider that the attacks performed before entry in the retention system have a low risk because the data stays in transit between network equipment and storage for a short period of time, the path is internal to the operator (thus is reasonably secure) and more important, the attacks are likely to take place only after a reasonable long time after the call is placed (otherwise the attacker would have refrained from making the call at all or would have attacked directly the network infrastructure to obtain an untraceable call).

The attacks on CDR data during the query may be:

- Impersonating authorized entities;
- Query alteration;
- Result alteration;
- Result disclosure.

All those can be stopped by proper signing and encryption of the request and the result.

The attacks after supposed deletion are disclosure attacks and would attempt to extract protected information from media used to store CDR data in the retention system. Encrypting the data is of course the main protection method but ciphertext-extensive attacks may exist which exploit the very large amount of encrypted data. To avoid these attacks write-once media should be destroyed properly and rewritable media should be properly erased after retention time expires.

We will focus on data stored in the retention system, because the CDR data stays there a rather long time (at least 6 months, according to EU recommendation) and the attacks may be either against the confidentiality or against the integrity of the data.

The CDR retention system has to take steps to prevent at least the following attack types:

- unauthorized disclosure of call detail, either for a particular call, a particular originator or a particular termination
- alteration of call detail (originator, termination, start/end time, other details)
- complete removal of one call or of all calls matching a particular originator, termination or time interval
- complete denial of service against the retention system

When evaluation the potential solutions we should also consider the size of the problem. One reasonably large operator may easily originate or terminate in excess of 10M calls per day (probably at least one order of magnitude more in the case of data services). At a minimum, a CDR will contain the originating subscriber number or identifier (at least 10 characters), termination or forwarding subscriber number (at least 10 characters), starting timestamp and duration (about 6 characters). To cover all situations (international dial numbers) additional space has to be reserved. In the case of mobile services additional location data is required and data services may have longer identification tokens. We estimate therefore the raw CDR to require at least 32 bytes of storage, leading to a minimum figure of 320MB of raw data to be stored for each day of traffic. This extends to about 10GB monthly and 120 GB yearly.

The above figures may appear small by today's standards in memory and storage capacity, but they are absolute minimum estimates of raw data. In a working system there should be protection which is obviously introduced by encryption and a working system should offer some sort of direct random access to the data, preferably optimized like an indexed database query. Both features have storage overhead and we consider it to be tenfold from extrapolating previous experiences. In the end the retention system will probably have to deal securely with about 500GB of data for a 6 month retention period on the assumption of 10M calls per day.

The above estimation may still appear small by comparison to a large business platform, but we should keep in mind that the CDR retention system is a non-profitable investment for the operator so one cannot expect it to use significant resources in terms of purchased hardware or software. Fortunately, the criteria for common queries are limited: by originator identifier (directly or indirectly mapped through user identity), by termination identifier and by time interval.

3 Potential solutions

According to the problem description we should figure how to store the data and how to efficiently retrieve the required CDR according to the typical query criteria.

3.1 Direct WORM storage

One obvious approach is to use WORM media, given that the write-once feature will provide tampering protection. This approach will not automatically solve the confidentiality issue, so some additional measures have to be taken. The main disadvantage of passive WORM media is capacity. Considering the above estimations, plain passive WORM media is not sufficient unless the time span for data written on one support is short. The time span should also be as short as possible, because data is protected only after write, and data in transit will be vulnerable to integrity attacks. Short time spans on one support will translate into many media units for the entire retention period, making the query process difficult and time-consuming. The speed of passive WORM media is also limited, affecting the query performance.

A better performing approach is to use special active WORM devices. These are special storage units presenting the WORM feature at the access port (Ethernet, SCSI) but backed by conventional magnetic storage. The WORM feature is protected by dedicated interface and firmware and the whole device is tamper-proof. This kind of device offers good protection but is usually expensive and its main use case is append to log. Our goal is to map database-like search capabilities over the protected data and the append to log restricted primitive is not friendly with common database engines and their storage management.

The lesser version of active WORM is a hybrid software system that achieves the write once feature by modification of the operating system block drivers or by a custom operating system [2]. The protection it provides may be reasonable but far weaker than the one offered by true passive optical media. This hybrid alternative fails to be database-friendly, too.

All the solutions based mainly on WORM devices will amount to periodically storing the CDR logs and sequentially scanning those for each query. This is not very efficient; especially considering that all the data that is sequentially scanned will be encrypted for confidentiality purposes.

3.2 Database backed by WORM storage

A slightly better solution would be to use a true database for query purposes and a passive WORM reference for result validation.

The CDR data would both be stored in the database and consolidated with a reasonable short interval to WORM storage (passive media should be OK). The initial query would be performed over the regular database and the integrity of the database itself might be checked against some hash checks stored on WORM media. The solution is fragile and may not provide either the required level of integrity or the desired level of performance.

Some DBMS (e.g. Oracle) offer the possibility to split the storage segments over read-write and read-only media. In such a simple scenario, the retention time window would be split in several fragments out of which only the last one is residing on read-write storage. Of course, management of splitting would be rather complicated and there are certain limitations in place which would limit the reasonable unsafe fragment to at least one week. The solution has the disadvantage of poor query performance for time frames located on read-only media, but the main disadvantage is that the unsafe fragment cannot be protected even if conventional row-level encryption is applied. Little can be done to actually prevent the rows in the read-write partition from being maliciously created, changed or deleted. Security mechanisms at database

level do exist but we are looking for an additional layer of trust, which would rule out any database authority from tampering the data.

By using encryption at row level one can simply prevent row changing and unauthorized row creation [3]. Efficient query is possible by index-scan over plaintext time attributes or by index-scan over pre-encrypted source or target fields. Unfortunately row deletion can be always performed at some authority level without being even detected.

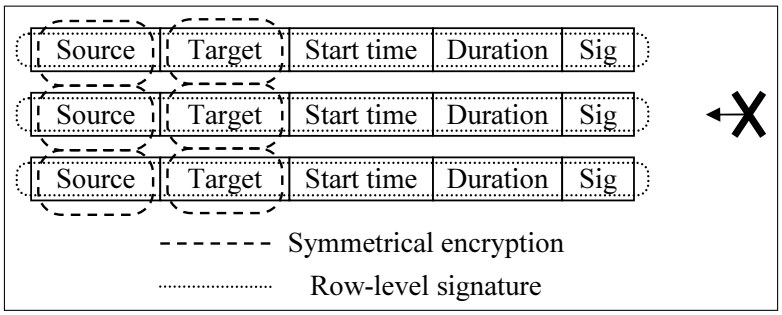


Figure 2: Row content is protected but rows can be deleted

3.3 Database with cross-linked rows backed by WORM storage

An improved solution is using results of our previous work, namely row chaining [4] [5]. Beside symmetrical encryption for confidentiality protection and row-level signature for row integrity and authenticity protection we use a chaining scheme that links one row to the previous one. The scheme can detect row deletion unless a whole block of rows from the end are collectively deleted. This would be solved by building the chain two ways, at the cost of revisiting previous records. Given the volume and the dynamic of data we consider a single chain is providing enough protection.

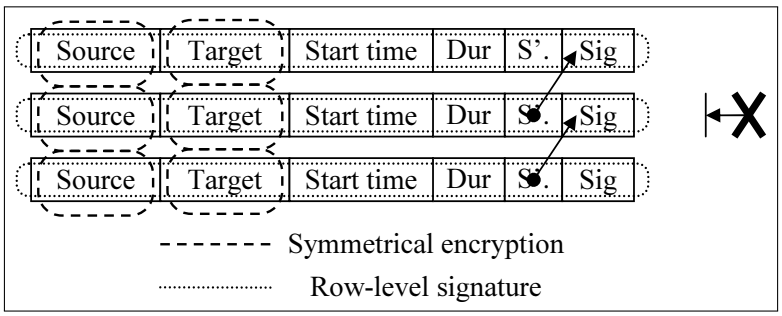


Figure 3: Rows protected by chaining cannot be deleted

Using the chaining scheme has the great advantage of relieving us from calling the WORM storage upon each query. The CDR data can now reside entirely in plain read-write database storage as its integrity is now protected by chains and signatures. When a query is presented, the results are quickly determined at regular database speed, and then the result-set is validated by walking the chain one or two steps for each record. The WORM storage is still required in case tampering is detected in any form and as a backup to avoid denial-of-service attacks (complete corruption of the database).

The above scheme has however a flaw and can be improved. One can maliciously alter the database indexes to skip certain records, and those records will never be returned in the result

set even if they are present in the database and the storage structure is still cryptographically intact. All checks will succeed, even for queries returning rows adjacent to the masked ones.

The solution is to build the chains not on subsequent rows as they are added to the database but on rows belonging to the same source or the same target. This means that the database will contain a chain of rows for each source and a chain of rows for each target. There is a new efficiency issue here, because we will need two chaining fields, one for source and one for target. The issue of protecting the last record in the chain becomes also apparently more sensitive, because the open ends of the chains are living longer. However, undetected deletion of a row requires that it is at the open end of both source and target chains, which is unlikely to happen for long in regular traffic.

4 Aggregation attack and countermeasures

So far we have considered that encrypting the source and the target fields offers appropriate protection against disclosure of the actual participants to a call. Unfortunately, in order to benefit from indexed search a particular source or target identifier has to be encrypted always with the same key, yielding the same encrypted value. The key does not have to be the same for all encryptions (actually should be changed over the source-target population) but once an identifier is encrypted in the system it should use the same ciphertext representation.

This invariance property required for indexing opens a path for aggregation attacks. Say the attacker has access to the database (either online or in archived form) and is looking for source A, but does not know the key. Actually it does not need to. If the attacker knows that A has placed calls at some particular moments in time it can aggregate through the data looking for that particular call pattern. Once it finds it all the past and future CDR involving A (at least as source) are accessible. The required number of such fixed points is not very high, as it was previously proven in aggregation attacks on blinded statistical data. The success of the attack depends on the traffic volume around the known time-points and on the number of the time-points available. The attacker may even optimize its chances by explicitly placing calls to A at low-traffic time intervals or if the attacker is itself A, by making calls at such well-chosen moments.

To provide some level of protection against this kind of attack we have to dilute the resolution the attacker has in the time domain when building a candidate set based on a given timestamp. When the attacker has a timestamp it can effectively build a set of sources or destinations that have placed or received a call at that moment in time. If the timestamp is say 11:00:00 AM, the set will be likely large (high traffic at busy hours) but if the timestamp is 03:00:00 AM the set will be considerably smaller. The time resolution dilution has to be made by sacrificing performance at some level. We will store in the Start time field a down-rounded timestamp and in an additional encrypted field the offset of the actual start time. This will reduce the resolution of the time searches and will entail some sequential processing to achieve exact query results. However, since the query specifies typically also the source or the target the performance penalty will be small (will use source or target indexes). On the other hand, the attacker has only the degraded start time to work with. To further refine the protection, the rounding should be made according to traffic patterns. Lower traffic intervals should be larger and higher traffic intervals may be smaller. This helps reducing the performance degradation we mentioned before and can be achieved by a Hamming-like algorithm applied once in a while to determine the optimal rounding intervals.

Even when using the rounding method to protect against data aggregation, special care should be taken with respect to the data storage pattern. All records have to be added incrementally with the starting timestamp (or eventually the end timestamp) because that is the way they come

from the network device logs. If the rounding algorithm performs for example the rounding at 23:00, 01:00, 05:00 the attacker may place a call shortly after 01:00 AM and look for records after but in close proximity of the Start time switch from 23:00 to 01:00. The solution is to randomly shift the rounding boundaries each day with at least 10% of the theoretical interval size.

5 Conclusions and Future Works

Designing and implementing a system which provides secure data retention of Call Detail Records is not a trivial task. There are several inter-related issues ranging from storage methods to data protection by encryption, and from efficient query to safe implementation. Unfortunately we have to note that in the current form, secure data retention has little driving force. The operators that manage the data have little to gain by properly implementing such a system and proper implementation is not necessarily cheap. The laws mandating the presence of such systems are rather vague on the required security and performance levels for such systems. The potential direct users have no direct means to encourage proper (read fast) implementations and the real owners of the call data (the subscribers) have also little means to encourage proper (read secure) implementations.

We however tried to outline within the limited size of this paper some of the requirements and the challenges related to such a system together with means to achieve a safe system with good performance.

Of course, each link in the chain has its very important contribution to the overall strength and there are many details left uncovered. Some directions may be: the proper protocol to be implemented between the retention system and the authorized partner requesting a query, proper and efficient protocol for destruction of CDR data at the end of retention period, implementation details on the chaining and on the time dilution algorithm, general implementation notes regarding secure split-role administration. Some of our future work may be related to these directions.

Bibliography

[1] ***, Directive 2006/24/EC of the European Parliament and of the Council Of 15 March 2006, *Official Journal of the European Union*, L105/54, 13.04.2006

[2] Y.Wang, Y.Zheng, Fast and secure WORM storage systems, *Proceedings of the IEEE Security in Storage Workshop (SISW)*, pages 11-19, 2003

[3] B. Schneier, Applied Cryptography, *John Wiley & Sons*, 1996.

[4] F. Vancea, C. Vancea, F. Vancea, C. Vancea, Practical Security Issues for a Real Case Application, *Int. J. of Computers, Communication and Control 3(S)*, pages 11-16, 2008

[5] F. Vancea, C. Vancea, Protecting data integrity with chained rows and public key cryptography. Comments on a real case, *Proceedings of RSEE 2008*

Int. J. of Computers, Communications & Control, ISSN 1841-9836, E-ISSN 1841-9844
Vol. V (2010), No. 5, pp. 968-975

Robust 2-DoF PID control for Congestion control of TCP/IP Networks

R. Vilanova, V. M. Alfaro

Ramón Vilanova
Department de Telecomunicació i Enginyeria de Sistemes
Universitat Autònoma de Barcelona
08193, Bellaterra, Spain,
E-mail: ramon.vilanova@uab.cat

Víctor M. Alfaro
Escuela de Ingeniería Eléctrica
Universidad de Costa Rica
San José, 11501-2060, Costa Rica.
E-mail: victor.alfaro@ucr.ac.cr

Abstract: This paper presents how Robust PID control can improve the performance of congesion control on TCP/IP networks. The proposed approach is compared with other control methods, such as PI control or RED/AQM, showing the advantages of the proposed technique.
Keywords: Congestion control, Active Queue Management, PID Control.

1 Introduction

Internet congestion control and congestion avoidance have been active research interests in the area of networking (see, for example [1] [2] [3]) during the last two decades. It has two components: (1) the end-to-end congestion control protocol, such as TCP [4], and (2) an active queue management (AQM) scheme implemented in routers. AQM signals congestion by discarding or marking packets. When congestion is detected by TCP, it will take actions to reduce the source sending rate. Normally, AQM objectives are: to stabilize the buffer queue length at a given target, thereby achieving predictable queueing delay, and to minimize the occurrences of queue overflow and underflow, thus reducing packet loss and maximizing link utilization. Thus, it is necessary to reduce as much as possible this problem. At present, there are methodologies to deal with this issue [5]: *congestion control* which is used after the network is overloaded and *congestion avoidance* which takes action before the problem appears. This paper deals with congestion control because it is where feedback control techniques can be openly and easily applied.

Recently several mathematical models of active queue management (AQM) schemes supporting transmission control protocol (TCP) flows in communication networks have been proposed [2] [3]. From these models a control theory-based approach can be used to analyze or to design AQM schemes. The more well known AQM scheme is probably RED [1]. RED can detect and respond to long-term traffic patterns, but it cannot detect congestion caused by short-term traffic load changes. In addition, it is well known that an appropriate tuning of RED parameters is not an easy task an may result in a non stabilizing controls scheme. This fact has motivated the research for alternative control approaches.

This paper presents the application of a Robust PID approach as an AQM controller. The controller can be easily tuned from the network parameters and it is compared with RED and the PI controller proposed in [3]. The performance under different load conditions shows the robustness and superiority of the presented approach. In addition the simple formulation of the PID controller also constitutes a motivation for implementation.

The rest of the paper is organized as follows. Next section presents the nonlinear model for a TCP router as well as the TCP control problem formulation. Section 3 reviews the RED and PI controller approaches for AQM control whereas section 4 presents the Robust 2-DoF PID approach. In section 5 a discussion and comparison is conducted. The paper ends with drawing some conclusions on the reported results.

2 AQM Router Dynamic Model and Control problem statement

In this section the dynamic nonlinear/linearized equations of TCP behavior developed in [3] are briefly reviewed as well as the purposes of AQM control stated.

2.1 Dynamic TCP Model

As in the literature, a network configuration consisting of a single congested router with a transmission capacity C is considered in this paper. TCP timeout mechanisms have been ignored for simplification. Using fluid-flow and stochastic differential equation analysis, the following coupled, nonlinear differential equations have been proposed as the dynamic model of the TCP behavior:

$$\begin{aligned} \dot{W}(t) &= \frac{1}{R(t)} - \frac{W(t)}{2}\frac{W(t-R(t-R(t)))}{R(t-R(t))}p(t-R(t)) \\ \dot{q}(t) &= \begin{cases} -C + \frac{N(t)}{R(t)}W(t) & q(t) > 0 \\ \max 0, -C + \frac{N(t)}{R(t)}W(t) & q(t) = 0 \end{cases} \end{aligned} \tag{2.1}$$

where

W	$\doteq$	average TCP window size (packets)
q	$\doteq$	average queue length (packets)
$R(t)$	$\doteq$	round-trip time$=\frac{q(t)}{C} + T_p$ (secs)
C	$\doteq$	link capacity
T_p	$\doteq$	propagation delay (secs)
N	$\doteq$	Number of active TCP sessions
p	$\doteq$	probability of packet mark

The queue length q, and window-size W, are positive bounded quantities; i.e., $q \in [0, \bar{q}]$ and $W \in [0, \bar{W}]$ where q and W denote buffer capacity and maximum window size respectively. In this formulation, the congestion window size W(t) is increased by one every round-trip time if no congestion is detected, and is halved upon a congestion detection. Moreover it has been assumed that the AQM scheme implemented at the router marks packets using Explicit Congestion Notification (ECN) to inform the TCP sources of impending congestion.

To linearize (2.1), it is assumed that the number of active TCP sessions and the link capacity are time-invariant, i.e., $N(t) \equiv N$ and $C(t) \equiv C$. In addition the dependence of the time delay argument $t - R$ on queue length q, is ignored and it is assumed to be fixed to $t - R_o$. Then, local linearization of (2.1) about the operating point results in the following equation:

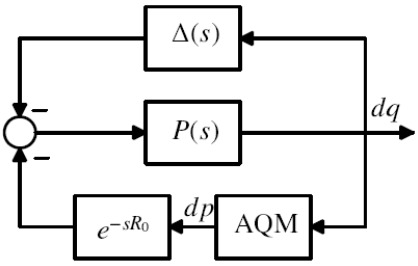


Figure 1: AQM Block Diagram of the linearized model along with High frequency uncertain dynamics

$$\begin{aligned} \dot{\delta W}(t) &= -\frac{N}{R_o^2 C}(\delta W(t) + \delta W(t - R_o)) \\ &- \frac{1}{R_o^2 C}(\delta q(t) - \delta q(t - R_o)) - \frac{R_o^2 C}{2N^2}\delta q(t - R_o) \\ \dot{\delta q}(t) &= \frac{N}{R_o}\delta W(t) - \frac{1}{R_o}\delta q(t) \end{aligned} \tag{2.2}$$

where $\delta W(t) = W(t) - W_o$ and $\delta q(t) = q(t) - q_o$ are the incremental variables with respect to an operating point. The operating point for a desired equilibrium queue length q_o is given by:

$$R_o = \frac{q_o}{C} + T_p \quad W_o = \frac{R_o C}{N} \quad p_o = \frac{2}{W_o^2} \tag{2.3}$$

This leads to a low order nominal model of the network dynamics that is accurate at a particular operating point $(R_o, q_o, C_o, W_o, p_o)$ given by:

$$P(s) = P_o(s)e^{-R_o s} = \frac{C^2/(2N)}{(s + \frac{2N}{R_o^2 C})(s + \frac{1}{R_o})} \tag{2.4}$$

By modelling the high frequency dynamics using a block $\Delta(s)$ such that

$$\Delta(s) = \frac{2N^2}{R_o^3 C^3} s(1 - e^{-R_o s}) \tag{2.5}$$

This term represents the high frequency, necessarily parasitic, network uncertainty in the model. Computational experience has shown that this can adequately capture certain deviations from nominal network performance. These considerations lead to the generation of a simplified feedback control system as shown in Fig. (1)

2.2 AQM Control problem

The function of an AQM control law is to mark packets (with probability p) as a function of measured queue length q. Marking of packets is consecutively used by the sender to throttle the amount of data sent; if no marked packets are received the window size is increased. Upon reception of a marked packet the window size is halved. The principal performance objectives for an AQM control law are:

1. Efficient queue utilization, to avoid overflow or emptiness of the queue buffer.

2. Regulated queueing delay, to minimize (optimize) the time required for a data packet to be serviced by the routing queue. The queueing delay is equal to q/C.

3. Robustness, to maintain closed-loop performance in spite of plant uncertainties, N, R_o and C.

3 RED and PI Approaches to AQM

this section reviews two of the more well established approaches for AQM control, say RED and the PI approach presented in [3].

3.1 RED approach to AQM

Random Early Detection (known as RED) was presented by [1]. A RED gateway calculates the average queue size, using a low-pass filter with an exponential weighted moving average. The average queue size is compared to two thresholds (minimum and maximum). When the average queue size is less than the minimum threshold, no packets are marked. When the average queue size is greater than the maximum threshold, every arriving packet is marked. If marked packets are in fact dropped, or if all source nodes are cooperative, this ensures that the average queue size does not significantly exceed the maximum threshold. When the average queue size is between the minimum and the maximum threshold, each arriving packet is marked with probability p, where p is a function of the measured queue length q. Hollot et al. in [3] proposed the following transfer function model for the RED controller:

$$C_{red}(s) = \frac{KL_{red}}{s+K} = \frac{K_{red}}{s/k_{red}+1} \tag{3.1}$$

where

$$L_{red} = \frac{p_{max}}{(max_{th} - min_{th})} \quad K_{red} = \frac{R_o^3 C^2}{(2N)^2} \quad k_{red} = -C\ln(1-\alpha_{red}) \tag{3.2}$$

and where α_{red} is REDs queue-averaging weight. The corresponding controller block diagram is shown in figure (2).

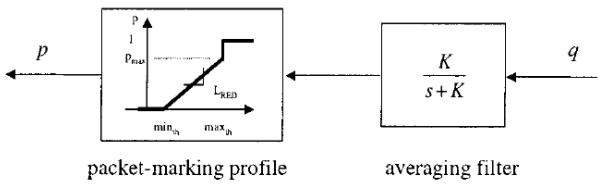


Figure 2: Block Diagram of RED as a cascade of low-pass filter and nonlinear gain element.

3.2 PI approach to AQM

According to [3] the transfer function of a PI controller can be written as:

$$C_{PI} = K_{PI}\frac{s/z+1}{s} \tag{3.3}$$

This controller is very well known by the control community. Its parameters can be tuned following methods proposed in the control literature. For example, [3] gave guidelines based on the Bode diagram tuning technique:

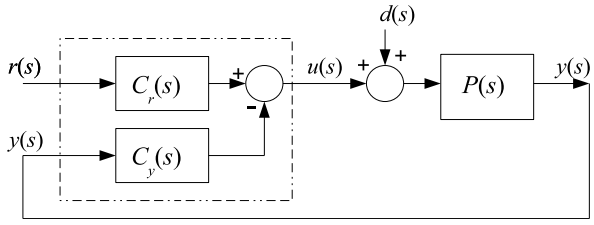


Figure 3: 2-DOF Control System

$$z = \frac{2N}{R_o^2 N} \quad K_{PI} = \omega_g z \left| \frac{j\omega_g + 1/R_o}{C^2/(2N)} \right| \quad \omega_g = \frac{\beta}{R_o} \tag{3.4}$$

where β determines the phase margin of the resulting nominal control system. The PI proposed in [3] is designed for a phase margin of about 30^o.

4 Robust PID Control

This section briefly presents the approach for Analytical Robust Tuning (ART) of 2-DoF PID controllers recently presented in [6]. Consider the *Two-Degree-of-Freedom* (2-DOF) control system of Fig. 3, where $P(s)$ is the *controlled process* transfer function, $C_r(s)$ the *set-point controller* transfer function, $C_y(s)$ the *feedback controller* transfer function, and $r(s)$ the *set-point*, $d(s)$ the *load-disturbance*, and $y(s)$ the *controlled variable*.

The output of the 2-DOF controller is given by

$$u(s) = C_r(s)r(s) - C_y(s)y(s) \tag{4.1}$$

For a PID_2 [7] it is

$$u(s) = K_c\left(\beta + \frac{1}{T_i s}\right) r(s) - K_c\left(1 + \frac{1}{T_i s} + T_d s\right) y(s) \tag{4.2}$$

where K_c is the *controller gain*, T_i the *integral time constant*, T_d the *derivative time constant*, and β the *set-point weighting factor* ($0 \le \beta \le 1$).

We will start right now with a Second-Order-Plus-Dead-Time (SOPDT) model of the form

$$P(s) = \frac{K_p e^{-L'' s}}{(T'' s + 1)(aT'' s + 1)}, \ \tau_o = \frac{L''}{T''} \tag{4.3}$$

The PID controller parameters are determined by the following equations for processes with parameters in the range $0.1 \le \tau_o \le 1.0$ and $0.15 \le a \le 1.0$.

$$\kappa_c = \frac{10\tau_i}{21\tau_c + 10\tau_o - 10\tau_i} \tag{4.4}$$

$$\tau_i = \frac{(21\tau_c + 10\tau_o)[(1+a)\tau_o + a] - \tau_c^2(\tau_c + 12\tau_o)}{10(1+a)\tau_o + 10a + 10\tau_o^2} \tag{4.5}$$

$$\tau_d = \frac{12\tau_c^2 + 10\tau_i\tau_o - (1+a)(21\tau_c + 10\tau_o - 10\tau_i)}{10\tau_i} \tag{4.6}$$

$$\beta = \min\left\{\frac{1}{K_c}, \frac{\tau_c T''}{T_i}, 1\right\} \tag{4.7}$$

The controller normalized parameters κ_c, τ_i and τ_d, and β depend on the model normalized dead-time τ_o and time constants ratio a, and on the design parameter τ_c. A minimum system robustness level is incorporated into the design process estimating a recommended maximum speed (τ_{cmin}) of the resulting closed-loop control system parameterized in terms of the maximum sensitivity function (M_s) by using

$$\begin{aligned} \tau_{cmin} &= k_{11}(M_s) + k_{12}(M_s)a^{k_{13}(M_s)} \\ & k_{11}(M_s) = 2.442 - 2.219M_s + 0.515{M_s}^2 \\ & k_{12}(M_s) = 10.518 - 8.990M_s + 2.203{M_s}^2 \\ & k_{13}(M_s) = 0.949 - 0.197M_s \end{aligned} \tag{4.8}$$

Combining the performance and robustness consideration above the design parameter may be selected in the range $\tau_{cmin} \leq \tau_c \leq 1.25 + 2.25a$. The range limits for the design parameter selection then combine the necessary restriction so that all controller parameters are positive and the accomplishment of a specified maximum sensitivity, with the necessity that the obtained response does not deviate too much away from the desired response, due of the dead-time approximation used in obtaining the tuning equations. For a more detailed presentation and discussion of the method please see [6].

5 Discussion

In order to illustrate the effectiveness of the Robust 2-DoF PID method, a numerical situation will be presented by taking the network simulation parameters of [3] where $q_o = 175$ packets, $T_p = 0.2$ seconds, $C = 3750$ packets/s (this corresponds to a 15MB/s link with an average packet size of 500 Bytes.). For a load of $N = 60$ TCP sessions we have $W_o = 15$ packets, $p_o = 0.008$, and $R_o = 0.246$. Therefore

$$P(s) = P_o(s)e^{-0.246} = \frac{1.17126\ 10^5}{(s+0.53)(s+4.1)}e^{-0.246} \quad \Delta(s) = 2.24\ 10^{-6}s(1 - e^{-0.246s}) \tag{5.1}$$

The corresponding RED controller parameters are (see [3]) $K = 0.005$ and $L_{red} = 1.86\ 10^{-4}$ whereas those of the PI controller $K_{PI} = 9.64\ 10^{-6}$ and $z = 0.53$. On the other hand, for the application of the ART method the parameters of the SOPTD model are needed. These result to be $a = 0.1297$, $L'' = 0.2467$ and $T'' = 1.9014$. And the controller parameters depend upon the desired robustness level expressed in terms of te M_s level. A minimum robustness is assured by $M_s = 2.0$ whereas highly robust systems are designed with $M_s = 1.4$. Here we take an intermediate level with $M_s = 1.6$. The resulting 2-DoF PID controller parameters are: $K_c = 4.4241\ 10^{-5}$, $T_i = 2.1443$, $T_d = 0.3436$ and $\beta = 0.6172$. Figure (4) shows the performance of the three presented control strategies applied to the nonlinear system (2.1). As it can be seen the RED controller cannot reach the new references. This is an inherent drawback because it lacks the corresponding integrator. On the other hand both the PI and ART-PID controllers reach the desired targets. On the right part of figure (4) it can de observed that the ART-PID controller reaches the desired set-point faster and without any overshoot. In addition, a test under different load conditions has been performed. In this case the number of TCP sessions

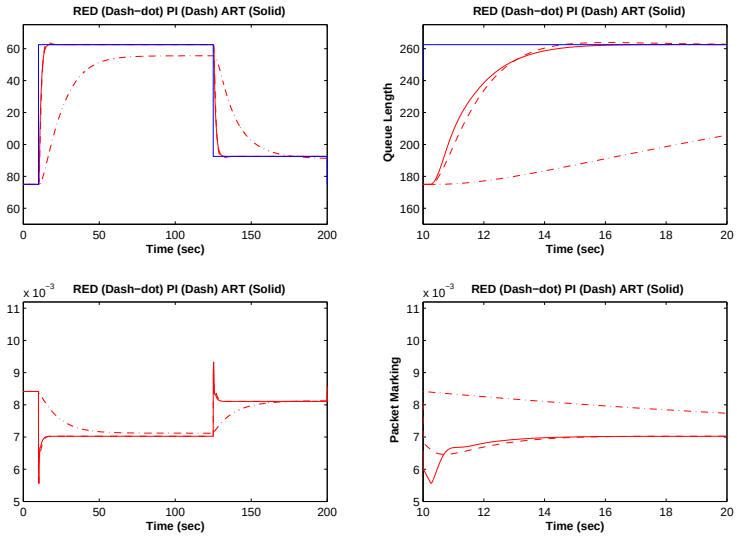


Figure 4: Changes on desired queue length. Comparison of PI, PID and RED performance. Plots on the right side show a zoomed version to better compare PI and PID

is a signal of the form $N(t) = N(1 + 0.01 * \sin(0.05t)) + \nu(t)$ being $\nu(t)$ a normally gaussian distributed random number of zero mean and variance $0.01 * N$. $\nu(t)$ is assumed to change with a sampling time of 5 sec. It can be seen that the ART-PID provides faster response to the load variation and faster recuperation of the desired queue length. Table (1) shows the mean value and standard deviation of the queue length error computed with respect to the desired target $q_o = 175 packets$. As a proof of performance, as we achieve a lower variation of the queue, a predictable performance will be expected. Therefore better QoS.

Table 1: Mean and standard Deviation of the queue length error

Controller	RED	PI	ART
mean	-0.43	-0.27	-0.1
std	10.9	9.2	5.0

6 Conclusions

In this paper the suitability of applying Robust PID controllers for the purpose of improving internet congestion control has been presented. The main advantage of the ART PID tuning is its one-parameter tuning. In addition this parameter is a direct specification of the desired robustness level. Therefore suitable for the sittuations with changes in load and system parameters. The performance has been compared to that of RED and a PI controller previously proposed in the literature.

Acknowledgment

This work has received financial support from the Spanish CICYT program under grant DPI2007-63356. Support from the Universidad de Costa Rica is greatly appreciated.

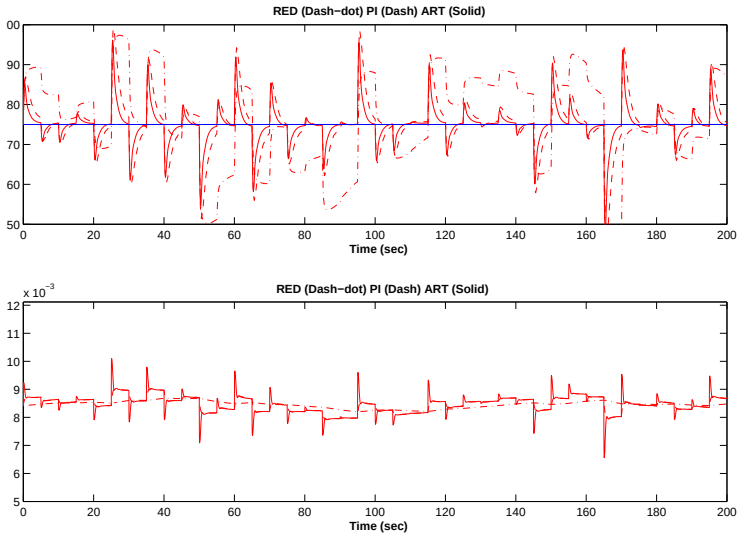


Figure 5: Regulation performance on a desired queue length of $q_o = 175 packets$ facing random variations of the number of TCP sessions

Bibliography

[1] S. Floyd and V. Jacobson, "Random early detection gateways for congestion avoidance.," *IEEE/ACM Transactions on Networking*, vol. 1, pp. 397–413, 1993.

[2] F. Kelly, "Mathematical modeling of the internet, in mathematics unlimited-2001 and beyond," *B. Enquist and W. Schmid, Eds, Berlin, Germany, Springer-Verlag.*, 2001.

[3] C. Hollot, V. Misra, D. Towsley, and W. Gong, "Analysis and design of controllers for AQM routers supporting TCP flows," *IEEE Trans. Automat. Contr.*, vol. 47, pp. 945–959, 2002.

[4] J. Postel, "Transmission control protocol," *RFC 793*, 1981.

[5] S. Ryu, C. Rump, and C. Qiao, "Advances in active queue management (AQM) based tcp congestion control," *Telecommunication Systems*, vol. 25, pp. 317–351, 2004.

[6] V. Alfaro, R. Vilanova, and O. Arrieta, "A single-parameter robust tuning approach for two-degree-of-freedom pid controllers," *The European Control Conference,Budapest, Hungary, August 23-26, 2009*, 2009.

[7] K. Åström and T. Hägglund, *Advanced PID Control*. ISA - The Instrumentation, Systems, and Automation Society, 2006.

Author index